



Published in final edited form as:

Psychol Assess. 2010 June ; 22(2): 382–395. doi:10.1037/a0019228.

Multidimensional Assessment of Criminal Recidivism: Problems, Pitfalls, and Proposed Solutions

Scott I. Vrieze and

Department of Psychology, University of Minnesota

William M. Grove

Department of Psychology, University of Minnesota

Abstract

All states have statutes in place to commit civilly individuals at high risk for violence. This note addresses difficulties in assessing such risk, but uses as an example the task of predicting sexual violence recidivism; the principles espoused here generalize to predicting all violence. As part of commitment process, mental health professionals, who are often psychologists, evaluate an individual's risk of sexual recidivism. It is common for professionals conducting these risk assessments to use several actuarial risk prediction instruments (i.e., psychological tests). These tests rarely agree closely in the risk figures they provide. Serious epistemological and psychometric problems in the multivariate assessment of recidivism risk are pointed out. Sound psychometric, or in some cases heuristic, solutions to these problems are proffered, in hope of improving clinical practice. We focus on how to make these tests' outputs commensurable, and discuss various ways to combine them in coherent, justifiable, fashions.

Keywords

Recidivism; violence; multidimensional assessment; actuarial prediction; clinical prediction; clinical statistical prediction; Bayes theorem

Note on Multidimensional Assessment of Recidivism Risk

One of us (WMG) was recently asked to offer testimony regarding the validity of sex offender recidivism prediction instruments, in particular the Minnesota Sex Offender Screening Tool-Revised (MnSOST-R; Epperson, Kaul, Huot, Goldman, & Alexander, 2003). The district attorney of a Minnesota county had brought a petition seeking the civil commitment, under a sexually dangerous persons statute, of an individual, Mr. John Smith as he shall be called, the facts of which pertinent to this note were as follows. The county claimed that Mr. Smith had sexually abused a female relative in childhood. He was never charged with this. The state claimed that Mr. Smith had "abused" a fourteen-year-old girl when he was sixteen by having sexual intercourse with her; he claimed he was a teenager himself at the time and the state did not dispute this. Mr. Smith was charged with criminal sexual conduct in the third degree for statutory rape, with a stay of sentence. When Mr. Smith was 21, a woman reported that Mr. Smith raped her; this report was delayed until one month after the alleged incident. Mr. Smith passed a polygraph concerning this matter;

© Copyright 2010 all rights reserved

Correspondence concerning this article should be addressed to Scott I. Vrieze, Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis, MN 55455. vrie0006@umn.edu.

charges were never filed against him. A year later it came to light that Mr. Smith was having relations with a fifteen year old girl; he claimed she told him she was older. If true, this claim would have made the sex legal. He was never charged with any crime concerning this matter. Latterly, the county subsequently undertook to commit Mr. Smith indefinitely as a sexually dangerous person (Note 1).

In the course of reviewing a 125-page report offered by an expert witness for the state, whom we shall call Dr. Fisbee, one of us (WMG) found that no fewer than eight different instruments had been administered for the purpose of arriving at a recidivism risk prediction, to fulfill the statutory requirement that the offender be judged to be “likely” to reoffend if allowed to be at large. The instruments were the Psychopathy Checklist-Revised (PCL-R; Hare, 2003), Violence Risk Appraisal Guide and Sex Offender Risk Appraisal Guide (VRAG and SORAG; Quinsey, Harris, Rice, & Cormier, 1993), Minnesota Sex Offender Screening Tool-Revised (MnSOST-R; Epperson et al., 2003), Static-99 (Hanson & Thornton, 1999), Rapid Risk Assessment of Sexual Offense Recidivism (RRASOR; Hanson, 1997), Sexual Violence Risk-20 (SVR-20; Boer, Hart, Kropp, & Webster), and the Historical, Clinical, Risk Management-20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997). The raw scores and corresponding risk estimates are given in the accompanying Table 1, column’s 1,2, and 3, as they appeared in the expert’s report. Many instruments yielded more than one risk estimate. For example the PCL-R was made to yield estimates of general criminal and of violent recidivism. In the case of the MnSOST-R, different recidivism probability estimates were given according to different assumed population base rates of recidivism (none of which actually match the state’s own figures on Minnesota offenders’ recidivism rate). For several instruments (VRAG, SORAG, Static-99, and RRA-SOR), recidivism risk projections over two or more time periods were given. Sometimes these estimates were quite broadly stated: “medium to high,” and some had suspiciously-high claimed precision, e.g., 92.9%. Two instruments gave it out that it was an iron-clad certainty that Mr. Smith would reoffend. Mr. Smith’s raw scores, along with the Bayes’ posterior probability of recidivism, are listed in columns 2 and 3 of Table 1. We will discuss the other columns in Table 1 shortly.

It is instructive to determine how raters scored the MnSOST-R, among the validated risk assessment tools, by examining Mr. Smith’s background. Reviewers of an earlier version of this paper objected that, from the description above, that he should have high scores on MnSOST-R, VRAG, SORAG, etc. Review of Dr. Fisbee’s report suggested that some arrests, charges not leading to conviction, and/or police interactions (e.g., police called but no arrest made) may have been erroneously scored by Dr. Fisbee as separate convictions on the risk instruments. This level of discrepancy is not altogether surprising. Opposing examiners in sex offender civil commitments have quite low reliability in the .4 to .6 range (Murrie, Boccaccini, Turner, Meeks, Woods, & Tussey, 2009). Low reliabilities between raters impugns expected achievable validity for both raters (reliability puts a ceiling on validity). Hence, in addition to the questions the authors had about combining risk indicators, there are questions of the veracity of the risk instrument scores. It is important to carefully distinguish these goals, and in this note we address only issues in combining the risk indicators. We take Dr. Fisbee’s scores as given, and attempt to combine those scores into some comprehensible overall estimate of risk. However, evidence that the predictors cannot be coherently combined can suggest that the scores were made in error.

The balance of this note will ask some important and, alas, difficult satisfactorily to answer questions about the interpretation of multiple estimates of risk on the same individual, and offer some thoughts about problems introduced by the availability of multiple indicia of this kind on the same person. Obviously, there are potentially great advantages offered to the forensic examiner (as there are to clinicians in general) by the opportunity to “triangulate”

on an assessment issue, to cross-check an inference, to get mutual confirmation (and, even more important, potential disconfirmation; Popper, 1959) from independent sources of information in clinical assessment. However, inextricably linked with this potential also come thorny problems in epistemology, probability, statistics, as well as specific, as yet insufficiently-addressed research issues in the multivariate assessment of sex offender recidivism risk.

It is these problems that we wish to address in this note. We cannot hope to give an in-depth treatment of these issues in the space available to us here. Indeed, a number of the problems do not, as far as we are aware, have complete or entirely satisfactory solutions at the present time. However, much good can come of forensic examiners, and clinicians in general, becoming more focally aware of the various problems and their potential to lead to seriously erroneous clinical inferences. Further, we hope to stimulate research in the direction of solving some of the epistemological and statistical problems, as well as encouraging the conduct of clinically relevant research on multi-instrument, incremental validity in sex offender recidivism prediction.

The Problems

Problem 1: Implausible probability values

First, some measures yield implausible risk estimates, e.g., the VRAG and SORAG figures of 100% risk, meaning claimed certainty that an individual will reoffend. Now, in general it is technically possible for an actuarial predictive instrument to yield a legitimate output probability of 1.0, for some sets of input data with some prediction problems. For example, $\text{Pr}(\text{dies \& under water \& no breathing apparatus \& three days duration} = 1.0)$. However, it presumably needs no argument that, on presently available scientific knowledge, there is no set of dispositional measurements on a person that would make it absolutely certain that he or she will (or will not, for that matter) commit a sex offense within a given time period. The most strongly driven individual, lacking even rudimentary impulse control, may be prevented from commission by a rare but sudden, complete, and permanent physical incapacity from acting on their impulses. Rare events do occur in forensic practice, even if they were not observed in the VRAG or SORAG probability calibration samples. This is why actuarial instruments for this prediction problem should avoid delivering zero or one probabilities; instead, they should deliver a substitute value for zero or one observed in the training sample, such as $(1/6)/(N + 1/3)$ for zero and $(N - 1/6)/(N + 1/3)$ for 1.0 (what are called “started fractions” in exploratory data analysis; Tukey, 1977).

This fairly obvious fact requires further enunciation because a predicted recidivism probability of 1.0 (or zero) also has *ipso facto* a claimed standard error of zero. If combined with other instruments’ predicted probabilities, not equal to one, employing optimal least-squares weights (i.e., which tend to be proportional to the inverses of standard errors), this would rob the other instruments’ scores of all weight. Combining probability estimates is discussed in more detail below.

Tentative Solution to 1

Estimates of probability equaling zero or one in a calibration study can be repaired by going back to the original study data and proceeding as follows. (The required data is provided for the VRAG only in a crude line graph on page 148 of Quinsey, Harris, Rice, & Cormier [1998]; it is not provided at all for the SORAG.) Of the 618 subjects in the original VRAG study, approximately seven (judging by visual inspection of the graph) scored 28 or higher, and all of these recidivated by the end of 7 years follow-up. However, as stated above, one does not believe that the population recidivism rate for individuals scoring ≥ 28 is actually 100%, on common sense grounds. The statistical literature offers the following advice, and

the total number of individuals n_x scoring within a given test score interval ($x_i, \dots, x_j, i < j$), to yield an estimated probability of recidivism:

$$p_r = \frac{n_r}{n_x} \quad (1)$$

one uses started counts

$$p_r = \frac{n_r + 1/6}{n_x + 1/3} \quad (2)$$

(Tukey, 1977). The resulting estimated probabilities p_r will never equal zero or one, but as $n_{X=x}$ becomes large p_r can closely approach these bounds.

