Law Hum Behav (2008) 32:266–278
DOI 10.1007/s10979-007-9092-x

ORIGINAL ARTICLE

Predicting Sex Offender Recidivism. I. Correcting for Item Overselection and Accuracy Overestimation in Scale Development. II. Sampling Error-Induced Attenuation of Predictive Validity Over Base Rate Information

Scott I. Vrieze · William M. Grove

Published online: 3 July 2007

Abstract The authors demonstrate a statistical bootstrapping method for obtaining unbiased item selection and predictive validity estimates from a scale development sample, using data ($N = 256$) of Epperson et al. [2003 *Minnesota Sex Offender Screening Tool—Revised (MnSOST—R) technical paper: Development, validation, and recommended risk level cut scores*. Retrieved November 18, 2006 from Iowa State University Department of Psychology web site: http://www.psychology.iastate.edu/~dle/mnsost_download.htm] from which the Minnesota Sex Offender Screening Tool—Revised (MnSOST—R) was developed. Validity (area under receiver operating characteristic curve) reported by Epperson et al. was .77 with 16 items selected. The present analysis yielded an asymptotically unbiased estimator AUC = .58. The present article also focused on the degree to which sampling error renders estimated cutting scores (appropriate to local [varying] recidivism base rates) nonoptimal, so that the long-run performance (measured by correct fraction, the total proportion of correct classifications) of these estimated cutting scores is poor, when they are applied to their parent populations (having assumed values for AUC and recidivism rate). This was investigated by Monte Carlo simulation over a range of AUC and recidivism rate values. Results indicate that, except for the AUC values higher than have ever been cross-validated, in combination with recidivism base rates severalfold higher than the literature average [Hanson and Morton-Bourgon, 2004, *Predictors of sexual recidivism: An updated meta-analysis*. (User report 2004-02.). Ottawa: Public Safety and Emergency Preparedness Canada], the user of an instrument similar in performance to the MnSOST—R cannot expect to achieve correct fraction performance notably in excess of what is achievable from knowing the population recidivism rate alone. The authors discuss the legal implications of their findings for procedural and substantive due process in relation to state sexually violent person commitment statutes and the Supreme Court's *Kansas v. Hendricks* decision regarding the constitutionality of such statutes.

Keywords Base rate · Bootstrap · Cutting score · Incremental validity · Minnesota Sex Offender Screening Tool—Revised · Recidivism · Sex offender

Sex offenses, and in particular sex offender recidivism, present considerable problems for law enforcement, penology, and society as a whole. A number of states have enacted laws that aim to reduce recidivism through involuntary, expensive, and typically prolonged civil commitment of ostensibly high-risk sex offenders, following such offenders' completion of their originally imposed criminal sentences. These have been legally and ethically controversial, mostly due to individual rights arguments (Janus 2000), but such laws have withstood Constitutional scrutiny (*Kansas v. Hendricks* 1997). A central feature of *Hendricks*-type statutes is a requirement that the individual to be committed be judged as ''likely'' to re-offend (Janus and Meehl 1997). Typically, attention is restricted to the likelihood of specifically sexual re-offenses; in any case, sexual recidivism is emphasized over non-sexual new

S. I. Vrieze · W. M. Grove (✉)
Department of Psychology, University of Minnesota, Twin Cities Campus, N218 Elliott Hall, 75 East River Road, Minneapolis, MN 55455-0344, USA
e-mail: grove001@umn.edu

S. I. Vrieze
e-mail: vrie0006@umn.edu

crimes. If one is to select offenders for potential commitment based on probable reoffending, then for reasons of efficiency and distributive justice, one needs as unbiased and accurate a prediction method as possible.

Predictive accuracy depends on the validity of the predictor, the base rate of the predictand, and the adaptation of predictor to predictand. The further the base rate departs from 1/2, the more valid a predictor must be to improve on ''betting the base rates,'' that is, predicting the modal outcome for all individuals (Meehl and Rosen 1955). Sex offender recidivism rates have been reported to range from 5% (Minnesota Department of Corrections 2000b) to 52% (Prentky et al. 1997), depending on the criterion and the follow-up interval. Hanson and Morton-Bourgon's (2004) meta-analytic estimate of 13.7% (N = 24,040 offenders; 84 studies) is a sex offender multiyear recidivism rate based on by far the most available data.

No matter how hard is recidivism to detect, the passage of sexually violent person (SVP) commitment laws in 17 states mandates that recidivism predictions about individuals be made. Clinical and ''actuarial'' (mechanical, formal) methods are currently used to make such predictions. The literature on clinical versus statistical prediction gives strong reason to conjecture that mechanical predictions of recidivism will perform as well as or better than clinical ones (Grove et al. 2000). However, few sex offender-specific studies have compared these methods directly (Litwack 2001). There are many actuarially-based methods in use, including the Violence Risk Appraisal Guide (VRAG; Harris et al. 1993), Sex Offender Risk Appraisal Guide (SORAG; Quinsey et al. 1998), Rapid Risk Assessment of Sexual Offense Recidivism (RRASOR; Hanson 1997), Static-99 (Hanson and Thornton 1999), and the Minnesota Sex Offender Screening Tool—-Revised (MnSOST—R; Epperson et al. 2003).

The MnSOST—R is used in 13 of the 17 states that have *Hendricks*-type commitment laws, and is reportedly used in more jurisdictions than the RRASOR or the SORAG (Dennis Doren, personal communication, October 10, 2005). This is concerning because there is relatively little validity information on the MnSOST—R. We are aware of three peer-reviewed studies: Barbaree et al. (2001), Bartosh et al. (2003) and Langton et al. (in press), whose sample overlaps substantially with that of Barbaree et al. Aside from the original development report (Epperson et al. 2003) we are aware of just two sources of validity information outside the peer-reviewed literature: a conference paper by Epperson et al. (2000); and a conference paper by P.R. Kropp (personal communication, April 18, 2006).

Epperson et al. (2000) presented a cross-validation sample at a sex offender treatment conference, but never published it in a peer-reviewed journal. About 220 Minnesota sex offenders released in 1992 were studied, excluding 50 individuals re-imprisoned for non-sexual offenses because they had little opportunity to commit a new sexual offense. The AUC was .73 ($p < .04$).

P.R. Kropp (personal communication, April 18, 2006) presented a cross-validation sample at a sex offender treatment conference, but also never published his results in a peer-reviewed journal. The MnSOST—R was administered to just 53 sex offenders incarcerated in Canadian federal prison. The point biserial correlation coefficient was .18. Given the study's recidivism rate of approximately .28, the equivalent AUC would be .73 ($p > .05$).

Bartosh et al. (2003) examined the MnSOST—R in Arizona sex offenders. They split subjects into $N = 73$ rapists and $N = 59$ extrafamilial child molesters. (The MnSOST—R is not designed to predict intrafamilial molester recidivism.) The MnSOST—R did not significantly predict sexual recidivism for rapists (AUC = .585, $p < .31$) or molesters (AUC = .586, $p < .36$).

