

Running head: Shrinkage in the MnSOST-R

The Importance of Cross-Validation in Actuarial Test Construction
for Correcting Inflated Recidivism Predictions: Shrinkage in the Risk
Estimates for the Minnesota Sex Offender Screening Tool – Revised.

Richard Wollert

Institute for the Study of Mandated Treatment

Requests for reprints and correspondence concerning this article should be sent to: Richard
Wollert, Ph.D., Institute for the Study of Mandated Treatment, 1220 SW Morrison St. #929,
Portland, OR 97205; Phone: (503) 241-0466

In press, to appear in the Journal of Threat Assessment

Abstract

Many states have passed laws for the civil commitment of sexually violent predators. Experts who testify at commitment hearings often base their opinions on actuarial tests for the prediction of sexual recidivism (ATSRs). Reliance on actuarial tests is promising, but they need to be accurate. Cross-validation is a widely recommended procedure for determining accuracy. In this paper the predictive accuracy of an ATSR called the Minnesota Sex Offender Screening Tool-Revised (MnSOST-R) was evaluated from a cross-validated perspective. The original predictions of the MnSOST-R were found to be greatly inflated. Dissemination of the MnSOST-R without adequate cross-validated research raises serious doubts about its adequacy. It also creates doubt in the professional community that other ATSRs might have been disseminated without sufficient attention to important test construction standards. Formulating a set of standards targeted specifically on the construction of ATSRs would be helpful to test developers, test users, and the court for evaluating their adequacy. Such a set of standards is proposed by the author. It is hoped that they will promote an adjustment of values so that test construction principles will be accorded the importance they deserve as ATSRs are developed in the future.

Keywords: Actuarial tests, civil commitment, sexually violent predators, Minnesota Sex Offender Screening Tool - Revised (MnSOST-R)

The Importance of Cross-Validation in Actuarial Test Construction
for Correcting Inflated Recidivism Predictions: Shrinkage in the Risk
Estimates for the Minnesota Sex Offender Screening Tool – Revised.

A large number of states have enacted legislation allowing for the post-prison civil commitment of sex offenders who are considered “dangerous” or “sexually violent” (Campbell, 2000; LaFond & Winick, 1998).

According to Covington (October 1997), the commitment process generally involves several steps. First, a petition to detain an inmate is filed by either the attorney general of the state where the inmate has been imprisoned or the “state’s attorney from the county of the original proceedings that caused the alleged sexually violent predator (SVP) to be jailed or confined” (p. 167). Then the court reviews the petition to determine if there is reason to believe that the inmate is eligible for commitment and a detention order is issued if this is the case. A hearing is then scheduled to assess whether there is “probable cause” to believe that the defendant meets the criteria for being classified as an SVP in that he “a) ... has engaged in harmful sexual conduct in the past, b) ... suffers from a mental disorder, and c) ... will likely engage in future acts of harmful sexual misconduct” (Janus & Meehl, 1997, pp. 33-34). If this condition is satisfied, a trial is held at which the state must show, beyond a reasonable doubt, that the defendant meets the SVP criteria. A commitment order is prepared if the defendant is found to be an SVP.

In most SVP trials it is not too difficult to show that a defendant meets the first two conditions listed above. The critical step, as far as most commitment decisions are concerned, therefore turns on whether it may be predicted beyond a reasonable doubt that defendants are

“likely”, and in some states “highly likely”, to commit another sexually violent offense if released.

Enormous stakes are involved, however, when it comes to making predictions about sexual violence at SVP trials. Janus and Meehl (1997), for example, have discussed the costs associated with erroneous predictions of violence and nonviolence. In particular, they have pointed out that excessively optimistic predictions may be followed by destructive acts of sexual violence while unduly pessimistic predictions result in lengthy and unnecessary deprivations of liberty associated with expensive and intrusive treatment. From their perspective, “both types of prediction errors are costly in many ways” (p.35).