Problem 2: What is the criterion?

Not all instruments were developed with the same criterion in mind. The VRAG, SORAG, and PCL-R were developed on violent offenders (which included sex offenders), and created to predict violent recidivism (which included sexual recidivism). The post-test probabilities listed in Quinsey, Harris, Rice, and Cormier (1998), and used by Dr. Fisbee in estimating Mr. Smith's likelihood of recidivism, are in fact probabilities of Mr. Smith committing *any* violent offense, including sex offenses. It goes without saying that the probability of committing another violent sex offense is lower than committing just any violent offense (i.e., the set of sexually violent crimes is a subset to the set of violent crimes), and Dr. Fisbee has without doubt erroneously inflated these posterior probabilities of sexual reoffense.

This point begs discussion of a larger issue. There is a general problem in the field with defining "recidivism." A number of considerations impinge on one's opinion about which kind of estimate to use, particularly important among which is the main purpose to which recidivism predictions are to be put: are they to predict outcomes until a statutorily mandated judicial review (e.g., six months or one year in many jurisdictions)? Until the expected end of a finite period of sex offender treatment (e.g., three years, as in the original Minnesota plan)? Or are they to predict outcomes for a lifetime, because commitment may in fact last this long? The Static-99 developers recognize this issue, and give probabilities of reoffense at 5 years, 10 years, and 15 years. While helpful, this does not relieve the examiner (or court) from selecting one particular time frame. Of course, all other tests must also deliver predictions for the same time frame (which they likely will not, see Table 1), otherwise the estimates are incommensurable.

Perhaps even more important is whether the recidivism rate should be based on reconviction and rearrest data, or if rates based on these events are underestimates that need adjustment upward given data on rates of un-reported (or unclosed) sex crimes. Some argue with considerable cogency that rearrest rates are underestimates of sex crimes (e.g., Doren, 1998). However, this does us little good from a predictive standpoint. This is because unless we can estimate with reasonable precision the actual number of crimes committed for which released offenders are never arrested or convicted, we have no way of knowing by what factor our recidivism rate (based on rearrests and reconvictions) should be multiplied. A subjective estimate of the multiplying factor would not be worthy of the adjective "scientific".

To find the "true" rate of recidivism some studies use re-arrest data, some reconviction. Some studies go so far as to use offender self-report, or they analyze child protection

records. Each method is tapping the hypothetical construct of “recidivism” in different ways, and each is prone to give different answers as to the true probability of recidivism.

One final point should be made here, suggested by a helpful reviewer of this article. In many validation studies offenders are evaluated upon leaving custody (e.g., prison), and followed for some period of time during which they may or may not recidivate. The test score obtained at the offender’s release is then validated against recidivism. It is therefore unknown based on existing validation studies how valid these tests are for predicting recidivism on community-dwelling sex offenders (e.g., a convicted sex offender who has served time, been released, and is now being evaluated pursuant to new charges). To be sure, an offender living offense free in the community for some significant period of time is less likely to commit a crime than an offender just released from prison, *ceteris paribus*. While less likely, it is unknown just how this community-dwelling variable would impact probability of recidivism.

Tentative Solution to 2

SVP evaluators can approach the problem of choosing a recidivism criterion in at least two ways. (1) They can refrain from using any data other than reconviction data. This tack has several advantages. a) As an estimate of recidivism it sensibly accounts for all civil rights (e.g., due process), and provides for legal safeguards against statements in court that are not valid. b) For good or ill, the arbiter of justice in civilized society is the judge or jury. The use of reconviction data as the criterion does not tempt the SVP evaluator to act as a fact witness, when they are supposed to be opinion witnesses.

However, many in the sex offense literature have argued that reconviction data is not enough, and the real issue in civil procedures is the “true” risk of sexual recidivism (Doren, 1998; Rice, Harris, Lang, & Cormier, 2006). By “true risk” we simply mean the risk of committing some future crime of interest (e.g., hands-on sexual offense). We do not presume that the commission of a future criminal act is taxonomic, as “sex crime” is not operationally definable due to its infinite extensibility. (Rice, et al. 2006 give several interesting and perhaps counterintuitive examples of seemingly non-sexual crimes deemed sexually motivated by the court.)

More important is that the true base rate of sex offenses (however one has defined it) is impossible to measure exactly. Unknown numbers of already-defined sex offenses go unnoticed every day. Victims under-report sex crimes. Even if victims regularly report sex crimes many offenders will not be caught and so will not contribute to the estimated recidivism rate. Alternative measures suffer their own infidelities. Offender self-report can be subject to a high rate of false negatives. Victim reports do not distinguish between high-rate offenders (offending against multiple victims) and low-rate offenders.

No single method is the “gold standard,” and if psychologists want accurately to determine the rate of “true” recidivism in the population they should engage in construct validation of recidivism to do so (see Cronbach & Meehl, 1955, for a description of construct validation). Dealing with recidivism as a hypothetical construct, instead of the conventional approach of using an operational definition (virtually always a yes/no dichotomy obtained from official records), is a step forward in measuring recidivism risk. (Note that this is not a criticism of the risk instruments *per se*, but in how they are have traditionally been validated.)

One such step forward would be a reconceptualization of recidivism as a many-valued variable (instead of the yes/no dichotomy). Continuity in recidivism estimates can be thought of as the probability that the psychological assessment yields an approximately accurate estimate of the proclivity for reoffending. For example, evidence at trial almost

never determines guilt or innocence with 1.0 probability. Rather, judges and juries determine whether the individual is criminally responsible based on a *degree* of confidence in guilt or innocence. In criminal court this degree of confidence must exceed the ill-defined criterion “beyond a reasonable doubt;” in civil court it ranges from “preponderance of evidence” to “clear and convincing evidence.” In any event, we as a field need not truncate the underlying degree of accuracy of the probability of recidivism, into a binary measurement of reconviction/non-reconviction. Instead, rating scales and assessment instruments can be used (as they were in Rice, et al. [2006]) more accurately to measure recidivism, and prediction instruments like the Static-99 can be validated against this quasi-continuous measure.

Even assuming that recidivism is indeed a categorical construct (the offender either committed a criminal act or they did not, or were reconvicted or not) the advantages of quasi-continuous measures of truly binary constructs are substantial. Grove (1991) has analytically shown that accuracy can be improved by measuring latent dichotomous variables with continuous metrics. Specifically, in the parameter space likely to manifest in clinical psychology one obtains greater predictive accuracy by quantifying dichotomous variables as continuous (even when they truly are dichotomous), rather than losing vital information by reducing measurement to yes/no. Again, continuous measurements can be thought of as measures of probability, where high scores indicate strong confidence of a prediction of recidivism and low scores the opposite.

This is borne out in recidivism research. Rice, et al. (2005) found that a polychotomous (4-level) item reflecting certainty of sexual recidivism performed better as a measure of recidivism than any single dichotomous criterion such as yes/no reconviction or yes/no reoffense. The VRAG more accurately predicted the polychotomous item than any single dichotomous criterion, suggesting that the hypothetical construct measured by the VRAG is something more than simple reconviction or rearrest. This example of convergent validity between the VRAG, reconviction data, and Rice, et al.’s (2006) polychotomous recidivism measure is an important first step in construct validation of an improved recidivism measure. Research in this area should focus on synthesizing different measures of recidivism, and investigate convergent and discriminant validity properties of the measures.

Ultimately, one might expect advances in risk measurement to mirror those observed in intelligence measurement. Originally, IQ testing was validated against clear measures of intelligence, such as education level, occupation status, etc., all of which were known to be fallible measures of intellectual prowess. Similarly, recidivism risk instruments are validated against official recidivism records, a known to be fallible source of criminal information. It would be unsurprising if a bootstraps procedure takes place in risk estimation, where the risk instruments ultimately supersede criteria, and become better measures of criminal behavior than any official re-arrest or reconviction record.

Problem 3: Near-uselessness of narrative reports of recidivism risk

Table 1 contains vague probability statements, e.g., “medium to high” for some predictors. Since this result is not quantified, comparing it to numerical values is difficult, and directly combining it with other numerical values impossible.

Tentative Solution to 3

We know of no entirely satisfactory solution to the second problem, which was an instrument delivering adjectives instead of numerical risk estimates. We recommend the redevelopment of the scale with tabulation of recidivism probabilities against overall scale scores.

Proponents of the ostensibly “hybrid” (Note 2) (actuarial + clinical judgment) SVR-20 may cite high area under the curve (AUC) of the receiver operating characteristic (ROC) as evidence that these “hybrids” are as accurate as actuarials (see Swetts & Pickett, 1962, for a description of the ROC and AUC; AUC is a popular measure ranging 0–1 and quantifies success in predicting a dichotomous criterion). One cannot compute an AUC under ROC without the instrument delivering numbers, a curious requirement for an instrument boasting non-quantitative risk estimates. Whatever AUC is found, that value’s accuracy can be expected only to drop in the conversion from numbers to words.