Barbaree et al. (2001) and Langton et al. (2006), when examined jointly, present interesting findings. Barbaree et al. studied the MnSOST—R in 150 Ontario sex offenders. The AUC was .65 ($p < .13$, using Bamber's 1975, formula for the variance of the AUC).

Langton et al. (in press) contains data previously reported in Langton (2003). Langton et al. (and Langton 2003) extended and refined Barbaree et al. (2001). Langton et al. studied 136 of 150 Barbaree subjects plus 218 new ones, 354 offenders in all. Hitherto unavailable file information was acquired, allowing incomplete or questionable MnSOST—R ratings to be recoded, and the MnSOST—R was completed for all the new subjects. The AUC for the $N = 354$ sample was .70 ($p < .001$). Langton's (2003) AUC for 117 of the 150 Barbaree subjects was .69 ($p < .05$) Langton (2003) conjectured that this improved result for the Barbaree et al. subset was due to more complete information and better-trained MnSOST—R raters.

Validity studies on other instruments are considerably more plentiful. Hanson and Morton-Bourgon (2004) list six studies (three peer-reviewed, N = 1,348) on the SORAG, 17 studies (six peer-reviewed, N = 5,266) on the RRASOR, and 21 studies (seven peer-reviewed, N = 5,103) on the Static-99. Notably, Hanson and Morton-Bourgon (2004) did not find significantly different AUCs (albeit in limited power studies) between the MnSOST—R (mean AUC = .66) and the other, better-researched instruments named above (mean AUC = .48–.77). However, their statistical methods were faulty. They failed to correct for non-independence of AUC estimates based on common samples, which necessitates that a special test (Hanley and McNeil 1983b) be combined with the usual independent-samples test (Hedges and Olkin 1985) they employed. This mistake may have masked true differences between AUCs. Hence, their failure to find different mean AUCs between the instruments studied, including the MnSOST—R, cannot be accepted at face value.

The literature is thus equivocal on the MnSOST—R's validity as a predictor of sexual reoffending. As Hanson and Morton-Bourgon (2004) state, further analyses are required.

When the MnSOST—R debuted, the work was never published in a peer-reviewed journal, causing concern in forensic circles (Campbell 2000; Janus and Prentky 2003; Wollert 2002, 2003). The MnSOST—R development sample data were acquired from the Minnesota Department of Corrections (DOC), and the present authors re-analyzed them with two aims:

1. estimating the effect of sampling error on overselection of items and overestimation of accuracy (AUC) in the development of the MnSOST—R, using a special form of statistical bootstrapping (Shao 1996); and
2. estimating the expected correct fraction of a recidivism prediction instrument under practical conditions, when that instrument has population accuracy (AUC) at several values representative of those reported for the MnSOST—R (and other sex offender recidivism prediction tools), at various postulated recidivism rates.

The first goal required development of a computerized version of a close approximation to the steps Epperson et al. (2003) followed to make decisions about item rating level weighting, item selection, and cutting score selection based on the development sample. The present authors initially developed this algorithm by close reading of the ''Final Report'' document describing the development of the MnSOST—R (Epperson et al. 2003). It became clear from this reading that certain information was omitted or still unclear, so Prof. Epperson was personally contacted for clarification. He generously explained in detail the procedures followed, including:

1. the identification of every potential item considered for inclusion in the MnSOST—R, out of the larger data set the Minnesota DOC delivered to us;
2. how missing values were imputed;
3. how certain items were recoded and combined a priori (i.e., not based on intercorrelations with each other, or validity correlations with recidivism) into summary items;
4. more detailed information about how the item coding level weighting procedure (called a ''modified'' Nuffield weight system) worked (see below); and
5. more detailed information about how the four stages of item selection employed by Epperson et al. were implemented (see below).

With a computerized algorithm embodying these procedures, one could mechanically apply Epperson et al.'s (2003) method to a sample of data possessing the requisite variables. The present authors used the algorithm repeatedly, applying it to samples drawn with replacement from the development data set, with the result that each sample yielded its own, unique MnSOST—R-type instrument. This is the heart of the statistical bootstrapping method explained in detail below.

The second study goal was to investigate the expected population performance of such an instrument as the MnSOST—R, having estimated AUC of a certain magnitude, when its application is subject to three kinds of sampling error: (a) error in estimating the AUC corresponding to error in estimation of test discriminating power, (b) error in estimating the recidivism base rate where the test will be used, and (c) error in choosing the test cutting score that minimizes classification errors, given the apparent recidivism rate.

The AUC has the advantage of being a prevalence-free measure of predictive accuracy and has been promoted by various investigators (Mossman 1994a, b; Rice and Harris 1995). Correct fractions (the epidemiologist's hit rate, or the proportion of recidivists and non-recidivists correctly identified as such by the test) are of interest in their own right, because decisions such as ''commit/don't commit'' are routinely being made, informed by predictions of ''will/won't reoffend,'' which correspond naturally to right/wrong tabulation (Amenta et al. 2003; Campbell 2003; Szmukler 2001). Alas, there is only a general correspondence between the AUC and correct fractions; the AUC only informs one about best-case classification accuracy—accuracy that is achievable in practice only if quite restrictive assumptions are *all* satisfied:

1. the AUC is estimated without error;
2. the mathematical model relating the form of the ROC, its area, and the distributions of recidivists' and non-recidivists' test scores is exactly valid, so that a certain equation relating the AUC to the difference between group test score distributions' means is exactly correct;
3. the recidivism base rate is estimated without error; and
4. the equation relating the forms of recidivists' and non-recidivsts' test score distributions, their mean separation, and the recidivism base rate, is solved for the exactly optimal cutting score that minimizes classification errors.

If one or more of these assumptions is materially false, then to that extent the classification accuracy achieved may fall short of what one might expect from a published AUC figure. In particular, suppose that sampling error intrudes—as it must. Then investigation beyond the AUC is required to determine what classification performance can be expected from a test, (a) whose AUC must be estimated, (b) which will then be employed in a population whose recidivism rate must also be estimated, (c) using an estimated cutting score to make predictions.