In light of these costs, it is not surprising that attorneys for both the state and the defense who are involved in SVP hearings usually retain mental health experts to assess the probability that defendants will sexually recidivate. Many of these experts, in turn, base their predictions heavily on “actuarial tests” that are referred to by such names as the Violence Risk Appraisal Guide (VRAG; Quinsey, Harris, Rice, & Cormier, 1998), Sex Offender Risk Appraisal Guide (SORAG; Quinsey et al., 1998), Rapid Risk Assessment for Sex Offender Recidivism (RRASOR; Hanson, 1998), Static-99 (Hanson & Thornton, 2000), and Minnesota Sex Offender Screening Tool – Revised Form (MnSOST-R; Epperson, Kaul, & Hesselton, 1999; Epperson, Kaul, & Huot, 1995).

These instruments share a number of characteristics. First, each includes several “predictor” items that were selected because, compared to other possible items, they were found to be highly correlated with either sexual or violent recidivism. Second, overall “risk scores” are obtained for each defendant to whom a test is administered by combining the scores achieved on each predictor item. Although some items may be weighted more heavily than

others, defendants who are positive for many items usually obtain scores placing them in a high risk group while defendants who are positive for only a few obtain low risk scores. Third, the developers of each test have compiled “experience tables” (Ohlin, 1951) from retrospective studies of released sex offenders that show the percentage of offenders in each risk group who have recidivated.

Using one of these tests to generate a prediction of sexual violence for an SVP defendant is fairly straightforward when an expert adopts a purely actuarial approach to interpret the test results¹. The expert first determines how the defendant scores on each item of the selected test and then calculates the defendant’s risk score from individual item scores. An experience table for the test is consulted to identify the recidivism rate associated with the defendant’s risk group. If the rate reaches a level that indicates to the expert that the defendant is likely or highly likely to recidivate, the expert communicates this to the relevant attorney and, at the attorney’s request, to the court.

The courts have not explicitly quantified the rates associated with either a high or very high likelihood of recidivism. On the basis of one decision, however, the authors of a recently published influential article adopted a benchmark of 50% as corresponding to the term “likely to recidivate” and 75% as corresponding to the term “highly likely to recidivate” (Janus & Meehl, 1997). The author has also consulted on about 20 cases where one of the primary issues was whether or not a defendant was “likely to recidivate”. In almost all of these cases, attorneys and experts have placed a great deal of emphasis on whether defendants fall in a risk group associated with a recidivism rate above 50%. This experience has suggested that, in practice, a risk level that exceeds 50% is probably the most widely accepted defacto definition of the term “likely to recidivate”.

Theoretically, the prospect of relying on true actuarial tests to predict sexual violence is promising in several respects. Predictive accuracy, for example, would probably be improved because actuarial tests have always been found to be more accurate than “intuitive” clinical judgment when it comes to making predictions about recidivism or parole outcomes (Grove, Zald, Lebow, Snitz, & Nelson, 2000). The reliance of actuarial tests on prespecified rules also holds out the promise that the evaluation process, even for a group as highly stigmatized as sex offenders, might be so fair and reliable that it would often be possible for defense and prosecution experts to agree in their opinions. Finally, by averaging actuarial results for representative samples of defendants, it might be possible to estimate the quantitative standards courts have adopted for classifying defendants as likely and highly likely to recidivate in a sexually violent manner.

In order to realize these gains, however, the tests that are used must be valid, reliable, and relevant for the purpose of predicting sexually violent behavior. Is there any way that test developers can assure the court that their instruments are truly valid, reliable, and relevant? Unquestionably. How? By following test construction guidelines that are emphasized in the vast literature on test development and reporting the results of these developmental efforts in manuals about the tests. A small but meaningful portion of this literature deals with the development of actuarial tests for predicting sexual recidivism (ATSRs). A large portion deals with the formulation of actuarials for the prediction of criminal behavior. An enormous portion deals with the construction of psychological tests, but is highly relevant for the construction of ATSRs because, being “prognostic, and evaluative devices” (American Psychological Association, 1985, p. 3) that are developed with the purpose in mind of providing an “objective

and standardized measure of a sample of behavior” (Anastasi, 1988, p. 23; Cronbach, 1970), they clearly meet the criteria for being considered psychological tests.