Developers of the SVR-20 may also argue that recidivism risk is not something that can be quantified by current methods, as “risk” includes not only future reconviction (or whatever criterion is of interest), but also includes severity and imminence of the future crime, with more severe and more imminent crimes resulting in increased risk. However, such risk locutions as “medium risk” do not accomplish this either. It is confused and confusing to suggest that one cannot quantify severity or imminence, and cannot then combine probabilities of future crime to yield an expected utility index. It of course can be defensibly done with a utility analysis, after eliciting utilities from stakeholders regarding imminence and severity, for example.

Problem 4: Base Rates and Cutting Scores

A fourth problem is that one instrument, the MnSOST-R, provides multiple predictions depending on the base rate of recidivism (however defined in the user manual for the scale). This correctly acknowledges that Bayes posterior probability of recidivism depends on the prior probability (base rate) of recidivism. However, there are three problems neglected by the multiple posterior probabilities delivered here, the last of which directly affects comparisons between, and combination of, the various recidivism probabilities across instruments.

1. None of the MnSOST-R prior probabilities used here, as a basis for calculating recidivism probabilities, actually correspond to the local jurisdiction’s (Minnesota’s) reported sex offender recidivism base rate ($P = .18$ in a 6.5 year follow-up; Minnesota Department of Corrections, 2000).
2. The optimal cutoff score for achieving maximally accurate classifications (“hitmax” cut) changes for each possible base rate (Meehl & Rosen, 1955); however, the same cutoff score, namely 13 (recommended by MnSOST-R developers Epperson, et al., 2003), was used by Dr. Fisbee for each base rate P in question. Indeed, the cutoff score Dr. Fisbee used $X_c = 13$ is not optimal for maximizing predictive accuracy for any base rate in the table (or for the 31.6% sample recidivism rate in the original MnSOST-R development study either).
3. None of the other instruments in Table 1 have had their cutting scores adjusted from whatever base rate held in calibration studies, to scores appropriate to the $BR = .18$ (over 6-7 years) of Minnesota offenders.

Note that there are two main issues here, not just one. The first issue was taken up in Problem 2 above; i.e., which measure of recidivism is the most appropriate for use in Sexually Violent Predator (SVP) evaluations.

A second and crucial issue is that whatever recidivism rate is considered appropriate for prediction purposes, it should be precisely and expressly established and consistently applied throughout the prediction process: in setting cutting scores, in calculating posterior probabilities, and with all prediction instruments. Dr. Fisbee failed to do this. As a result, the obtained probabilities are very likely incorrect and, if correct, are only so by mere luck.

In particular, if score points (or score ranges, e.g., scores at or above a certain point) are being used to generate recidivism probabilities), then a single table from a calibration study which has been based on, say, $P = .35$ (as in the MnSOST-R middle table entry), must have its entry adjusted for the known or postulated recidivism rate in the target population (e.g., .18 in Minnesota). Crucially for multidimensional assessment, a corresponding adjustment must be made to each instrument's predicted probabilities, so that each prediction scale is working off the same base rate. Otherwise, the estimated recidivism probabilities are incommensurable, and one cannot meaningfully judge whether they are mutually confirmatory or disconfirmatory.

Tentative Solution to 4

Getting all instrument-based recidivism probabilities on the same basis (tied to the same base rates, however those base rates are defined) is central to making them comparable and hence combinable into a single recidivism prediction. Imagine the target population (i.e., the narrowest population to which Mr. Smith belongs) has a recidivism rate of $P_{\text{prior}} = .4$. Then the following simple manipulation of Bayes theorem makes the appropriate adjustment. Begin with Dr. Fisbee's estimated .88 posterior probability of recidivism for Mr. Smith, taken from Table 1 (under the .35 base rate row). This is equal to:

$$O_{\text{post}} = \frac{p_{\text{post}}}{1 - p_{\text{post}}} = \frac{.88}{1 - .88}.$$

Now we transform to remove the effect of the MnSOST-R's development sample base rate P_{prior} , and obtain the base-rate-free likelihood ratio Ω :

$$O_{\text{post}} = \frac{P_{\text{prior}}}{Q_{\text{prior}}} \Omega = \frac{.88}{1 - .88} = \frac{.35}{1 - .35} \Omega$$

where

$$\Omega = \frac{286}{21}$$

Here, P_{prior} is the prior probability of recidivism in the calibration study for the recidivism prediction instrument, in this instance equal to .35, and $Q_{\text{prior}} = 1 - P_{\text{prior}}$. Next multiply Ω by the applicable prior odds ratio, here $.4/(1 - .4)$, obtaining the new posterior odds

$$O_{\text{post}} = \frac{P_{\text{prior}}}{Q_{\text{prior}}} \Omega = \frac{.4}{1 - .4} \frac{286}{21} = \frac{114.4}{12.6}$$

Finally, transform the odds ratio into a probability:

$$p_{\text{post}} = \frac{O_{\text{post}}}{O_{\text{post}} + 1} = \frac{114.4/12.6}{(114.4/12.6) + 1} = .90$$

A point estimate of p_{post} is welcome, but difficult to interpret without knowledge of its probability of error. We need a confidence (or predictive) interval on this quantity, for each measure being used. Then comparison of these intervals across measures for a given

individual, and taking into account the correlation between risk probabilities delivered by these measures, will tell us whether the measures are producing coherent estimates or one or more pairs of measures are disconfirming each other by undercutting. We take this issue up under the Tentative Solutions to 6,7,8,9.

It is deplorable but common practice to eschew reporting of such calculations (e.g., Boer, Hart, Kropp, & Webster, 1997). Hanson and Thornton (1999) were conscientious enough to provide frequency tables for the RRASOR and Static-99, from which these calculations can be made by the sophisticated reader; the same cannot be said for the creators of the HCR-20, SVR-20, VRAG, PCL-R, or SORAG. As far as we know, score frequency tables are not available for sex offenders and sexual recidivism for the VRAG, SORAG, or PCL-R; nowhere in the literature could we find this information for the SVR-20 or HCR-20 either (including the SVR-20 manual).

Problem 5: What is the optimum cutting score?

The MnSOST-R scoring procedure involves application of a cutting score which classifies examinees as either future recidivists or non-recidivists, whereas the VRAG, SORAG, RRASOR, Static-99 all directly deliver posterior probabilities of recidivism. The cutting score method can be used directly to maximize classification accuracy, whereas the other methods cannot. The cutting score method has a long history in medical and psychopathological tests, where the clinic is interested in yes/no disease status, and also interested in maximizing the accuracy of yes/no diagnoses (i.e., predictions of present disease state).

The purpose of SVP evaluations is somewhat different, and it may well be that the court is interested in the best estimate of the offender's likelihood to recidivate (and not interested in classification accuracy alone). To be sure, if posterior probabilities are reported without concern for cutting score placement, and judges are responsible for determining if posterior probability of recidivism is "high enough" to warrant commitment, then the jurisdiction will never learn its classification accuracy. In other words, they will never know the error rate of SVP predictions, which incidentally is one of the seven Daubert criteria for admissibility of evidence (*Daubert v. Merrill Dow Pharmaceuticals*, 1996; Grove & Barden, 1999). The effect of these scoring methods is to allow the judge, who in most cases does not have psychometric or statistical expertise (otherwise they would not have requested an expert witness) to place the test cutting score because they decide what risk level constitutes high risk. This will differ from judge to judge and result in suboptimal classification accuracy. This is acceptable if the judge is aware of cutting score impact on long-term accuracy, and also aware that methods exist whereby expected accuracy (or expected utility) can be maximized. Judges and other legal professionals are not psychometricians, and cannot be expected to have specialized measurement training.

Tentative Solution to 5

Setting a cutting score for a measure, to ensure maximally accurate predictions using that measure, is in principle not difficult. For example, one can calculate the hit rate with the cutting score set at every value for the predictor variable that occurs in the data. Next, select the cutting score with the desired hit rate. We expand on these issues in our solutions to the next few problems.

Thusfar, we have mostly ignored an important aspect of base rates, cutting scores, and classification accuracy. These problems are confounded by utility analysis. One might argue that the primary purpose of SVP evaluations is to reduce the societal burden of sex offenders and sexual offenses. If true, then society (nor, presumably, the court) is interested in

classification accuracy, but rather in minimizing the cost of sexual offenses. While we do not have space to address this issue in detail here, we will highlight the primary considerations. In a utility approach one considers the relative costs of prediction errors. The results of false positive errors (punishing a offender who would not recidivate if free) may be less costly than the results of false negative errors (freeing offenders who will recidivate if free). Hence, if one wants to limit costs, one would be more willing to make false positive errors (less costly) than false negative errors (more costly). Costs act just as base rates do, and cutting scores are set so as to maximize cost-weighted prediction errors. If societal burden is the primary purpose to which violence prediction is put, then a prime area of future research is in eliciting cost information from stakeholders, and incorporating that information into the prediction process.

Problem 6: How should we combine scales or items to arrive at an overall recidivism probability?

The procedure used by Dr. Fisbee appears to have been: perform, for each instrument, a comparison of the raw score to a per-instrument cutoff score, so that a certain score range for each instrument is considered “high risk.” Having done this for all instruments, across instruments Dr. Fisbee seems to have then tallied a box score, in effect, counting how many instruments’ scores fall in the high risk range. If most (or, preferably, all) scores fall in the high risk range, this is considered to be a consistent result and supports an inference that the individual is at high risk of recidivism, and further creates justified high confidence in that inference.