Method

Reproduction of Epperson et al. (2003) Analyses

To try to reproduce Epperson et al.'s results, we followed their steps as precisely as possible (Epperson et al. 2003; D. Epperson personal communication, June 20, 2001). This is the procedure:

1. *Identification of candidate predictors.* The investigators started with about 825 potential predictor items, 178 of which were actually considered for inclusion in the MnSOST—R. (We retrieved these variables from spreadsheets provided by the Minnesota DOC.) Certain items were combined and/or recoded a priori; we started with the combined/recoded items in our analyses.
2. *Modified Nuffield weights* (Nuffield 1982). Each coding level for each categorical item had its associated recidivism rate calculated; weights were assigned to each level, as rounded integer multiples of the difference between the level-specific recidivism rate and the overall item's recidivism rate, divided by .05. Adjacent coding levels not differing in weights, or levels containing less than 10% frequency, were combined and their combined weight recalculated; this process continued until no changes occurred. Any nonmonotonicity existing in coding-level weight assignment at the end of the procedure derives from empirical, nonmonotonic relationships between item rating levels and recidivism rates. Epperson (personal communication June 20, 2001) indicated that subjective-theoretical considerations were taken into account in deciding whether to combine rating levels; after consultation, one of the present authors (WMG) wrote an algorithm and accompanying computer program capturing as closely as possible his stated policies. The final part of this step was to delete all items whose every rating level had the same weight.
3. *Screen items for Pearson r with recidivism.* Among items surviving step (2), drop any whose Nuffield-weighted scores' correlations with the criterion are not significant, $p < .10$.
4. *Block-entry logistic regression.* Surviving items were divided into four groups. Item groups were entered successively into logistic regression to predict new sex offenses, with the Nuffield-weighted item scores serving as predictors for each individual. The groups were: (a) dynamic (potentially treatment- or development-changeable) variables (e.g., discipline history, chemical dependency treatment while incarcerated, sex offender treatment while incarcerated, age at prison release), (b) chronicity (e.g., adolescent antisocial behavior, number of sex offense convictions, number of different age groups against which offenses committed, length of sex-offending history), (c) sex offense topographic variables (e.g., use or threat of force in any sex offense, 13–15 year old victim in any sex offense, any sex offense committed in public, victim a stranger, multiple acts in any single offense, sex offense committed while under penal supervision), and (d) variables reflecting instability (e.g., unstable employment, history of substance abuse). Each variable having partial χ^2 with $p < .15$ was retained; no retained variable was ever later removed.

   Epperson et al. (2003) only classified those 16 items into blocks for logistic regression analysis that survived their steps (1)–(3); the other 178 – 16 = 162 items had no group assignment. We created an ''Other'' category for these items. We entered the items in this block last, after blocks (a)–(d), if its items survived steps [1]–[3] in our re-analysis, into the logistic regression. To retain an ''Other'' item in the logistic regression, we required a partial χ^2 with $p < .15$.

All items surviving step (4) constituted the final instrument. The sum of Nuffield-weighted item scores constituted the MnSOST—R score for each individual.

Consistent Model Selection Bootstrapping

A major purpose of this study was to estimate the validity of the MnSOST—R, without capitalizing on chance as did Epperson and colleagues. Bootstrapping (Shao 1996) can give an approximately unbiased estimate of accuracy using the same cases to develop a scale and to test its validity. This method relies on repeatedly sampling with replacement from the study sample as if it were a population, calculating a statistic of interest (in this case, the AUC) for each ''pseudo-replicate'' sample.

Ordinary bootstrapping, in which a sample of size N is drawn with replacement from a ''population'' of size N, works well for some types of models, but fails badly for model selection problems like the one involved in building an instrument. A helpful reviewer has raised the question Whether ordinary bootstrapping would be more appropriate than Shao's (1996)? The reason that ordinary bootstrapping is inappropriate for this study is the large-scale winnowing of the variables of the data set; the Epperson et al. (2003) development procedure reduced 178 candidate variables to 16. Ordinary bootstrapping is designed to deal with an unselected set of variables (e.g., items) and to estimate sampling variability associated with the weights attached to the items, and performance statistics associated with the items (AUC, etc.). Introducing variable selection brings in an additional source of sampling variability: one bootstrap pseudoreplicate sample can have a different set of variables from another bootstrap sample, not just a different set of weights. Shao (1996) has proven that the ordinary bootstrap produces the correct model (i.e., a model with all those variables which have non-zero weights in the population and only those variables with non-zero weights in the population) with probability approaching zero as N approaches infinity. Shao's method of bootstrapping was developed specifically to produce (a) the correct model, in terms of selected variables with probability approaching one as N approaches infinity and (b) asymptotically unbiased estimates of the model weights (i.e., weights of variables obtained in the logistic regression analysis that is the final step in Epperson et al.'s [2003] and our procedure). In other words, ordinary bootstrapping produces a *statistically inconsistent* result, whereas Shao's (1996) bootstrapping produces a *statistically consistent* result that is to be preferred in this type of analysis.

Shao's (1996) procedure has three steps: (1) drawing numerous (in our study, 1,000) pseudo-replicate samples with replacement, each of size m, from the whole sample of N. m is chosen much smaller than N, such that $m/N \rightarrow 0$ as $N \rightarrow \infty$, e.g., $m = N^{3/4}$. Here, $m = 64$. (2) For each pseudoreplicate sample, carry out the whole series of MnSOST—R item selection steps (1)–(4) above; record which items pass all four steps and their Nuffield weights, as well as the χ^2/df of the final model (instrument). The latter is recorded because the logistic regression χ^2 is what is optimized by variable selection; dividing by the df compensates for different instruments having differing numbers of items. (3) Evaluate the χ^2/df of each model in the whole $N = 256$ sample. This is analogous to applying a sample-based procedure to the whole population. (4) After obtaining 1,000 models' χ^2/df values, one examines the distribution of the corresponding AUCs, as these are of much greater interest: particularly the middle (mean, median) of this distribution and its spread (SD, confidence intervals).

Translation of Population AUC to Expected Correct Fraction under Sampling Error

It has already been explained how sampling error, as well as less than certain distribution of test scores, degrades the perfect relationship between population AUC and attainable classification correct fraction. To investigate the extent of this problem, we simulated the distribution of expected population correct fractions, employing the following steps:

1. A suitable quantitative specification of the non-recidivist and recidivist population score distributions on the prediction instrument is developed (probability density functions, cumulative distribution functions, quantile functions, and pseudo-random number generators). Concretely, the two population distributions are arrived at in stages as follows. First, the non-recidivist ($N = 166$) and recidivist ($N = 90$) samples (plus 11.05, added for convenience to make all MnSOST—R scores positive) are used to calculate kernel density estimates (KDEs) over the range 0–40 via unbiased cross-validated bandwidth estimation and an optcosine kernel (chosen because it is approximately mean-square-error efficient; Scott 1992). Next, cubic spline interpolation is used to ''fill in'' the two frequency curves (density estimates times $\hat{P}$ and $\hat{Q}$, respectively) and put both subpopulations on the same X-axis values, .001 score points apart. Finally, separate multinomial approximations to the spline-interpolated KDEs are used for looking up probability density function, cumulative density function, and quantile function values, and for generating pseudorandom observation values. By following these steps, we avoid assuming a particular parametric form (e.g., Gaussian, gamma) for the MnSOST—R-like instrument score distributions, instead forcing the score distributions to be as much like the Epperson et al. (2003) sample distributions as possible, subject to the constraint that the two distributions be smooth.
2. The simulation proper begins by estimating values $\hat{A}_z$ (the sample AUC value) from its asymptotic distribution according to Bamber (1975) with $N = 256$ and $P \approx .32$, the proportion of recidivists in Epperson et al. (2003). Estimated $\hat{A}_z$-values were simulated from population AUC values chosen to equal either .5813, .6796, or .77. The former is the present analysis's bootstrap AUC estimate; see below. .6796 is the weighted mean AUC for the MnSOST—R (Hanson and Morton-Bourgon 2004), and .77 is the original reported AUC for the MnSOST—R (Epperson et al. 2003); these values thus essentially cover the range of reported AUC values for the MnSOST—R. Given heteroscedastic binormal ROC assumptions, $\hat{A}_z$ values dictate corresponding estimated mean differences $\hat{d}_a$ between recidivist and non-recidivist test score distributions according to the following mathematical transformation,