Accessing guidelines for the development of psychological tests is a simple matter because these guidelines are available in a single source that is updated periodically (American Psychological Association, 1985; American Psychological Association, 1999). In contrast, although a number of observers have raised relevant questions and issues (Campbell, 2000; Dawes, Faust, & Meehl, 1989; Gottfredson, 1987; Hanson, 1998; Heilbrun, 1992; Rogers, 2000), a comparable source of standards for the construction of actuarials has never been compiled. This is a very unfortunate oversight, as the formulation of such a set of standards would enhance predictive accuracy in SVP cases by enabling the court to evaluate the scientific merits of ATSRs on a test by test basis, thereby minimizing the administration of invalid or misleading tests.

If standards for the construction of ATSRs were ever formulated, some guidelines would undoubtedly emphasize the importance of collecting test data on a fresh sample of subjects once the data used for test development were analyzed. Three considerations, discussed by Simon (1971), point up the significance of this procedure, known as “cross-validation”. First, the multiple regression procedures that are used to extract predictor items from a large pool of items administered to an initial subject sample capitalize on factors that are often related to chance. Recidivism estimates for the lowest and highest score groups of a test under development are consequently almost always inflated. In turn, the correlation between the risk scores of a test and its “criterion” – that is, the outcome the test predicts – is inflated. Second, the extent of inflation for each risk group may be determined by collecting data from a follow-up sample of subjects that is representative of the population to which the test will be applied. Third, on the

basis of the fresh data, the true test-criterion correlation may be estimated with greater certainty and the original recidivism table for the test may be revised to accurately represent its performance.

An intuitive understanding of the importance of cross-validation may be gained from an examination of Figure 1, based on data presented by Simon (1971, p.109). One test that he and his colleagues developed to predict failure while on probation was found to have a strong correlation with the criterion in the initial subject sample ($r=.41$), apparently subdivided the sample into 4 distinct risk score groups, and generated an experience table that seemed much more accurate than the overall failure rate for making predictions about subjects with the highest and lowest scores. The performance of this test, based only on data from the construction sample, is represented by the dotted and dashed line in Figure 1. Follow-up data were then collected to verify the test's performance. An analysis of this data documented a high degree of "shrinkage" in the test-criterion correlation, so that it was much lower than originally estimated ($r=.12$). The follow-up study also showed that the test was capable of subdividing the sample into only two score groups, and that predictions based on the test were not any more accurate than simply using the failure rate as the best estimate for everyone in the sample. The dashed line in Figure 1 represents the performance of the cross-validated test. Simon's data clearly underscore the value of cross-validation for verifying test performance.

The author recently evaluated the predictive accuracy of an ATSR known as the MnSOST-R from a cross-validated perspective. The significance of the MnSOST-R for SVP cases is discussed in the next section, and a couple of factors are summarized that suggest it yields inflated estimates of sexual recidivism. Cross-validated data that verify this flaw are presented at the end of this section. The implications of these results for the MnSOST-R are

described in the final section. The concern that a significant subgroup of experts from the professional community share about the adequacy of ATSRs will be discussed, and the importance of test construction standards for addressing doubts about the scientific merits of ATSRs will be considered.

Use of the MnSOST-R in SVP Cases: Reasons Why Its Risk Estimates
Are Inflated and Cross-Validation Data Verifying This Flaw

The MnSOST-R (Epperson et al., 1995; Epperson et al., 1999) is widely used by behavioral scientists to determine the chances that sex offenders being considered for civil commitment will sexually recidivate. One of the reasons for the test's popularity, compared to alternative instruments, may be that it seems to be a powerful tool for differentiating between subjects. For example, as column F of the developmental experience table presented as Table 1 indicates, 6-year predictions derived from the MnSOST-R range from 0% to 88% and risk scores for the MnSOST-R appear to subdivide this continuum into 6 distinct categories. In contrast, 15-year predictions for an alternative test known as the "Static-99" (Hanson & Thornton, 2000) only go from 7% to 52% and the predictions from adjacent levels are in some cases nearly identical.

Insert Table 1 About Here

A couple of methodological considerations suggest that the MnSOST-R's recidivism predictions might be inaccurate. The first of these is that recidivists were "oversampled" during the developmental phase of the instrument's construction so that the failure rate reached 35% (Epperson et al, 1995). Because of this inflated base rate, the risk predictions could also easily have been inflated. The second consideration is that the MnSOST-R was distributed for use

without first being tried out on a second sample (Hanson, 1998). As noted earlier, this is a problem because “cross-validation” is necessary to control for inflated estimates and is recognized by virtually all prediction experts as a critical step for compiling a truly accurate prediction table ².