This is reasoning of precisely the same type that interprets data from related studies in a research domain as favoring a substantive theory by tallying number of null hypothesis significance tests (NHSTs) found significant (e.g., at $p < .05$) and, if “enough” go the right way, one is encouraged about the theory’s verisimilitude. This is well known to be bad scientific thinking (Meehl, 1978). In fact, according to a neo-Popperian philosophy of science, even one well conducted study result that goes the “wrong way,” i.e., opposite the others, is bad news for the theory, if all are supposed to be testing the same hypothesis. Meehl (1990) acknowledged, of course, some situations in which it is rational to set aside a disconfirmatory finding, but these require quite special justification (e.g., auxiliary theory regarding instrumentation was false; substantive theory had repeatedly survived experiments designed to put theory at strong risk of refutation [if false], so it had a great deal of “money in the bank”); appeal to NHST vote-counting is *not* such a justification.

In the recidivism prediction problem, all the delivered numbers are supposed to provide point estimates of (essentially) the same quantity—i.e., the probability of one or more new sex offenses being committed over the same or comparable time periods. Hence, the estimates should all agree (within psychometric error) across the various instruments, once adjusted to measure the same phenomenon (common base rate, common time period). That the agreement should not merely lie within population sampling error is easily seen because there is no sample; there is only one individual, namely Mr. Smith. To the extent that one obtains risk estimates for an offender that differ more than can be attributed to psychometric error (32.7 to 100 percent!), the estimates actually *undercut* each other as knowledge claims.

To grasp how the various recidivism probabilities can be used as consistency tests for each other, we turn to a favorite illustrative example of Imre Lakatos (and Paul Meehl). This concerns the calculation of Avogadro’s number (originally Lohschmidt’s number) serving as evidence for the real existence of atoms. In his classic book *Atoms*, Perrin (1913) gave eight qualitatively different methods to calculate the number of molecules in a mole, ranging from the fact that the sky is blue to a simple undergraduate method involving a pipette and Petri dish. Amazingly, each method gave approximately the same number (6×10^{23}) and, to

paraphrase the philosopher of science Wesley Salmon (1984), this either points us to a natural principle or a “damned strange coincidence.” In this case, Perrin, who had started out as a believer in the theory that atoms were convenient fictions, “converted” to the school of thought that atoms were real. Otherwise, there was no accounting for the fact that so many Avogadro’s number estimates closely converged; if atoms *are not* real, then counts of atoms are not counting anything, and there is no reason for estimates thereof to be consistent. (For a more sophisticated discussion of numerical consistency via epistemically different paths, see Meehl [1978, 1979] on consistency tests, Salmon [1984] on “damned strange coincidences”). The connection to disconfirmation: when eight *qualitatively similar* methods give notably discrepant results, a logical conclusion is that we are either measuring different constructs (e.g., one measure is of atoms, the other is of elementary particles), or that some serious systematic error is entering into at least some of our measurements, and perhaps different errors beleaguer different measures.

Instead of commenting directly, we first present three more related, but different, problems.

Problem 7: Predictor overlap

Another problem is that the delivered numbers for recidivism risks are not independent in several different ways, the impact of which is hard to assess. First, the item content of the recidivism prediction instruments overlaps significantly, indeed often quite considerably: e.g., the SORAG and VRAG contain the PCL-R; the Static-99 contains all questions from the RRASOR. Of course, empirical correlation is not a strict function of content overlap, even though *ceteris paribus* higher content overlap will usually lead to higher correlation. Only the remaining, nonartifactual covariation can be parsed into confirmatory versus disconfirmatory covariation.

This final problem is perhaps the most difficult to examine, as correlations between scales have not been widely reported. Multivariate correlations (partial correlations) are reported only in a few select studies (e.g., Seto, 2005). Knowledge of the scales’ interrelationships is vital to understanding their incremental validity, and determining the effectiveness of the approach by Dr. Fisbee. Let X represent the Static-99, Y the outcome (recidivism or non-recidivism), and let Z be some other measure. We administer Z to Mr. Smith and scored it. Now we wish to know if X will give us any further information about Mr. Smith; that is, if it will *add* any accuracy to our final prediction, over and above the accuracy obtained with Z alone. To determine the unique, incremental effect of X , we can compute the partial correlation between X and Y given that we know Z . Formally, this partial correlation is represented by $r_{X,Y|Z}$. Fortunately for our purpose in this note, one can easily find the partial correlation between recidivism and any measure x , given $n - 1$ other variables, and thereby obtain an estimate of recidivism risk using *all* instruments. All necessary information is contained in the $n \times n$ covariance matrix.

Problem 8: Epistemological Considerations in Predictor combination

Before one can arrive at a statistical formulation of the problem involved in combining possibly discrepant probability estimates, one must recognize that there is a significant epistemological problem. A fundamental principle in inductive logic is the Total Relevant Evidence (TRE) requirement, generally formulated as follows: in drawing conclusions about a matter of fact, one is required to base conclusions on all evidence that is probabilistically relevant to the conclusion. “Relevant” means here that the likelihood ratio (symbolized here by Ω_X), conditioned on all other known facts with regard to future recidivism over some defined time interval, does not equal 1.0. (The likelihood ratio is simply the probability among recidivists of showing a particular pattern of instrument scores, divided by the probability among nonrecidivists of showing precisely the same pattern. A conditional

likelihood ratio is symbolized $\Omega_{X|Y,Z}$. “Total” means that all facts for which the conditional likelihood ratio is other than 1.0 must be considered. In practice, one confines attention to facts for which one knows the conditional likelihood ratio to be *materially* different from 1.0 (as measuring weakly contributory facts is not worth the trouble). In the present context, it is unknown how much *relevant* evidence each instrument provides; given the substantial content overlap and empirically-observed score correlations one would expect some instruments to be at least partially redundant, hence perhaps providing little if any *materially relevant* evidence to recidivism prediction (see Nunes, Firestone, Bradford, Greenberg, & Broom, 2002; Seto, 2005). To the extent that redundancy is high, supernumerary scales may be given undue weight at the expense of scales that have appreciable incremental validity.

The epistemological dilemma is as follows: the several risk prediction scales yield several recidivism probabilities. Even if they were to yield probabilities calibrated to the same prior probability of recidivism, and for precisely the same follow-up interval, it is too much to expect that they will always agree. Indeed, the present case demonstrates striking disagreement of values (risk probabilities ranging from 0.327 to 1.0, an over 3:1 range). Taken at face value as claims to knowledge of the probability of reoffense, all instruments except VRAG and SORAG contradict each other, and so each forms a *prima facie* reason to disbelieve the others; they undercut each others’ statuses as knowledge claims.

However, no psychologist in their right mind takes such point estimates as these as having accuracy to the tenth percentage point or even to the nearest percent. Barring complicating factors like unknown redundancies between instruments, the solution to the epistemological problem is to construe the various risk estimate *confidence intervals*, not the point estimates themselves, as undercutting defeaters (as such things are known in nonmonotonic reasoning; Frankish, 2005) when inconsistent, and so prevent the forensic psychologist from delivering a risk estimate using those particular measures. When the predictive intervals all overlap, they are not defeaters after all, but instead deliver consistent (i.e., reinforcing) estimators of risk that may even serve to narrow the width of the estimated risk interval.

The uncertainty in a recidivism instrument’s probability prediction is a function of three sources: (a) uncertainty in the recidivism rate estimate (stemming from sampling error in the population prior probability estimate), (b) uncertainty across time and observers in instrument score (psychometric error), and (c) uncertainty in the statistical model (model error relating to item selection, item weighting, and cutting score placement, leading to uncertainty in the likelihood ratio—the conditional probability of a particular instrument score, given that an individual is a recidivist or not). We present a formula below that deals with (a) and (b). However, (c) is a more difficult problem calling for approaches such as Bayesian model averaging (Hoeting, Madigan, Raftery & Volinsky, 1999) or model selection procedures (Shao, 1997) such as consistent model selection bootstrapping (Shao, 1996; Vrieze & Grove, 2008) that go beyond the scope of this article.

Tentative Solutions to 6,7,8

These three problems are heavily intertwined, as are their solutions, thus we present the rest of this article in an attempt (tentatively) to solve problems 6-8, which are the thorniest problems with the least satisfying solutions.

There are two ways to combine multiple instrument scores: clinical judgment and mechanical (statistical, actuarial) prediction. Grove, Zald, Lebow, Snitz, & Nelson (2000) meta-analyzed 617 effect sizes from 137 studies of comparative validity. These studies concerned the prediction of health and human behavior. A main result obtained was that when hit rates are used as a criterion validity index (used for some analyses because it is easy to understand), actuarial data combination was about 12% more accurate than clinical

judgment, on average. We therefore strongly advocate the conduct of actuarial prediction studies that will yield optimal predictor weights; and in these same studies the calculation of optimal cutting scores for various base rates.

Some of the risk instruments used by Dr. Fisbee were developed by different researchers, and scoring guidelines change depending on the developers. The VRAG, SORAG, RRASOR, and Static-99 developers (at least implicitly) suggest that evaluators obtain a raw score for their client. Then, by comparing this raw score to a development sample of offenders with similar raw scores, the evaluator assigns the observed recidivism rate for those offenders to the present client. In this way, one obtains a probability of recidivism for clients with that test score.