$$\hat{d}_a = \Phi^{-1}\left(\hat{A}_z\right)\sqrt{1 + \frac{\sigma_n}{\sigma_{sn}}},$$

where Φ^{-1} is the standard normal cumulative distribution function, and σ_n and σ_{sn} are the standard deviations of the non-recidivists and the recidivists, respectively. We checked whether these assumptions were critical by seeing whether we could reproduce the observed mean difference between non-recidivists and recidivists in Epperson et al. (2000) even though the score distributions were *far* from normal. The formula yields (after appropriate transformation to account for scale) the correct mean difference to within .1, out of an observed mean difference of 5.52 (less than 2% error);

3. $N\hat{P}$ is simulated binomially, then divided by N to obtain $\hat{P}$ values. $\hat{Q} = 1 - \hat{P}$;
4. instrument score data X_{ni} (i = 1,...,N_r) for recidivists and X_{nri} (i = 1,...,N_{nr}) for non-recidivists are simulated from separate population distributions developed as described in step 1. For each simulated value $\hat{A}_z$ from step 2, with its corresponding value of $\hat{d}_a$, the mean difference between to-be-sampled score distributions is adjusted from its observed (Epperson et al. 2000) sample value of 5.52 raw score units to $\hat{d}_a$ SD units according to the $\hat{d}_a$ calculated during step 2; the variance was kept equal to the originally observed variances in Epperson et al. Pseudorandom scores on the prediction instrument were generated separately for recidivists ($N_r = N\hat{P}$ in number, rounded to nearest integer) and non-recidivists (N_{nr} = 256 minus rounded $N\hat{P}$ in number) by use of a multinomial pseudorandom number generator, followed by linear transformation to the appropriate MnSOST—R score metric;
5. the estimated optimum cutting score $\hat{X}_c$ for a given pair of non-recidivist and recidivist score samples (in relative abundance $\hat{P} : \hat{Q}$) was then sought. This was accomplished in two steps. First, KDEs (unbiased cross-validated bandwidth, optcosine kernel as in step 1) were calculated, and were cubic spline interpolated to a common basis on the interval (0, 40) at .001 intervals. Second, the region from the highest mode of the non-recidivist KDE to the maximum score in either sample was examined for an intersection of the $\hat{P}$- and $\hat{Q}$-weighted KDEs (known as the Hitmax cut; Meehl 1965). In this way, we avoided assuming that the forms of the test score distributions were known to a user of the prediction instrument. This Hitmax cut is taken as the estimated optimal cutting score $\hat{X}_c$;
6. apply the estimated sample cutting score $\hat{X}_c$ to the *parent* test score distributions, i.e., the true distributions reflecting no sampling error, as these were specified in step 1. That is, one approximates via numerical integration,

$$\alpha \approx 1 - \int_0^{\hat{X}_c} KDE_r \text{ and}$$

$$\beta \approx \int_0^{\hat{X}_c} KDE_{nr},$$

where α is the epidemiologists' sensitivity, β is specificity, $\hat{X}_c$ is the cutting score, and KDE_r or KDE_{nr} is the spline-interpolated kernel density estimate for the recidivist or non-recidivist *population* from step 1, respectively. From these quantities, the correct fraction, in the parent *mixed* population, is simply CF = $P\alpha + Q\beta$. Note that P and Q here are *not* estimated in this equation (and so do not have ''hats'' on them), and likewise these KDEs are only subject to numerical approximation error, *not* sampling error (and so do not have ''hats'' on them);
7. repeat steps (2) to (6) 1,000 times per combination of true $A_z \times P$ values, to obtain a distribution of expected population correct fractions for 1,000 sample-estimated cutting scores $\hat{X}_c$. Examine the middle (mean, median) and spread (quantiles) of each parameter combination-specific CF-distribution. Also, compare the correct fractions to those obtainable by ''betting the base rate,'' i.e., always predicting the most frequent outcome (if $\hat{Q} > .5$, predict non-recidivism for all offenders, otherwise predict recidivism for all offenders), regardless of instrument score (Meehl and Rosen 1955).

The correct fraction above and beyond that achievable from base rate information alone is a critical measure of a test's incremental validity, for a specific combination of A_z and P. The CF will reflect real life conditions, including sampling error, which always intrudes into the estimation of (1) recidivism rates and (2) cutting scores that are apparently optimal for local recidivism rates.

A reviewer has questioned the relevance of the CF, as unlike A_z it is *base rate dependent*. The answer is that in application, the instrument user does not experience ''base rates in general,'' which is what A_z measures predictive accuracy for. Instead, the clinician experiences a particular base rate—the one in the population from which their examinees are drawn. There is an associated cost for a misprediction at that base rate, not at ''base rates in general.'' It is true, as pointed out by one reviewer, that a proper decision analysis would consider not only the proportion of erroneous classifications but also the costs associated with these errors—distinguishing positive from negative classification errors as we do not. However and unfortunately, there is no general agreement among decision makers about the relative importance of false positive and false negative prediction errors. Therefore, we followed custom in the statistical discrimination literature and assumed zero–one loss: zero relative cost if no prediction error occurs, versus a cost of one if either a false negative or a false positive prediction error occurs. Absent a legislative clarification of how important is mistakenly committing a low-risk sex offender, as compared to letting a high risk sex offender live at large in the community, we are loath to impose our private evaluations of relative importance.[1]

The bootstrapping procedure used to achieve our first study goal was written in the Statistical Analysis System (SAS Institute 2005), using the macro facility, Output Delivery System, DATA step, and several statistical procedures. The AUC-correct fraction simulation was written in the R graphics and statistics programming environment (Venables et al. 2002). All code is available on request. The use of human subjects data was approved by the University of Minnesota Institutional Review Board (IRB) as an exempt secondary use of data; original approval was granted to Epperson and colleagues by the Minnesota DOC IRB (Steven Huot personal communication, May 15, 2006).