Shortly after a “final report” on the MnSOST-R was distributed an attempt was made to correct some of the problems associated with oversampling by estimating what the appropriate recidivism predictions would be for a population with an overall recidivism rate of 21% (Epperson et al., 1999). For the top three categories (i.e., risk groups IV, V, and VI from Table 1), the average rate reduction was about 13%. Since a cross-validation wasn’t carried out, however, the accuracy of even these supplementary estimates has remained in doubt.

Lack of a cross-validation study has therefore delayed compilation of a traditional prediction table and assessment of the accuracy of the initial MnSOST-R predictions. Usually, the issue of cross-validation is taken up after a test’s developmental results have been published in a peer-reviewed source. The MnSOST has been under development since 1991 (Epperson et al., 1999), however, and has yet to be published in such a source. Nonetheless, it has been weighted heavily by expert witnesses in many civil commitment cases throughout the United States. This practice seems highly questionable in the absence of cross-validated evidence. Given this situation, the only responsible course of action is to begin assessing the cross-validated status of the MnSOST-R immediately with whatever relevant data are available.

Recently, the developers of the MnSOST-R distributed a set of predictive results via the internet that combined the data for their original subjects with data for a non-selected cross-validated (CV) group that had about a 22% failure rate (Epperson, 2000). Although the new report pooled the data and didn’t present a full experience table, it gave enough data so that it

was possible for the author to compile the experience table for the combined samples. This is presented as Table 2. It was then possible to compile the table for the CV sample – which is a true prediction table – by subtracting the failures and successes in columns C and D of Table 1 from the successes and failures in columns F and G of Table 2. This result is presented as Table 3.

Insert Tables 2 and 3 About Here

Comparing column F from the first and third tables, it is clear that the recidivism rates obtained with the MnSOST-R developmental sample were inflated and that they underwent a dramatic amount of “shrinkage” on cross-validation. Since the CV sample has a 23% recidivism rate it is clear from the third table that the supplementary recidivism estimates that were proposed for a population with a 21% failure rate were also inflated.

Discussion

The extent of shrinkage in the recidivism predictions generated by the MnSOST-R holds serious practice implications. For example, using a reasonable benchmark of 75%, in the past experts could argue that subjects in the two highest MnSOST-R risk groups were highly likely to recidivate. On the basis of the cross-validated results this is no longer justified. In fact, the results in Table 3 suggest that the current MnSOST-R contains only three distinguishable groups (i.e., I and II, III and IV, V and VI), rather than six. A reasonable argument might even be advanced that the MnSOST-R is irrelevant for civil commitment cases. The reasoning behind this position is that the benchmark for civil commitment should exceed 50%, yet none of the estimates from the prediction table for the MnSOST-R reach this level. Finally, psychologists and other mental health experts who have used the cut scores disseminated by the developers of the MnSOST-R in past SVP cases might choose to avoid its use in the future. This suggestion is

based on two considerations. The first is that psychologists are, from an ethical standpoint, required to “rely on scientifically...derived knowledge” when engaging in “professional endeavors” like assessments (American Psychological Association, 1992, p. 4). The second is that, as the results of the present cross-validated analysis suggest, it is questionable whether the procedures for the compilation of the experience table for the MnSOST-R meet the standard of scientific derivation.

About the same time that the study at hand was completed, the developers of the MnSOST-R announced that cross-validation data were available for another 115 subjects plus the 95 subjects in the initial cross-validation sample (Epperson, Huot, Hesselton, Alexander, & Graham, 2000). The author requested data on the number of subjects in each score group and their recidivism rates, but these data were not released (Epperson, personal communication, April 5, 2001). Had the data for the additional subjects been analyzed, however, a meaningful difference in the present results would not have been obtained. The reason for this is that the recidivism rates reported by Epperson and his colleagues (2000) for the 220 subjects in what they referred to as the “full CV sample” was 20% whereas the recidivism rate for a subset of the full sample they called the “full 6-year risk sample” was 26%. These rates are almost the same as the 23% rates for the 95 subjects presented in Table 3. Mathematically, it is simply not possible for the risk levels for the larger groups to approximate the original risk levels in Table 1 unless their recidivism rates were much greater than 23% (Janus & Meehl, 1997).