On the other hand, some test developers (notably developers of the MnSOST-R) suggest that the evaluator use a test cutting score, and predict “recidivism” for clients who score above the cutting score. One can also obtain a probability of recidivism for these clients by calculating the recidivism rate of those scoring above the cutting score.

There are two different ways of calculating posterior probabilities of recidivism depending on how the test is scored. We first illustrate the former method by constructing an interval for Mr. Smith’s RRASOR score of 4, which under a Minnesota base rate of .18 translates according to Hanson (1997) to 35% probability, and $O_{\text{post}} = .34/(1 - .34) = .52$ odds, of recidivism (for five year follow-up). Plugging this value into a convenient closed-form method of confidence interval estimation provided by Newcombe (1998),

$$\text{c.i.} = \frac{2np + z^2 \pm 1 \pm z \sqrt{z^2 \pm 2 - \frac{1}{n} + 4p(np \mp 1)}}{2(n + z^2)} \quad (3)$$

where n is the number of individuals with the score in question (139 in this case [Hanson, 1997]), p is the proportion of recidivists with that score (here $p = .34$), $q = 1 - p$, z is the appropriate quantile from the normal distribution (e.g., 1.96), and note that $\pm \neq \mp$; for the upper bound use summation on the first three “ \pm ” and subtraction on the \mp and vice versa for the lower bound. For the RRASOR score of 4 we obtain the following 95% confidence interval [.26, .43]. Mr. Smith’s score of 8 on the Static-99 yields a posterior probability of .39 with 95% c.i. [.31, .48], using data from Hanson and Thornton (1999).

Calculating c.i.’s for the VRAG and SORAG is more difficult because, as noted above, the normative data for the VRAG and SORAG consisted of both sexually violent offenders and merely violent offenders. Additionally, the prediction criterion consisted of any violent recidivism (instead of just sexual recidivism) and the normative data is therefore not wholly appropriate for the sexual recidivism question. To bypass these problems while remaining true to what we see as Dr. Fisbee’s intent, we have applied the VRAG and SORAG “bins” (see Quinsey, Harris, Rice, and Cormier, 1998) to 390 sexual offenders from the Massachusetts Bridgewater hospital dataset described extensively in Knight and Thornton (2007). The subjects were sex offenders, and the outcome was sexual recidivism at five years, which is the same as with the RRASOR and Static-99 above, and comparable to the 6.5-year Minnesota follow up (MN Department of Corrections, 2006). Mr. Smith’s VRAG score of 28 falls in bin 9, which contained only two offenders and a recidivism rate of 50% (which translates to 33% in MN with its base rate of .18). Our sample was not small, and the fact that bin 9 contained only two offenders casts doubt on the VRAG bin 9’s usefulness in predicting sexual recidivism. The 95% confidence interval for this probability was [.003, .93], which provides almost no information about Mr. Smith’s risk. The SORAG performed somewhat better, yielding a probability of recidivism of .38, and an interval [.17, .63].

In Mr. Smith's case, these obtained posterior test probabilities seem to cohere; each of the above 95% confidence intervals overlap, and thus reinforce each other as risk estimates. It appears at first blush as though the Static-99 and the RRASOR results are mutually reinforcing. However, we might expect such strong coherence even before calculating their correlation matrix because the Static-99 includes each of the RRASOR's items, and the scales consistently correlate highly (.74 in the Massachusetts sample). Of course, reliability imposes a ceiling on validity, and this .8 correlation is about as high as possible given the imperfect reliabilities of the Static-99 and RRASOR. That is, for all intents and purposes, the Static-99 and the RRASOR are redundant, and any discrepancy between their 95% confidence intervals should be taken as strong evidence that one, the other, or both, the Static-99 and RRASOR measurements are seriously flawed and should not be used in the instant examinations. The same can be said for the VRAG, SORAG, and PCL-R, which also share items (i.e., PCL-R), and correlate at least as highly as the Static-99 and RRASOR ($r(\text{VRAG}, \text{SORAG}) = .93$, $r(\text{VRAG}, \text{PCL-R}) = .77$, $r(\text{SORAG}, \text{PCL-R}) = .78$ in the Massachusetts sample).

The preceding discussion only addresses sampling error of the test score distributions. It does not account for sampling error in the base rate itself. That is, in all c.i. intervals we have here assumed the base rate was known to be .18, when in fact this is only an estimate and is itself subject to variability. We address this shortcoming in what follows.

The MnSOST-R and certain other measures define recidivism as reconviction, over a certain follow-up interval. It will be highly useful to redevelop all measures to predict a common criterion. In many instances this is only a matter of reanalyzing data already on hand.

To employ the eight quantitative instruments listed in Table 1 as their developers intended one would use the VRAG and SORAG score bins, post-test probabilities associated with individual RRASOR and Static-99 scores, and a summary scale and cutting score for the MnSOST-R. These data combination methods will unfortunately always result in suboptimal long run classification accuracy of sex offenders, and more false-positive (predicting reoffense for those who truly would not) and false-negative (predicting that an offender will go straight when they truly will commit another crime) errors than need be. The reason? Suboptimally placed cutting scores.

To avoid these extra errors, one uses cutting scores which, when properly set according to the test score distributions, the local base rate, and any cost-benefit analysis one may be able to conduct, can serve to maximize the accuracy of recidivism predictions. The MnSOST-R developers were on the right track, but did not succeed in correctly placing their suggested cutting score of 13 (assuming their intent was to maximize predictive accuracy of the instrument). We illustrate in what follows how to set optimal cutting scores for each instrument individually. (We assume throughout this note, without evidence to the contrary, that the costs of false-positive and false-negative errors is equal.)

To carry out our first suggested solution to problem 7, a number of steps are necessary. Initially, the test score distributions need to be determined. The usual approach here involves score frequency tables (which is usually represented visually by histograms). Score frequency tables (and histograms) include necessary raw data, but are quick and dirty ways to represent test score distributions. Histograms, for example, throw away large amounts of information; each bin in a histogram is discrete, and pays no attention to other aspects of the histogram, including sizes and locations of adjacent bins.

Prediction instruments are often thought to measure an underlying continuum of risk, and instrument scores are discrete measures of this underlying continuum. Optimal cutting score

placement depends on the true instrument score distributions, and ragged histograms often shown in instrument development articles do not allow for precise cutting score placement.

The typical histogram is in fact a member of a larger class of methods called kernel density estimates. A histogram's kernel is a non-overlapping rectangle with unit height and width equal to the histogram bin width. The kernel need not be rectangular; other examples include triangular kernels and Gaussian-distributed kernels. The Gaussian kernel, for example, is centered at each score x and depending on the bandwidth σ of the chosen kernel distributes Gaussian weights to each datapoint across the range of x , with the most weight assigned at x itself. This method not only solves the continuity problem of histograms, but for every score x takes into consideration, for the density of x , not just the number of offenders who scored x , but also the scores of all other offenders under study. Readers interested in density estimation are referred to Scott (1992), Siminoff (1996), Silverman (1996), or Wand and Jones (1992). Individuals interested in employing density estimation in the R programming language are referred to Venables and Ripley (2002) and Crawley (2007), or the extensive online help available in the R help archive and message boards. R is a freely available software program. To download R, go to CRAN, the Comprehensive R Archive Network, at <http://cran.r-project.org>. A host of introductory materials are available on the website.

For every test other than the HCR-20 (for which we lacked adequate information) we computed kernel density estimates of score distributions for each the recidivists and non-recidivists. We chose to use an Epanechnikov kernel because of its relatively low mean integrated squared error (assuming correct kernel bandwidth; Scott, 1992). We chose to use a bandwidth selector proposed by Sheather and Jones (1991), due to its general popularity among statisticians (Venables & Ripley, 2002), fast rate of convergence, as well as its typically visually appealing results. Specifically, we used the "SJ-dpi" bandwidth in the R environment's *density()* function. This bandwidth selector broke down for the RRASOR and Static-99 and returns densities with as many modes as there are score values, due to the relatively small number of possible scores (the development report gives RRASOR scores ranging from 0 to 5, and Static-99 scores ranging from 0 to 6). We thus multiplied the obtained Sheather and Jones bandwidth by small integer values (less than 6) until we obtained a unimodal and adequately smoothed density estimate.

We further smoothed the obtained densities with cubic spline interpolations in order to have tens of thousands points of support (for all intents and purposes a continuum). Next we multiplied each smoothed density by its appropriate base rate of recidivism ($P = .18$ and $Q = 1 - P$, per the Minnesota Department of Corrections [2000] study), and found the point where these two relativized density distributions cross, which is the cutting score that maximizes classification accuracy (hitmax cut). All offenders who score above the hitmax cut have $> .5$ probability of reoffense; all who score below it have $< .5$ probability of reoffense. (The cutting score can be moved so that offenders predicted to recidivate have $> .75$ probability of recidivism, but is then no longer the cutting score that maximizes classification accuracy under our equal cost scenario.) In addition, one can compute the odds of recidivism for an offender with score x on the instrument by simply dividing $f_r(x_r)$ for recidivists by $f_{nr}(x_{nr})$ for non-recidivists. This method is superior to that used with the VRAG and SORAG, where scores are collapsed into quasi-arbitrary bins, the location and width of which determines in large part the posterior probability of recidivism. Instead of getting a probability of recidivism for bin 9 of the VRAG (which includes all scores ≥ 28 , and was likely composed of around seven offenders), we can get a probability for Mr. Smith who scored exactly 28, no more, no less.