Results

The main result of our reanalysis appears in Table 1. The m = 64-based model having the lowest N = 256-based residual χ^2/df for fit of the hierarchical regression model is displayed, since this was the criterion Epperson et al. used to decide on final item selection. The actual value of this statistic is not unbiased. SPSS (SPSS, Inc. 2005) variable names, as received in Minnesota Department of Corrections system files, are given for the items, as well as an explanation of which each item measures. The reader will immediately note that whereas 16 items were selected by Epperson et al., only four were selected by the bootstrap. Clearly, the bootstrap results are consistent with the hypothesis that overfitting (overselection of items) has taken place during scale development by Epperson et al. This would tend to make the performance of their scale less stable than desirable in repeated applications to the same population, let alone applications across distinct populations (e.g., different base rates, let alone different offender characteristics). Such overfitting by Epperson et al. would not be surprising—indeed it would be predicted—because Shao (1996) has shown that the probability of selecting the correct model using a procedure like that of Epperson et al. goes to zero as N goes to infinity.

Direct comparisons of the items selected, and the AUCs estimated, must be made with special care. The procedures followed by Epperson et al. differed in one important way from the one we followed: in assigning (modified) Nuffield weights, Epperson et al. allowed subjective, theory-mediated considerations to influence their decisions. While we strove to write an algorithm that captured as much of their policies as a mechanical procedure could do, based on consultations with Epperson, it is of course in principle impossible for any computer program exactly to mimic a human judge's decisions. As the Nuffield weight assignments were the first step in Epperson et al.'s multistep scale-building procedure, any differences between their procedure and that of the present authors, even if relatively small, might have a ''snowball'' effect on subsequent stages, ultimately resulting in significant differences in items selected and the AUC calculated. The question of comparability (or lack thereof) between the Epperson et al. results and those of the present authors is taken up in the ''Discussion.''

Bootstrap Analysis

First, we describe the models generated in the 1,000 m = 64 bootstrap pseudoreplicate samples, in terms of the variables per model, and then in terms of the models themselves. Next we describe the models' performances.

A total of 178 items went into the selection process. About 95 of these were actually retained in at least one model, while 85 did not pass muster for any model. Two models retained no variables at all. The five-number summary regarding items per model was 0, 5, 7, 8, and 15. (A five-number summary gives the minimum, first quartile, median, third quartile, and maximum.) No individual item was included in as many as 22% of the 1,000 models developed. Clearly, quite different variables were being selected across different pseudoreplicate samples.

The data can also be looked at in terms of models per item. The number of models in which an item participated was 0–218. The five number summary was 0, 0, 23, 51, 218. Considering only those items that were retained in at least one model, the five number summary was 1, 12, 40, 59, 218.

The main result of interest is, of course, the performance of the models (or instruments) developed. One is interested both in typical performance and the distribution about the average.

The distribution of whole-sample (N = 256) AUCs for those 998 pseudoreplicate models that retained at least one item was as follows: M = .5813, M ± SD = (.5395, .6231). The 95% bootstrap basic percentile confidence interval was (.4882, .6744), which notably includes .5, the chance-level

[1] One might imagine that headway on this problem could be made by assuming certain values, e.g., values in a certain range, for the ratio of relative costs of false positive versus false negative classification errors. However, it can readily be proven that there always exists some value of this ratio for which *any* instrument, no matter how weakly it outperforms a mere flip of the coin in predicting sex offender recidivism, has huge utility. Contrariwise, there always some (other) value of this ratio for which such an instrument, no matter how strongly it outperforms a flip of the coin (as long as it is not perfectly accurate, $A_z = 1$), has huge disutility. Proof omitted to save space, but available from the authors on request.

Table 1 Items selected by the bootstrap model selection analysis

Number	SPSS variable ID	Nature of item
1	Sohx	Length of offending history
2	Cdtx	Chemical dependency treatment (completed or in)
3	Supfail	Number of supervision failures (e.g., probation violations)
4	Xcdecep	Offense in which compliance achieved through deception
	AUC = .5813	

value for an AUC; but the 90% c.i. lower limit just excludes .5. That is, the bootstrap analysis rejects H_0: $A_z = .5$ at $.025 < p < .05$. The five-number summary for whole-sample AUCs is .4002, .5628, .5863, .6077, .6867; one can say that a typical AUC in our analyses ran about eight and one-half percent better than chance-level discrimination.

Effect of Sampling Error on Translation of Population AUC into Achievable Correct Fraction

The final analysis investigated the effect of sampling error on the ability to turn a given level of AUC into an effective population correct fraction, by selecting a suitable cutting score for a local recidivism base rate. As described in ''Methods''—for an assumed range of population AUC and recidivism base rate values—the AUC, recidivism base rate, and optimum cutting score were estimated in 1,000 simulated finite samples ($N = 256$ each for the AUC and cutting score estimators, and $N = 2{,}000$ for the recidivism rate estimator). (Assuming that recidivism rate estimation is done on $N = 2{,}000$ is quite a generous concession on behalf of the test; the Minnesota recidivism rate estimate is based on less than one tenth this number, $N = 192$; Minnesota Department of Corrections 2000a.) The optimum cutting score was sought by kernel density estimation followed by interpolation for the intersection of the estimated base rate-weighted densities.

Table 2 gives the average population correct fractions for various combinations of AUC and recidivism rate values. The pattern is quite simple. Except for $P = .4$, the CF for the instrument is either less than that of betting the base rates or never exceeds it by more than a very few percent (e.g., 3.7% at most). At $P = .4$, the instrument ''beats the base rate'' by amounts varying from 5.8% (AUC = .5813) to 8.5% (AUC = .77). Note that a base rate of .4 is over *three times* that of the weighted mean sex offender recidivism rate reported on the largest meta-analysis to date (Hanson and Morton-Bourgon 2004). Also note that the AUC of .77 is the highest AUC ever reported not only for the MnSOST—R but also, to our knowledge, for any sex offender recidivism prediction instrument; it is not cross-validated. Concentrating on what are much more typical values: at the average AUC reported for the

Table 2 Correct fractions in 1,000 simulated trials of estimating cutting score and applying it to parent populations

AUC	Base rate	Five number summary: correct fraction					Expected correct fraction	Correct fraction for ''betting the base rate''
		5%	Q1	Median	Q3	95%		
.5813	.05	.9477	.9496	.95	.9505	.95	.9496	.95
	.1	.8996	.9	.9	.9022	.9031	.9002	.9
	.137	.8630	.8630	.8630	.8701	.8709	.8645	.863
	.2	.8	.8	.8055	.8176	.8179	.8067	.8
	.3	.7	.7155	.7270	.7426	.7460	.7248	.7
	.4	.6159	.6483	.6606	.6882	.7015	.6582	.6
.6796	.05	.9476	.9492	.95	.9505	.95	.9495	.95
	.1	.8996	.9499	.9001	.9029	.9031	.9004	.9
	.137	.8630	.8641	.8670	.8708	.8709	.8657	.863
	.2	.8022	.8044	.8145	.8178	.8179	.8097	.8
	.3	.7244	.7272	.7386	.7439	.7461	.7317	.7
	.4	.6664	.6670	.6766	.6966	.7017	.6750	.6
.77	.05	.9476	.9485	.9499	.9499	.9501	.9492	.95
	.1	.8996	.9	.9001	.9029	.9031	.9006	.9
	.137	.8630	.8641	.8670	.8708	.8709	.8668	.863
	.2	.8022	.8103	.8145	.8179	.8179	.8130	.8
	.3	.7244	.7346	.7386	.7456	.7461	.7374	.7
	.4	.6664	.6773	.6853	.7008	.7017	.6849	.6

MnSOST—R (also very close to the average AUC reported for other recidivism prediction instruments such as the SORAG, VRAG, Static-99, etc., in the Hanson and Morton-Bourgon meta-analysis), and at the average recidivism base rate reported in that meta-analysis, namely 13.7%, the simulated instrument has an incremental validity measured in CF, compared to betting the base rates, that is estimated to be .8657 – .863 = .0027. This is, as far as the present authors are concerned, so close to zero as to make no difference.