Oversampling is a useful technique for identifying items that are capable of differentiating recidivists from nonrecidivists. It can lead to some very negative consequences, however, when it is not paired with cross-validation (Simon, 1971). Still, the lack of appreciation that test developers in the criminal justice field have sometimes shown for the

lessons of history make it an important principle to restate every so often. The shrinkage in the risk estimates for the MnSOST-R may, from this perspective, be seen as one of the latest installments in this ongoing saga.

This episode might not be too troubling if actuarials were usually ignored when it came to making commitment decisions. Actuarials have been considered in all of the SVP hearings involving the author, however, and, with one exception, all of the commitment candidates who received a high risk estimate on any actuarial were classified as SVPs. Furthermore, upon analyzing the results of actuarial testing for 13 cases, the author found that the MnSOST-R consistently generated the highest risk estimates when compared to the RRASOR, Static-99, and a modified form of the VRAG (Wollert, 2000). Finally, a stunning difference in risk estimates was apparent between the actuarials, with the predictions based on the MnSOST-R averaging close to 75% and the predictions from the other instruments clustering at about 35%. These considerations suggest that predictions based on the MnSOST-R may have unnecessarily placed the freedom of many individuals in jeopardy.

The fact that the MnSOST-R was developed and disseminated for use in the absence of adequate cross-validated research clearly undermines confidence in the test itself. More generally, it raises doubts within the professional community as to whether critical quality control standards were overlooked when other ATSRs were developed. To examine this issue further, the author recently consulted with a number of social workers, psychologists, and psychiatrists throughout the country who are frequently involved in SVP cases but argue against the use of ATSRs at this point in their development. In addition to pointing out many reservations about ATSRs, they indicated that other experts shared their views and provided a number of manuscripts, preprints, or reprints that elaborated on the limitations of ATSRs

(Campbell, 2000; Campbell, February-March 2001; Donaldson, 2000; Otto, Borum, & Hart, 2001; Otto & Petrila, in press; Petrila & Otto, in press; Rogers, 2000).

These exchanges suggest that a significant portion of the professional community will continue to advance the argument that ATSRs, as a class of tests, should not be used because of flaws in their development. Although an uncritical acceptance of these instruments is obviously problematic (Rogers, 2000), a blanket rejection of ATSRs might squelch continued research on those that are genuinely promising. A set of standards targeted specifically on the construction of ATSRs may be helpful to the court when evaluating the strengths and limitations of these assessment instruments. Such a set of standards, particularly one that is succinct but considers a wide range of issues in a logical sequence, might also provide a framework the professional community might use for differentiating potentially accurate tests from ones that are misleading.

On the basis of consultations with other professionals and his reading of the literature on ATSRs and test construction, the author has formulated a set of standards that speaks to this issue. These standards are listed in the Appendix. Since they are offered as a starting point for collegial discourse, they will undoubtedly undergo revision and expansion over time. Still, they would seem to be useful at the present time in a couple of respects. For one thing, they offer at least a provisional framework that ATSR developers might consider in discharging their ethical responsibility to “use scientific procedures and current professional knowledge for test design, standardization, validation, reduction or elimination of bias, and recommendations for use” (American Psychological Association, 1992, p. 7). For another, potential ATSR users might also refer to them in discharging their responsibility to “have a clear rationale for the intended uses of a test ... in terms of its validity and contribution to the assessment and decision-making process” (American Psychological Association, 1999, p. 114).

Overall, a plausible argument might be made that the timeframe for the dissemination of ATSRs has been driven to a significant extent by the needs of the criminal justice system and that the importance of conducting research based on scientific principles of test construction has, as a result, been downplayed. The mounting criticism of ATSRs indicates that an adjustment of this value system is needed so that test construction principles are once again accorded the emphasis they deserve. It is hoped that the formulation of these special guidelines will facilitate this transition.