Table 1, column four, gives for each scale our estimated probability that Mr. Smith will recidivate. If this probability is greater than $.5$, it means that Mr. Smith scored above the

hitmax cutting score X_{HM} (the cutting score that maximizes classification accuracy, or “hits”), and indicates that if maximum accuracy is desired, the expert witness should predict recidivism for Mr. Smith or not.

Point estimates are useful, but we also want to present some indication of how error-prone those estimates are. Here we wish to improve on the confidence intervals above, and account for sampling error in the base rate, as well as in test score distributions. The base rate P is binomially distributed, and its variance is defined as $np(1 - p)$, where p is the base rate estimate and n is the number of subjects in the sample used to estimate p . In the Minnesota Department of Corrections (2000) report $p = .18$ and $n = 128$; $\text{Var}(p) = 18.89$. One way to compute the confidence interval would be to take normal distribution quantile (e.g., 1.96 for 95% c.i.) and calculate $np \pm 1.96 \sqrt{18.89}$. In this case $\text{CE} = [14.5, 31.6]$ which translates to a 95% confidence interval for $P_{\text{recidivism}}$ of [.11, .25].

Instead of using the normal approximation 1.96, we non-parametrically bootstrapped the confidence interval for the base rate $P = .18$ as well as the post-test probability of recidivism for each instrument under consideration. We report in Table 2 some summary statistics of the bootstrap pseudo-replicate distributions, including the minimum, the lower end of a 95% confidence interval (2.5%), the first quartile, median, mean, 3rd quartile, upper c.i. (97.5%), and maximum. As is readily apparent from the table, many risk instrument's 95% confidence interval overlaps, although this appears to be largely because every instrument has a relatively wide interval. The PCL-R has the largest interval (.52 units); the RRASOR and SVR-20 have fairly small intervals (.14 and .16 units respectively). These small intervals result from the large N available for the RRASOR, and from Mr. Smith's low score of 14 for the SVR-20, which means the SVR-20 posterior probability is calculated from the middle portion of the distribution, where a large number of offenders (whether recidivists or not) reside. This is in contrast to every other instrument, where Mr. Smith scored extremely high values and was therefore located in the high-end tails of the test score distributions, which are exceedingly unstable.

While many of the intervals overlap, not all do. Notably, the Static-99's interval conflicts with the SVR-20's, and the RRASOR's. The PCL-R's interval conflicts with the RRASOR's (and almost with the SVR-20's). This finding calls into serious question measures taken by the Static-99, RRASOR, PCL-R, and SVR-20, as the test's predictions essentially undercut (with high probability) each other as knowledge claims of Mr. Smith's true probability of reoffense. There may be some kind of serious systematic error intruding into these instruments' measures of Mr. Smith's risk. Equally likely, the measures are not measuring the same construct, at least with respect to Mr. Smith. Future research would benefit from latent structural analyses (e.g., confirmatory factor analyses) of these instruments, to determine exactly what each measure shares in common, and in where their measurements diverge. (Preliminary analyses in the Massachusetts dataset suggest two largely uncorrelated factors for these measures.) This would be especially useful for the RRASOR and Static-99, as they share some items and one thus does not expect their predictions to be discrepant. Overlap and redundancy between scales, and their impact on risk assessment, is the focus of the rest of our article.

We used for analyses in Table 2 the following datasets: for the RRASOR and Static-99 we combined data from the Massachusetts's sample with that from the respective development reports (Hanson, 1997; Hanson & Thornton, 1999). For all other instruments we used only the $N = 390$ Massachusetts dataset.

Finally, we construct a point estimate and confidence interval for the probability of recidivism given all tests combined (except the HCR-20, for which we can find very little

information in the published literature), accounting for the sometimes substantial correlations between them.

As what counts is the correlation between the recidivism probabilities the instruments generate (i.e., not the instrument raw scores), we would seek figures on the correlation between, say, the MnSOST-R probabilities and the Static-99 probabilities. Alas, these are not to be found in the literature. As a proxy for probability correlations, one can locate some figures on correlations between raw scores on risk instruments.

We have meta-analyzed what few published studies report raw score correlations between instruments. The resulting values are reported in Table 3. Point biserial correlations between tests and recidivism were not meta-analyzed by the present authors, but were computed from effect sizes d listed in Hanson and Morton-Bourgon's (2004) meta-analysis, under assumption that the base rate equals .18. These figures were obtained by the equation

$$d = \sqrt{2} \Phi^{-1}(.72) \quad (4)$$

since $AUC = \Phi(d/\sqrt{2})$ for a binormal ROC model, and 0.72 is a typical-to-high AUC in Hanson & Morton-Bourgon's (2004) meta-analysis of the MnSOST-R and other actuarial

risk prediction tools. Given d , we can calculate $r_{pbis} = \frac{d}{s_x} \sqrt{n_1 n_0 / (N(N-1))}$, the point biserial correlation between X and the criterion Y , where s_x is the standard deviation of a predictor X (this SD can be presumed equal to 1 without loss of generality), N is the total number of observations, and n_0 and n_1 are the numbers of recidivism and non-recidivism events in the calibration sample, respectively.

Hanson and Morton-Bourgon's (2004) meta-analysis reported instrument validities based on thousands of offenders and scores of studies, and represent the best estimates of instrument validity to date. In contrast, it is disconcerting that so few studies report correlations, which are simple to compute, easy to report, and contain very important information about relationships between predictor variables. What we need however is not only correlations between instruments, but partial correlations of instruments with recidivism, given all other instruments under consideration.

Based on two studies of partial correlation of measures with recidivism (Seto, 2005; Nunes, Firestone, Bradford, Greenberg, & Broom, 2002), we could with some justification assume that once a first measure taken into account, a second measure adds nothing to the accuracy of recidivism predictions. If we did make this assumption, we would adjust the probability estimate for a second instrument score, conditional on the first, to equal the target population base rate (.18 for Minnesota) for every individual, irrespective of his or her instrument score. Since this is a radical step based on only two studies, we instead assume for the nonce that correlations reported in the literature provide meaningful information about the true relationships between these instruments.

We can now use our correlation matrix to calculate optimal regression weights for each independent variable (i.e., risk instrument) for the dependent variable of recidivism. We multiply these weights by Mr. Smith's observed instrument scores and use a logistic link function to determine Mr. Smith's posterior probability (given *all* instruments) of recidivism under the .18 Minnesota base rate.

Preliminarily, we must note that the correlations reported in Table 3 do not tell the whole story. Risk instruments have imperfect reliability, and observed correlations between them

will be attenuated due simply to measurement error. A simple disattenuation function corrects this and delivers correlations between the tests' true scores (assuming classical test theory holds for these instruments, which is a problematic assumption, see Lumsden, 1976):

$$\widehat{\rho}_{X,Y} = \frac{r_{X,Y}}{\sqrt{\rho_{X,X} \cdot \rho_{Y,Y}}},$$

where $\widehat{\rho}_{X,Y}$ is the disattenuated correlation coefficient between X and Y , $r_{X,Y}$ is the observed correlation, and $\rho_{X,X}$ and $\rho_{Y,Y}$ are the reliabilities of instruments X and Y , respectively. We conservatively estimated a reliability coefficient of .9 for the Static-99, RRASOR, and SORAG; .85 for the PCL-R; .8 for the MnSOST-R; and .75 for the SVR-20 (for example reliability coefficients see Langton, et al. [2007] and Knight and Thornton [2007]), and applied this correction to the original correlation matrix, which resulted in the disattenuated matrix displayed in Table 4:

As seen in Table 4, the disattenuated correlation between the SORAG and VRAG is .995, which suggests that these two tests are nearly, if not entirely, redundant. Barring future research to the contrary, practicing clinicians should not interpret SORAG and VRAG results as mutually reinforcing. As the VRAG appears redundant with the SORAG, was originally developed on violent offenders (Quinsey, Harris, Rice, & Cormier, 1998), and only later used for predicting sexual recidivism of sex offenders, we chose to exclude the VRAG scores from all further analyses.

Concerns about classical test theory assumptions are significant. Reliabilities obtained in practice are always underestimates of reliability as defined in classical test theory (i.e., correlation between truly parallel tests), and thus inflate the disattenuation statistic (Lumsden, 1976). To avoid these issues, we chose to conduct the final analysis on the observed correlation matrix reported in Table 3. We used the following algorithm. (1) Invert the matrix. (2) rescale the matrix so the diagonal is a unit vector; (3) multiply entire matrix by -1 . The resulting square matrix contains the partial correlations for each variable given all other variables. Now, the partial correlations between each instrument and recidivism are the regression coefficients for that instrument. Next, (4) convert Mr. Smith's raw scores on each instrument to standardized scores, using whatever dataset is available for means and variances (we used the Massachusetts dataset). (5) Enter these values into the logistic regression model,

$$\text{Pr(Recidivism|Instrument Scores)} = \frac{\exp(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6)}{1 + \exp(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6)},$$

where b_0 equals zero and is the intercept, b_1, \dots, b_6 are the coefficients for the instruments, and x_1, \dots, x_6 are Mr. Smith's scores on these instruments. For purpose of illustration the instruments are numbered in the following order: (1) Static-99; (2) SORAG; (3) SVR-20; (4) RRASOR; (5) PCL-R; (6) MnSOST-R. All data was taken from the Massachusetts dataset, as well as the RRASOR and Static-99 development reports (Note 3). Mr. Smith's scores yield the following result

$$.399 = \frac{\exp\{0 - .095(2.3) + .078(2.8) + .575(-.17) + .186(1.2) - .466(2.7) - .308(.51)\}}{1 + \exp\{0 - .095(2.3) + .078(2.8) + .575(-.17) + .186(1.2) - .466(2.7) - .308(.51)\}}.$$

That is, according to the best information available to us, Mr. Smith has a 40% chance of committing a future sex crime and, if Dr. Fisbee is interested in minimizing the number of errors he makes over his career, he would predict that Mr. Smith will not recidivate sexually. Unfortunately, it is impossible at present to construct a defensible confidence interval for this estimate, at least without assuming the tests are jointly multivariate normally distributed. Under such an assumption one could conduct a parametric bootstrap to estimate the 95% c.i.; we do not believe this a defensible approach. The other, more easily justifiable solution would be to conduct a traditional logistic regression with raw data and compute standard errors for the predicted probabilities. Raw multivariate data from the meta-analyzed studies were not available to us during the drafting of this article, preventing us from following this tack.