The reason the figures in the table appear as they do—viz., quantiles of CF running close to but less than P is as follows. For A_z = .5813, .6796, and .77, substantial distribution overlap occurs. Coupled with modest recidivism rates and right-skewed scores, recidivist distributions generally ''nestle'' beneath the non-recidivist distribution's right tail, rather than intersecting with the recidivist distribution; an optimal cutting score cannot be found at an intersection of the recidivist and non-recidivist test score distributions. Instead, a default cutting score at an extremely high value of X has to be taken. This makes the instrument perform much like, and about as well as, betting the base rate.

Discussion

This study adds a rather negative finding on the validity of the MnSOST—R to an already equivocal record of previous efforts to cross-validate this instrument.

Comparison to the Literature

The reader may view the bootstrap-estimated population AUC of .58 with skepticism, reasoning that it does not match up to other cross-validation studies' AUC values. One anonymous reviewer pointed out, for example, that our value of .5813 differs significantly (by Hanley and O'Neil's [1983a] test) from the .73 value reported by Epperson (2000). However, no other study's AUC differs from our .5813 result at the $p < .05$ level. Most important, our value's confidence interval overlaps with that of the most carefully conducted, comprehensive review of the prediction literature to date; Hanson and Morton-Bourgon's (2004) weighted confidence interval for the MnSOST—R: our 90% interval's upper end is at .67, their lower end is at .64.

Comparison with Epperson et al. (2003)

An anonymous reviewer opined that there is no fair comparison between Epperson et al.'s results and ours, due to methodological differences. This is too strong. Some generalizations and plausible quantitative limits can be placed on the direction, and probable size, of the difference between Epperson et al.'s predictive validity, and our predictive validity, based on the clinical versus mechanical prediction literature (Grove et al. 2000). This is because the former investigators used a partially judgment-based and partially statistically-driven scale-development procedure, whereas the present authors used an entirely mechanical, statistics-driven procedure; Grove et al. included studies that examined such hybrid prediction schemes.

We can examine statistics on the effect sizes (ESs) from Grove et al. (2000), as the best comprehensive source on how well subjective versus mechanical data combination works.[2] The average advantage enjoyed by purely statistical data combination was .086; the 5% and 95% quantiles ran from .50 to –.12, measured in an ES metric that allowed intercomparison of various kinds of predictive accuracy statistics, including correlations, correct fractions, AUCs, and others. When translated into correlations for convenience, this works out to an average of $r = .0858$, approximately (quantiles –.1194, .4621). Hence, the reader is most plausibly entitled to expect that the predictive validity of Epperson et al.'s items at Stage 1 was attenuated by almost .09 in correlation collectively, of course this being made up by Epperson et al.'s failure to cross validate their whole scale's predictive accuracy, resulting in a considerably inflated AUC.

Caveat Regarding the AUC as Sole Measure of Predictive Accuracy

An AUC statistic tells a researcher or clinician how well a test performs across the whole potential range of disorder (recidivism) base rates and optimum cutting scores for those base rates. It can lull the clinician into thinking that, if the AUC is suitably high, the test will perform satisfactorily in a given population, i.e., at a given base rate. This is far from necessarily so; a sufficiently high AUC only ensures that a test *can* perform well in a population with a certain base rate, *if* the cutting score is appropriately set. The point of our third set of analyses is to bridge the gap between population AUC and attainable correct fraction, taking into account sampling errors that degrade test performance. These analyses show that, unless the recidivism rate is more than three times higher than the aggregate figure reported in the literature

[2] The Grove et al. (2000) meta-analysis concerned 136 head-to-head comparisons of judgment accuracy (or prediction) in health and human behavior, across a very wide variety of domains (diagnosis, prognosis, criminology, mental health, etc.), employing a wide variety of judges (mental health professionals, medical doctors, etc.) at varying levels of ostensible expertise and training (from novices to those fully trained, and with many years of experience).

(Hanson and Morton-Bourgon 2004), an instrument with predictive accuracy like the MnSOST—R will not provide predictive accuracy (measured in CF) exceeding that of betting the base rate by more than about three and one-half percent, an amount which we consider trivial. This is particularly so when one considers the amount of effort expended to obtain the measure, and the importance of the decision about an individual's life which may in significant part depend on a recidivism prediction predicated on the score from this instrument.

In view of the fact that base rate predictions require no expert input, no time spent evaluating offenders, and deliver predictions about as accurate as elaborate actuarial schemes like the MnSOST—R across a wide range of clinically relevant recidivism rates, it would seem that base-rate predictions have much to recommend them in the field of sex offender recidivism prediction. Naturally, we are not so naïve as to expect that such predictions will find widespread adoption. After all jurists, like legislators and the public, are almost certainly a good deal more worried about false negative predictions than they are about false positives; and base-rate predictions make *only* false negative predictive errors.

Nevertheless, base-rate predictions do provide a rational, objective benchmark for the assessment of predictive accuracy of sex offender recidivism predictions. An elaborate, expensive actuarial scheme for predicting recidivism that cannot outperform the simple prediction of the modal outcome for each offender hardly inspires confidence.

To justify use of such a method as the MnSOST—R despite its inferior correct fraction (or essentially equivalent correct fraction, and very considerably greater expense), an appeal on some other basis must be made. The only rational basis that comes to mind is to an instrument's long run *cost-weighted error rate*, i.e., the expected disutility of its false-negative errors times that error rate, plus the expected disutility of its false-positive errors times that error rate—as compared to the expected disutility of false-negative errors for betting the base rate, times its false-negative error rate (viz., P). Evaluating disutilities requires a careful analysis of the tangible and intangible costs of false-negative and false-positive decision errors: costs of needless yet expensive civil commitments; costs to victims, families, and society of preventable new sex offenses; and so on.

The authors are unaware of a single serious attempt to undertake an analysis of this kind, let alone any concrete justification balancing competing error costs, such that a reasoned appraisal of disutilities favors the use of prediction instruments having accuracies either no better than betting the base rates, or at best minimally more accurate. Emotional appeals on the side of the (undeniably) great costs to victims, families, and societies are ersatz relative to a careful, balanced decision analysis.