References

- American Psychological Association (1985). Standards for educational and psychological testing. Washington, D.C.: Author.
- American Psychological Association (1992). Ethical principles of psychologists and code of conduct. Washington, D.C.: Author.
- American Psychological Association (1999). Standards for educational and psychological testing. Washington, D.C.: Author.
- Anastasi, A. (1988). Psychological testing. New York: Macmillan.
- 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. (February/March 2001). Actuarial scales and evidentiary reliability. Sex Offender Law Report, 19.
- Covington, J. R. (October 1997). Preventive detention for sex offenders. Illinois Bar Journal, 166-175.
- Cronbach, L. J. (1970). Essentials of psychological testing. New York: Harper & Row.
- Dawes, R. M., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. Science, 243, 1669-1674.
- Donaldson, T. S. (2000). Accuracy of the Static-99. Unpublished manuscript.
- Epperson, D. (2000). Description of the MnSOST-R cross-validation sample. Available: <http://psych-server.iastate.edu/faculty/epperson/crossvalidationupdate.html>
- Epperson, D., Kaul, J., & Hesselton, D. (1999). Minnesota Sex Offender Screening Tool – Revised (MnSOST-R): Development, performance, and recommended risk level cut scores. Unpublished manuscript.

Epperson, D. L., Kaul, J., & Huot, S. (1995). Predicting risk of recidivism for incarcerated sex offenders. Paper presented at the annual ATSA conference, New Orleans, LA.

Epperson, D, Kaul, J., Huot, S., Hesselton, D., Alexander, W., & Gilman, M. (2000). Cross-validation of the MnSOST-R. Paper presented as the annual ATSA conference, San Diego, CA.

Gottfredson, S. D. (1987). Prediction: An overview of selected methodological issues. In D. Gottfredson & M. Tonry (Eds.), Prediction and classification: Criminal justice decision-making (pp. 21-51). Chicago: University of Chicago Press.

Grove, Zald, Lebow, Snitz, & Nelson (2000). Clinical versus mechanical prediction: A meta-analysis. Psychological Assessment, 12, 19-30.

Hanson, R. K. (1998). What do we know about sex offender risk assessment? Psychology, Public Policy, and Law, 4, 50-72.

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

Heilbrun, K. (1992). The role of psychological testing in forensic assessment. Law and Human Behavior, 16, 1992.

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

LaFond, J. Q. & Winick, R. (1999). Sex offenders and the law. Psychology, Public Policy, and Law, 4, 3-24.

- Ohlin, L. (1951). Selection for parole. New York: Russell Sage Foundation.
- Otto, R. K., Borum, R., & Hart, S. (2001). Professional issues concerning the use of actuarial tests in sexually violent predator evaluations. Unpublished manuscript, University of South Florida.
- Otto, R. K. & Petrila, J. (in press). Admissibility of testimony based on actuarial scales in Sex offender commitments. Sex Offender Law Report.
- Petrila, J. & Otto, R. (in press). Admissibility of expert testimony in sexually violent predator proceedings. In A. Schlank (Ed.), The sexual predator. Kingston, NJ: Civic Research Press.
- Pedhazur, E. J. (1982). Multiple regression in behavioral research. New York: Holt, Rinehart, and Winston. (1982).
- Quinsey, V. L., Harris, G., Rice, M. & Cormier, C. (1998). Violent offenders. Washington, D.C.: American Psychological Association.
- Rice, M. R. & Harris, G. (1997). Cross-validation and extension of the Violence Risk Appraisal Guide for child molesters and rapists. Law and Human Behavior, 21, 231-241.
- Rogers, R. (2000). The uncritical acceptance of risk assessment in forensic practice. Law and Human Behavior, 24, 595-605.
- Simon, F. H. (1971). Prediction methods in criminology. London: Her Majesty's Stationery Office.
- Thorndike, R. L. (1982). Applied psychometrics. Boston: Houghton Mifflin.
- Wilkins, L. T. (1969). Evaluation of penal measures. New York: Random House.
- Wollert, R. W. (2000). The value of archiving actuarial data from commitment cases for analyzing test validity, expert judgment, and average sexual recidivism risk. Unpublished manuscript.