Conclusions

We begin summarizing by providing, very briefly, an outline of the problems and their solutions.

1. Test scores should not deliver observed, but implausible, probability values. Simple corrections can be made to observed values to address this issue.
2. Test scores must be calibrated to predict the same criterion (e.g., arrest versus reconviction; five years versus 10 years hence), otherwise the scores are difficult to compare. We argue that a future direction in recidivism prediction is construct validation of the outcome variable itself, recidivism. We liken this potential for development to IQ testing.
3. Test scores should deliver quantities, such as probability of recidivism. Locutions such as “high risk” convey in an unnecessarily limited sense the information about an individual’s risk to recidivate (e.g., imminence, severity, probability).
4. Test scores should *all* be adjusted to reflect the local base rate (or a set of defensible base rates). Cutting scores should be set accounting for this base rate.
5. Optimum cutting scores can be determined, if merely by tabulating the range of possible hit rates for any test/base rate combination. Utilities also impact cutting score and cost minimization may be considered the goal of violence prediction.
6. Finally, the answers to our last four problems are heavily interdependent, and are best answered jointly. Measurements from instruments purporting to measure the same construct should agree, within measurement error tolerance. When they do not, a deep epistemological problem is broached, which cannot be satisfactorily solved by box scores or averaging. Instead, it indicates that something seriously wrong has occurred in the measurement of Mr. Smith, and discrepant predictions undercut each other as knowledge claims. To complicate matters, test content often overlaps between tests, the effect of which on test score consistency is unknown.

We provide several ways (parametric and nonparametric) to compute confidence intervals for prediction estimates, to determine if they are mutually confirming or disconfirming. Finally, we suggest that conduct of a logistic regression based on meta-analyzed correlation weights is one defensible way to completely combine information from multiple measures, and arrive at a single estimate of risk.

From the forgoing, the reader can see that predicting sex offender recidivism, as problem in applied psychology, has quite a number of potential pitfalls. It remains to be seen whether any new scale development or redevelopment incorporate any of the problem solutions outlined here. If they do, it still remains to be seen whether a worthwhile improvement in recidivism prediction, from the approximately .7 AUC reported for actuarial instruments by

Hanson and Morton-Bourgon (2004). If they do not adopt at least a number of these recommendations (such as an attempt to combine tests), one can expect results following the old adage: “Keep doing what you’ve been doing, and you’ll get what you’ve been getting.” Our opinion is that an AUC of .7 is too low for principled use in forensic settings, and in determining incarceration for indefinite periods. This is especially true given that the AUC’s reported in the literature present an upper bound on the accuracy obtainable by a working clinician. One expects, given clinician (i.e., human) fallibility in determining base rates, scoring instruments, applying cutting scores, combining the results from diverse tests, and making clinical/professional judgements during the entire process, that long-term clinical field accuracy will fall short of AUC’s reported in the literature. We only hope to have assisted in informing clinical practice of a handful of error-inducing pitfalls and complications that routinely arise in the assessment of risk.

Reference Notes

1. Some of the details concerning the case have been altered to disguise the identity of the subject of the petition in question.
2. See Meehl (1954) for a sophisticated discussion of why no such “hybrids” exist, and why the SVR-20, for example, is an instantiation of clinical judgment, and can be expected to incorporate all those human biases, errors, and heuristics we expect of clinical judgment (Garb, 1999). This likely contributes to its low interrater reliability, as discussed below.
3. We ignore the multilevel nature of this model to prevent further complications, as this algorithm is for illustrative purposes. Technically, there is clustering within samples in the present analysis, as well as in the original Static-99 and RRASOR development reports (which also ignored the multilevel nature of the data).

Acknowledgments

The first author was supported by NIMH Training Grant 5T32MH017069-27.

References

- Barbaree HE, Seto MC, Langton CM, Peacock EJ. Evaluating the predictive accuracy of six risk assessment instruments for adult offenders. *Criminal Justice and Behavior* 2001;28:758–762.
- Bartko JJ. On various intraclass correlation coefficients. *Psychological Bulletin* 1976;83:762–765.
- Boer, DP.; Hart, SD.; Kropp, RP.; Webster, CD. Manual for the Sexual Violence Risk-20: Professional Guidelines for Assessing Risk of Sexual Violence. Vancouver, Canada: The British Columbia Institute Against Family Violence; Simon Fraser University Mental Health, Law, and Policy Institute; 1997.
- Crawley, MJ. The R book. Hoboken, NJ: Wiley; 2007.
- Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychological Bulletin* 1955;52:281–302. [PubMed: 13245896]
- Daubert, v. Merrell Dow Pharmaceuticals, Inc. 509 U.S. 579. 1993
- Dawes RM. The robust beauty of improper linear models in decision making. *American Psychologist* 1979;34:571–582.
- Doren D. Recidivism Base Rates, Predictions of Sex Offender Recidivism, and the “Sexual Predator” Commitment Laws. *Behavioral Sciences and the Law* 1998;16:97–114.
- Ducro C, Pham T. Evaluation of the SORAG and Static-99 on Belgian sex offenders committed to a forensic facility. *Sexual Abuse: A Journal of Research and Treatment* 2006;18:15–26. [PubMed: 16598662]
- Epperson, DL.; Kaul, JD.; Huot, S.; Goldman, R.; Alexander, W. Minnesota Sex Offender Screening Tool Revised (MnSOST-R) technical paper: Development, validation, and recommended risk level

- cut scores. 2003. Retrieved September 3, 2007, from Iowa State University Department of Psychology Web site: http://www.psychology.iastate.edu/~dle/mnsost_download.htm
- Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 1973;33:613–619.
- Frankish, K. Non-monotonic reasoning. In: Brown, K., editor. *Encyclopedia of language and linguistics*. 2. Oxford: Elsevier; 2005.
- Garb, HN. *Studying the Clinician: Judgment Research and Psychological Assessment*. Washington, D.C.: American Psychological Association; 1999.
- Grove WM. When is a diagnosis worth making? A comparison of two statistical prediction strategies. *Psychological Reports* 1991;68:317.
- Grove, WM. Base rates, Bayes Theorem, and performance of diagnostic tests. In: Grove, WM., editor. *Mathematical aspects of diagnosis*. in preparation
- Grove WM, Barden RC. Protecting the integrity of the legal system: The admissibility of testimony from mental health experts under Daubert/Joiner/Kumho analyses. *Psychology Public Policy and Law* 1999;5:224242.
- Grove WM, Zald DH, Hallberg AM, Lebow B, Snitz E, Nelson C. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment* 2000;12:19–30. [PubMed: 10752360]
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36. [PubMed: 7063747]
- Hanson, RK.; Morton-Bourgon, KE. *Predictors of sexual recidivism: An updated meta-analysis*. Ottawa: Public Safety and Emergency Preparedness Canada; 2004. User report 200402
- Hanson, RK.; Thornton, D. *Static 99: Improving Actuarial Risk Assessments for Sex Offenders*. Ottawa: Department of the Solicitor General of Canada; 1999. User report 1999-02
- Hanson, RK. *The Development of a Brief Actuarial Risk Scale for Sexual Offense Recidivism*. Ottawa: Department of the Solicitor General of Canada; 1997. User report 1997-04
- Hare, RD. *The Hare Psychopathy Checklist-Revised (PCL-R)*. Toronto: Multi-Health Systems; 2003.
- Harris GT, Rice ME, Quisey VL, Lalumiere ML, Boer D, Lang C. A multisite comparison of actuarial risk instruments for sex offenders. *Psychological Assessment* 2003;15:413–425. [PubMed: 14593842]
- Hoeting JA, Madigan M, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. *Statistical Science* 1999;14:382–417.
- Knight, RA.; Thornton, D. *Evaluating and improving risk assessment schemes for sexual recidivism: A long-term follow-up of convicted sexual offenders*. 2007. Retrieved July 19, 2008 from: www.ncjrs.gov/pdffiles1/nij/grants/217618.pdf
- Kyburg, HE, Jr. *The logical foundations of statistical inference*. Dordrecht, Netherlands: D. Reidel; 1974.
- Langton CM, Barbaree HE, Seto MC, Peacock EJ, Harkins L, Hansen KT. 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* 2007;34:37–59.
- Looman J. Comparison of two risk assessment instruments for sexual offenders. *Sexual Abuse: A Journal of Research and Treatment* 2006;18:193–206. [PubMed: 16598660]
- Lumsden J. Test theory. *Annual Review of Psychology* 1976;27:251–280.
- Meehl, PE. *Clinical versus actuarial prediction: A theoretical analysis and review of the evidence*. Minneapolis, MN: University of Minnesota Press; 1954.
- Meehl PE. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology* 1978;46:806–834.
- Meehl PE. Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry* 1990;1(108141):173–180.
- Meehl PE, Rosen A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin* 1955;52:194–216. [PubMed: 14371890]
- Minnesota Department of Corrections. *Sex offender policy board and management study*. St Paul, MN: Author; 2000.