Optimal cutting score placement depends in part on levels of competing disutilities. Cutting scores currently used in fieldwork are suboptimal, whether due to promulgation of fixed cutting scores without regard to local recidivism rates (as in Epperson et al. 2003), or to insufficiently large samples to obtain stable estimates of optimal cutting scores for local jurisdictions. (Reliable estimates of quantiles of score distributions can require *thousands* of cases per group.) When already misestimated cutting scores are credulously adopted without concern for disutility, it is safe to say that one simply does not have an inkling of a jurisdiction's cost outcome, even though these are data of the gravest import. One can expect, however, that jurisdictions will experience classification correct fractions considerably lower than AUC figures appearing in the literature portend, whether users realize it or not.

One reviewer claimed that comparisons of the CF of an instrument like the MnSOST—R to that of betting the base rate were irrelevant since for (unstated) ethical and legal reasons, it was not possible to adopt ''betting the base rate'' as a decision strategy. We would argue that there are, in fact, circumstances, where one could in effect adopt such a strategy, by acting so as to (in effect) always render modal predictions/decisions (for $P < .5$) in the sex offender commitment arena: an expert witness who concludes that it is not possible to carry out the prediction of ''likely'' sex offender recidivism with the requisite degree of accuracy, using admissible evidence, in accordance with the requirements of statute, can refuse to take on such cases on behalf of the State. After all, working on such cases is hardly compulsory.

We add here that regardless of whether one actually contemplates adoption of ''betting the base rates'' as a decision strategy, comparing the CF of an instrument to the CF of betting the base rate is, quite frankly, a not very demanding validity hurdle, beyond the hurdle cleared by establishing that the instrument is more valid than pure chance, which is all that establishing $A_z > .5$ by significance testing accomplishes. In fact, establishing that an instrument outperforms the base rates is a demonstration of incremental validity—a showing that the instrument tells the clinician something more than what readily available information, on hand prior to acquiring examinee-specific data, would have allowed one to predict. If an instrument will not allow one materially to outperform the base rate (the base rate is one kind of readily available pretest information), then one's test really has not much going for it.

Implications for Forensic Practice

Our analyses suggest that single actuarial instruments, with validities similar to the MnSOST—R and recidivism rates below .4, do not suffice for the sex offense recidivism risk prediction problem. Unfortunately, there is currently no known way to improve on predictions made with single actuarial instruments:

(a) clinical prediction, including clinically-mediated actuarial outcomes, is expected to perform poorer than ‘pure’ actuarial prediction; and
(b) combining, via statistically optimal procedures like logistic regression, measures like the MnSOST—R with other instruments (e.g., Static-99) has thus far not been demonstrated to increase accuracy beyond that obtained with only one instrument (Seto 2005).
(c) instruments with accuracies comparable to MnSOST—R employed in populations with base rates similar to that investigated in our correct fraction simulation (i.e., below .4) will add only negative to negligible incremental validity to the base rate strategy;

Irrespective of the nascent state of the science, courts solicit risk predictions and practitioners must be aware of and acknowledge prediction instruments’ infelicities and report to the court *accurate* predictions, even if that means reporting that their predictions, no matter how the predictions are derived, function at chance-level, should (a) through (c) prove true.

In states like Minnesota, licensed psychologists are ethically mandated by the Board of Psychology to report ‘‘[A]ny reservations or qualifications concerning the validity or reliability of the conclusions formulated and recommendations made, taking into account...the limitations of scientific procedures...the impossibility of absolute predictions;...[and] a notation concerning any discrepancy, disagreement, or conflicting information regarding the circumstances of the case that may have a bearing on the psychologist’s conclusions...’’ (Minnesota Board of Psychology 2005). In Minnesota, and we expect elsewhere as well, it is expressly an ethical as well as professional obligation that, should (a) through (c) prove true, psychologists either (1) refrain from making sex offense recidivism predictions, (2) express serious reservations about such predictions, or (3) expressly adopt an explicit regrets ratio that countervails low base rates and renders the risk instrument sufficiently valid.

Respondents to sexually violent predator civil commitment petitions are protected, like all those at trial, by substantive and procedural due process safeguards. Indeed, questions of due process were among the central issues of the Supreme Court’s decision in *Kansas v. Hendricks* (1997). The Court opined that substantive due process is not violated by civil commitment statutes as long as statutes require a finding of the respondent’s dangerousness *and* a finding of an inability to control that dangerousness. Such inability may simply be volitional impairment secondary to a mental abnormality or personality disturbance that renders the respondent dangerous beyond his/her control. Under (a) through (c) above, there is currently no way to differentiate between high-risk and low-risk offenders. Unless States with civil commitment laws are at the present time willing to commit without discretion any member of the entire class of sex offenders, and the average level of dangerousness of this entire class is sufficient for a particular State legislature’s purposes and intent, enforcement of the statutes cannot meet criteria set out in the *Hendricks* opinion, thus inculpating civil commitment statutes under *Hendricks* as in violation of the due process clause of the Fourteenth Amendment.

The Court further opined that civil commitment statutes were constitutional as long as procedural due process was observed, i.e., that proper procedures and evidentiary standards were upheld prior to civil commitment. Again, given (a) through (c) above, instruments like the MnSOST—R do not meet even the most basic evidentiary standard, that of relevance. When the MnSOST—R, and other risk prediction methods of comparable or worse accuracy, are used in jurisdictions with low base rates they offer negligible to negative incremental validity over what is already known about the likelihood of reoffense prior to obtaining the offenders’ test scores, that is, the tests are not useful in deciding any issue of fact; *they possess at most paltry probative value at that point in the fact-finding process*. Instruments performing like the MnSOST—R are irrelevant, and should not be admissible as evidence at trial. The MnSOST—R, and similar instruments, fail to meet the most basic test for admissibility in, for example, the Federal Rules of Evidence Rule 702 (1975).

Were tests such as the MnSOST—R probative in commitment trials, they may still be excludable under Federal Rule 403 (1975) because a tests’ probative value may be substantially outweighed by the danger of unfair prejudice. To take an example, if the MnSOST—R were used to predict future offending in a population with a base rate of ~14%, the average respondent at trial for commitment has an ~86% chance of being correctly identified by the test as either future recidivist or non-recidivist. The remaining ~14% of the time, the respondent is approximately three times more likely to be incorrectly identified a recidivist as compared to being incorrectly identified a non-recidivist. When the test errs, it errs systematically against the respondent, constituting a prejudiced outcome.

In *Hendricks*-like statutes, State legislatures have attempted to establish by fiat a ‘‘fact on the ground,’’ namely that courts possess, by acting with assistance of experts, the ability to detect, with a degree of accuracy sufficient for their purposes, the presence of a legally requisite degree of likelihood of committing a future offense, whatever may be the legislature’s intent as to the meaning of ‘‘likely’’ in that jurisdiction. The present analyses suggest that the state legislatures have, in effect, attempted to establish by statute the proof of an empirical non-fact: Namely that the MnSOST—R, and instruments with comparable AUCs, along with clinical, judicial, and/or lay intuition and judgment, when used in populations with base rates less than .4, can assist in the determination of likelihood of reoffense.