Footnotes

¹ Dawes, Faust, & Meehl (1989) succinctly describe this method, stating that “to be truly actuarial, interpretations must be both automatic (that is prespecified or routinized) and based on empirically established relations” (p. 1668).

² A consensus on this point exists among the developers of ATSRs, experts on the construction of actuarial tests in general, and authorities on psychological testing. For further discussion on the importance of cross-validation, the interested reader is referred to such varied sources as the American Psychological Association (1999), Anastasi (1988), Dawes, Faust, and Meehl (1987), Gottfredson (1987), Hanson (1998), Pedhazur (1982), Rice and Harris (1997), Simon (1971), Thorndike (1982), and Wilkins (1969).

Authors Note

Requests for reprints and correspondence concerning this article should be sent to the author at the Institute for the Study of Mandated Treatment, 1220 S.W. Morrison St. #929, Portland, OR 97205. The telephone number for the Institute is (503) 241-0466.

Table One

Experience Table for the MnSOST-R Developmental Sample

A. Risk Groups	B. Risk Group Scores	C. Number of Failures	D. Number of Successes	E. Total Number of Subjects	F. 6 -Year Recidivism Rates (C/E)
VI	≥ 13	14	2	16	.88
V	8 to12	26	15	41	.63
IV	4 to7	29	35	64	.45
III	0 to3	14	52	66	.27
II	-5 to -1	7	52	59	.13
I	\leq to -6	<u>0</u>	<u>10</u>	<u>10</u>	<u>0</u>
All		90	166	256	.35

Note: Successes and failures were compiled from the cumulative frequencies presented in the second and third columns of Table 4 presented by Epperson et al. (1999).

Table Two

Experience Table for the MnSOST-R Cross-Validation and Developmental
Sample (N = 351 Subjects)

A. Risk Groups	B. Risk Group Scores	C. Percent of All Subjects ^a	D. Number of Subjects (C x 351)	E. 6 -Year Recidivism Rate ^b	F. Number of Failures (D x E)	G. Number of Successes (D - F)
VI	≥ 13	.07	25	.70	18	7
V	8 to12	.17	61	.59	36	25
IV	4 to7	.24	85	.39	33	52
III	0 to3	.26	91	.20	18	73
II	-5 to -1	.21	75	.09	7	68
I	≤ to -6	.04	<u>14</u>	<u>0</u>	<u>0</u>	<u>14</u>
All			351	.32	112	239

^a From Col. 3 of "MnSOST-R Total Score Categories" (Epperson, 2000, p.2)

^b From Col. 4 of "MnSOST-R Total Score Categories" (Epperson, 2000, p.2)

Table Three

Prediction Table for the MnSOST-R Cross Validation Sample

A. Risk Groups	B. Risk Group Scores	C. Number of Failures	D. Number of Successes	E. Total Number of Subjects	F. Recidivism Rates (C/D)
VI	≥ 13	4	5	9	.44
V	8 to 12	10	10	20	.50
IV	4 to 7	4	17	21	.19
III	0 to 3	4	21	25	.16
II	-5 to -1	0	16	16	0
I	\leq to -6	<u>0</u>	<u>4</u>	<u>4</u>	<u>0</u>
All		22	73	95	.23