- Nunes KL, Firestone P, Bradford JM, Greenberg DM, Broom I. A comparison of modified versions of the Static-99 and the Sex Offender Risk Appraisal Guide. *Sexual Abuse: A Journal of Research and Treatment* 2002;14:253–269. [PubMed: 12087686]
- Perrin, J. *Les Atomes*. Paris: F Alcan; 1913.
- Popper, KR. *Logic of scientific discovery*. London: Routledge; 1959.
- Quinsey, VL.; Harris, GT.; Rice, MO.; Cormier, CA. *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association; 1998.
- Rice ME, Harris GT, Lang C, Cormier C. Violent sex offenses: How are they best measured from official records. *Law and Human Behavior* 2006;30:525–541. [PubMed: 16770703]
- Salmon, WC. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press; 1998.
- Scott, DW. *Multivariate density estimation: Theory practice and visualization*. New York: Wiley; 1992.
- Seto MC. Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. *Psychological Assessment* 2005;17:156–167. [PubMed: 16029103]
- Shao J. Bootstrap model selection. *Journal of the American Statistical Association* 1996;91:655–665.
- Shao J. An asymptotic theory for linear model selection. *Statistica Sinica* 1997;7:221–264.
- Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society series B* 1991;53:683–690.
- Silverman, BW. *Density estimation for statistics and data analysis*. Boca Raton, FL: Chapman & Hall; 1996.
- Simonoff, JS. *Smoothing methods in statistics*. New York: Springer-Verlag; 1996.
- Swetts, JA.; Pickett, RM. *evaluation of diagnostic systems: Methods from signal detection theory*. New York, NY: Academic Press; 1982.
- Tukey, PW. *Exploratory data analysis*. New York: Addison-Wesley; 1977.
- Venables, WN.; Ripley, BD. *Modern applied statistics with S*. 4. New York, NY: Springer; 2002.
- Vrieze SI, Grove WM. Predicting sex offender recidivism. I. Correcting for item overselection and accuracy overestimation in scale development. II. Sampling error-induced attenuation of predictive validity over base rate information. *Law and Human Behavior* 2008;32:266–278. [PubMed: 17610052]
- Wand, MP.; Jones, MC. *Kernel Smoothing*. Boca Raton, FL: Chapman & Hall; 1995.
- Webster, CD.; Douglas, KS.; Eaves, D.; Hart, SD. *Assessing The Risk of Violence, Version 2*. Burnaby, British Columbia, Canada: Simon Fraser University and Forensic Psychiatric Services Commission of British Columbia; 1997. HCR-20.

Table 1

Dr. Fisbee's Predictions, Followed by the Present Author's Kernel Density Estimated Predictions and Recommended Cutting Scores.

Scale	Score	Prediction	KDE Prediction (5 year)	X_{Hitmax} BR = .18
Static-99	8	.39 (5 year)	.51	out of range ^a
		.45 (10 year)		
		.52 (15 year)		
SORAG	38	1.00 (violent; 7-year)	.30	out of range ^a
SVR-20	14	"Medium to High"	.17	31
RRASOR	4	.33 (5 year)	.18	out of range ^a
VRAG	28	1.00 (violent; 7-year)	.26	32
PCL-R	29.5	.81 (general)	.47	32
		.37 (violent)		
MnSOST-R	18	.88 (.35 base rate)	.26	out of range ^a
		.78 (.21 base rate)		
		.70 (.15 base rate)		
HCR-20	32	.92	N/A	N/A

^aIndicates a hitmax cutting score that was beyond the test's score range. (Kernel density estimates extrapolate beyond observed data according to kernel properties, the normal kernel ranges from $-\infty$ to ∞ , just like the normal distribution.)

Table 2

Bootstrapped distributions of Mr. Smith's Probability of Recidivism Using KDEs.

Scale	Score	Min	1st Q	Median	Mean (95% c.i.)	3rd Q	Max
Static-99	8	.22	.41	.47	.47 (.29, .66)	.53	.73
SORAG	38	.02	.21	.27	.29 (.09, .53)	.35	.92
SVR-20	14	.06	.14	.17	.17 (.10, .26)	.20	.32
RRASOR	4	.06	.13	.15	.15 (.09, .23)	.17	.28
VRAG	28	.001	.17	.26	.27 (.04, .58)	.36	.82
PCL-R	29.5	.16	.40	.49	.49 (.26, .78)	.58	.96
MnSOST-R	18	.04	.20	.27	.28 (.11, .51)	.34	.71

"1st Q" and "3rd Q" indicate the 1st and 3rd quartiles, the 95% c.i. denotes the 95% confidence interval, the Mean is Mr. Smith's expected probability of sexual reoffense.

Table 3

Meta-analyzed Correlation Coefficients

Test Pair	r	95% CI	k	N	Q	Studies
Static-99:						
—SORAG	.637	[.61, .66]	6	1917	6.1	1,2,3,4,5,6
—SVR-20	.491	[.41, .56]	1	390	0	1
—RRASOR	.799	[.78, .82]	3	1254	22.3*	1,2,4
—VRAG	.473	[.43, .52]	3	1254	.2	1,2,4
—PCL-R	.466	[.40, .53]	2	602	.1	1,7
—MnSOST-R	.589	[.54, .63]	2	744	4.5*	1,2
—Recidivism	.242		21	5103	44.2*	9
SORAG:						
—SVR-20	.693	[.64, .74]	1	390	0	1
—RRASOR	.425	[.38, .47]	3	1254	.4	1,2,4
—VRAG	.915	[.91, .92]	3	1254	4.2	1,2,4
—PCL-R	.761	[.73, .79]	2	602	1.7	1,7
—MnSOST-R	.518	[.46, .57]	2	744	.5	1,2
—Recidivism	.184		5	1348	4.2	9
SVR-20:						
—RRASOR	.172	[.07, .27]	1	390	0	1
—VRAG	.634	[.57, .69]	1	390	0	1
—PCL-R	.742	[.69, .78]	1	390	0	1
—MnSOST-R	.509	[.43, .58]	1	390	0	1
—Recidivism	.296		6	819	19.0*	9
RRASOR:						
—VRAG	.197	[.14, .25]	3	1254	5.7*	1,2,4
—PCL-R	.135	[.06, .21]	2	602	<.1	1,7
—MnSOST-R	.425	[.37, .47]	3	1000	.2	1,2,8
—Recidivism	.227		18	5266	55.8*	9
VRAG:						

Test Pair	<i>r</i>	95% CI	<i>k</i>	<i>N</i>	<i>Q</i>	Studies
—PCL-R	.651	[.60, .69]	2	602	45.6*	1,7
—MnSOST-R	.400	[.34, .46]	2	744	<.1	1,2
—Recidivism	.200		5	1147	8.1	9
PCL-R:						
—MnSOST-R	.384	[.31, .45]	2	602	1.9	1,2
—Recidivism	.111		13	2783	14.36	9
MnSOST-R:						
—Recidivism	.254		4	813	4.76	9

Note: Where *r* is the meta-analyzed correlation; *k* the number of studies; *N* the total number of subjects from the *k* studies; *Q* the homogeneity statistic. Studies indicates from which publications 1-9 below the correlation statistics were obtained.

* *p* < .05

1. Knight and Thornton (2007).
2. Langton, Barbaree, Seto, Peacock, Harkins, and Hansen (2007).
3. Duero and Pham (2006).
4. Harris, Rice, Quimsey, Lalumiere, Boer, Lang (2003).
5. Looman (2006).
6. Nunes, Firestone, Bradford, Greenberg, and Broom (2002).
7. Barbaree, Seto, Langton, Peacock (2001).
8. Epperson, Kaul, Huot, Goldman, and Alexander (2003).
9. Hanson and Morton-Bourgon (2008).

Table 4

Disattenuated Meta-analyzed Correlation Coefficients

	Static-99	SORAG	SVR-20	RRASOR	VRAG	PCL-R	MnSOST-R	Recidivism
Static-99	1							
SORAG	.710	1						
SVR-20	.597	.843	1					
RRASOR	.840	.210	.210	1				
VRAG	.520	.995	.772	.159	1			
PCL-R	.533	.872	.929	.154	.774	1		
MnSOST-R	.698	.611	.657	.502	.471	.467	1	
Recidivism	.242	.184	.296	.223	.200	.111	.253	1

Note: Table values assume following (conservative) estimates of instrument reliabilities: .9 for Static-99, SORAG, VRAG, and RRASOR; .85 for PCL-R; .8 for MnSOST-R; .75 for SVR-20. See Langton et al. (2007) and Knight and Thornton (2007) for example reliabilities.