Acknowledgments The authors thank Greg Bal of the Iowa Public Defenders Office for facilitating obtaining the data set, and Eric Janus for useful critiques of related work. We are grateful to Steven Huot and most especially to Douglas Epperson for his patience in answering our questions, and for the latter's willingness to meet with us. Prof. Epperson is, of course, not responsible for any errors that may remain in this article.

References

Amenta, A. E., Guy, L. S., & Edens, J. F. (2003). Sex offender risk assessment: A cautionary note regarding measures attempting to quantify violence risk. *Journal of Forensic Psychology Practice, 3*, 39–50.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology, 12*, 387–415.

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

Bartosh, D. L., Garby, T., Lewis, D., & Gray, S. (2003). Differences in the predictive validity of actuarial risk assessments in relation to sex offender type. *International Journal of Offender Therapy and Comparative Criminology, 47*, 422–438.

Campbell, T. W. (2000). Sexual predator evaluations and phrenology: Considering issues of evidentiary reliability. *Behavioral Sciences and the Law, 18*, 111–130.

Campbell, T. W. (2003). Sex offenders and actuarial risk assessments: Ethical considerations. *Behavioral Sciences and the Law, 21*, 269–279.

Epperson, D. L., Kaul, J. D., Huot, S. J., Hesselton, D., Alexander, W., & Goldman, R. (2000, November). *Cross-validation of the Minnesota Sex Offender Screening Tool—Revised (MnSOST—R)*. Paper presented at the meeting of the Association for the Treatment of Sexual Abusers, San Diego, CA. Retrieved March 30, 2006 from http://www.psychology.iastate.edu/faculty/epperson/atsa2000/sld001.htm.

Epperson, D. L., Kaul, J. D., Huot, S., Goldman, R., & Alexander, W. (2003). *Minnesota Sex Offender Screening Tool—Revised (MnSOST—R) technical paper: Development, validation, and recommended risk level cut scores*. Retrieved November 18, 2006 from Iowa State University Department of Psychology web site: http://www.psychology.iastate.edu/~dle/mnsost_download.htm.

Federal Rules of Evidence Rule 702, Pub. L. No. 93-595, §1, 88 Stat. 1937 (1975).

Federal Rules of Evidence Rule 403, Pub. L. No. 93-595, §1, 88 Stat. 1932 (1975).

Grove, W. M., Zald, D. H., Hallberg, A. M., Lebow, B., Snitz, E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19–30.

Hanley, J. A., & McNeil, B. J. (1983a). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29–36.

Hanley, J. A., & McNeil, B. J. (1983b). A method for comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*, 839–843.

Hanson, R. K. (1997). *The development of a brief actuarial risk scale for sexual offense recidivism*. (User report 1997–04.). Ottawa: Department of the Solicitor General of Canada.

Hanson, R. K., & Morton-Bourgon, K. E. (2004). *Predictors of sexual recidivism: An updated meta-analysis*. (User report 2004–02.). Ottawa: Public Safety and Emergency Preparedness Canada.

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

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior, 20*, 315–335.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Janus, E. S. (2000). Sex predator commitment laws: Constitutional but unwise. *Psychiatric Annals, 30*, 411–420.

Janus, E. S., & Meehl, P. E. (1997). Assessing the legal standard for predictions of dangerousness in sex offender commitment proceedings. *Psychology, Public Policy, and the Law, 3*, 33–64.

Janus, E. S., & Prentky, R. A. (2003). Forensic use of actuarial risk assessment with sex offenders: Accuracy, admissibility and accountability. *American Criminal Law Review, 40*, 1443–1499.

Kansas v. Hendricks (1997). 117 S. Ct. 2072.

Langton, C. M. (2003). Contrasting approaches to risk assessment with adult male sexual offenders: An evaluation of recidivism prediction schemes and the utility of supplementary clinical information for enhancing predictive accuracy (Unpublished doctoral dissertation, Institute of Medical Science, University of Toronto, Toronto, 2003). *Dissertation Abstracts International*. 64(4-B), 1907. (UMI No. 2003-95020-071).

Langton, C. M., Barbaree, H. E., Seto, M. C., Peacock, E. J., Harkis, L., & Hansen, K. T. (in press). Actuarial assessment of risk for reoffence among adult sex offenders: Evaluating the predictive accuracy of the Static-2002 and five other instruments. *Criminal Justice and Behavior*.

Litwack, T. R. (2001). Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and the Law, 7*, 409–443.

Meehl, P. E. (1965). *Detecting latent clinical taxa by fallible quantitative indicators lacking an accepted criterion.* (Report No.-PR-65-2). Minneapolis: University of Minnesota, Research Laboratories of the Department of Psychiatry.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194–216.

Minnesota Board of Psychology. (2005). Minnesota Board of Psychology psychology practice act. Minneapolis: Minnesota Board of Psychology.

Minnesota Department of Corrections. (2000a). *Sex offender policy board and management study*. St. Paul: Author.

Minnesota Department of Corrections. (2000b). *Sex offender supervision: 2000 report to the Legislature*. St. Paul: Author.

Mossman, D. (1994a). Further comments on portraying the accuracy of violence prediction. *Law and Human Behavior, 18*, 587–593.

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

Nuffield, J. (1982). *Parole decision-making in Canada: Research towards decision guidelines*. Ottawa: Solicitor General of Canada Research Division.

Prentky, R. A., Lee, A. F. S., Knight, R. A., & Cerce, D. (1997). Recidivism rates among child molesters and rapists: A methodological analysis. *Law and Human Behavior, 21*, 635–658.

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

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63*, 737–748.

SAS Institute, Inc. (2005). *SAS language reference: Dictionary*, Version 9. Cary, NC: Author.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. New York: Wiley.

Seto, M. C. (2005). Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. *Psychological Assessment, 17*, 156–167.

Shao, J. (1996). Bootstrap model selection. *Journal of the American Statistical Association, 91*, 655–665.

SPSS, Inc. (2005). *SPSS 14.0 Base user's guide*. New York: Prentice-Hall.

Szmukler, G. (2001). Violence risk prediction in practice. *British Journal of Psychiatry, 178*, 84–85.

Venables, W. N., Smith, D. M., & the R Development Core Team. (2002). *An introduction to R*. Bristol, UK: Network Theory, Ltd.

Wollert, R. W. (2002). The importance of cross-validation in actuarial test construction: Shrinkage in the risk estimates for the Minnesota Sex Offender Screening Tool—Revised. *Journal of Threat Assessment, 2*, 87–102.

Wollert, R. W. (2003). Additional flaws in the Minnesota Sex Offender Screening Tool—Revised. *Journal of Threat Assessment, 2*, 65–78.