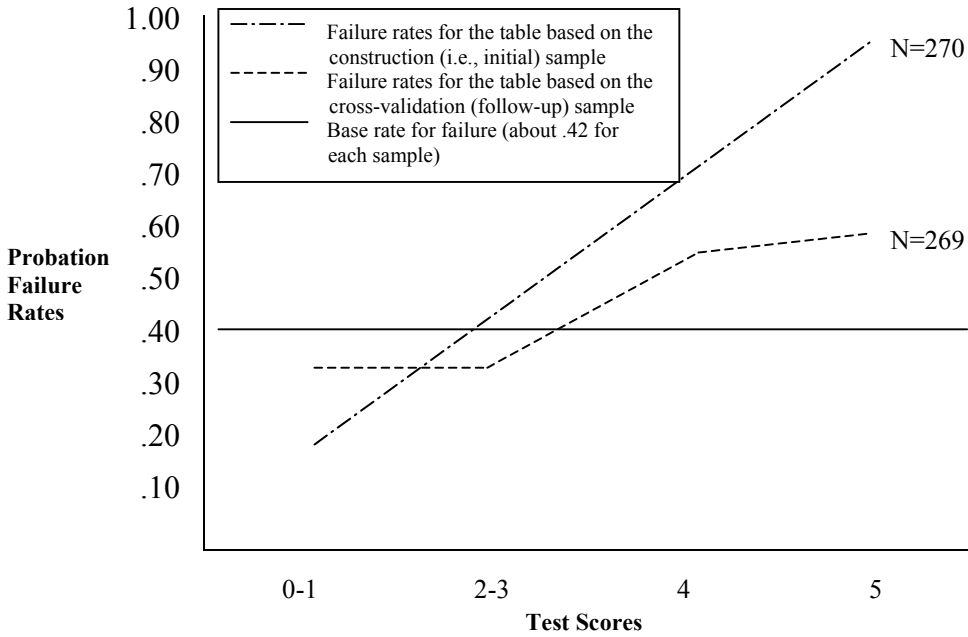


Figure 1 An Illustration of the Importance of Cross-Validation for Identifying Inflated Risk Estimates and Shrinkage. (Source: Simon, 1971, p. 109)

Appendix

A Framework of Standards for Evaluating the Adequacy of ATSRs

1. The purpose of the test should be stated clearly.
2. Some criterion measures should reflect only sexually violent recidivism.
Consideration should be given to developing criterion measures of a continuous nature (e.g., different levels of sexual dangerousness) in addition to dichotomous measures (e.g., recidivism versus nonrecidivism).
3. The perspective used to specify the content of the initial pool of possible predictor items should be described and justified in relation to achieving the test's purpose.
4. The initial item pool should be scored for a large and representative developmental sample drawn from the population or populations to which the test is to be applied.
5. Data should be reported that describe subject attrition rates and missing data. Steps should be taken to avoid or correct attrition and missing data problems.
6. Predictor items should be selected from the item pool so that adequate consideration is given to minimizing problems associated with significant inter-item correlations.
7. A formula, or test, for predicting each criterion measure should be developed on the basis of analyzing the item pool for the developmental sample.
8. Adequate inter-rater reliability for scoring both the test items and the criterion items on the developmental sample should be obtained.
9. A reasonably strong correlation between the test and its criterion should be obtained.

10. The standard error of measurement (SEM) should be calculated.
11. The SEM should be used to determine the "true" test score interval associated with a 90% level of confidence for each "observed" score that defendants might obtain. Tests with wide confidence intervals should be revised.
12. If subjects are assigned to risk groups, the efficiency of the test for subdividing subjects into distinguishable risk groups should be determined.
13. The failure rate for each risk group should be calculated and compiled in a preliminary experience table.
14. Cutting scores should be determined for classifying defendants as "likely" and "highly likely" recidivists.
15. The predictive accuracy of the test at each cut point should be formulated in terms of the rate of false positives, false negatives, true positives, and true negatives. Specificity, sensitivity, and overall test accuracy should also be calculated.
16. The predictive accuracy of the test at each cut point should be compared against the accuracy achieved by simply predicting the failure baserate for each defendant.
17. A cross-validation study should be conducted, administering the original pool of items to a fresh sample of subjects (see standards 4-16).
18. The results of the cross-validation study should be fully reported (see standards 5-16).
19. The original test should be revised if the results with the cross-validation sample differ markedly from the results with the developmental sample.

20. A true prediction table should be compiled if the cross-validation study replicates the findings with the developmental sample.

21. Inter-rater reliability for scoring the test items in an applied setting, using defense and prosecution experts as raters, should be determined and reported. If poor inter-rater reliability and wide confidence intervals are obtained, appropriate steps should be taken to resolve this problem.
 22. Any other research that is needed prior to disseminating the test should be conducted.
 23. The development of the test should be described in an article accepted by a journal that makes publication decisions based on a peer-review system.
 24. Once the above standards are met, a formal manual should be compiled that summarizes whatever information a user might find useful for evaluating the adequacy of a test, administering it, and interpreting its scores.
-