

Running head: Flaws in the MnSOST-R

Additional Flaws in the Minnesota Sex Offender Screening Tool - Revised:

A Response To Doren and Dow (2002)

Richard Wollert, Ph.D.

Institute for the Study of Mandated Treatment

Correspondence concerning this article should be addressed to Richard Wollert, Ph.D., Licensed Clinical Psychologist, Institute for the Study of Mandated Treatment, 1220 SW Morrison St., #930, Portland, OR 97205; e-mail: rwwollert@aol.com; 503.757.7712 (OR) or 360.737.7712 (WA).

Abstract

Wollert's (2002) paper on actuarial tests that are used for the prediction of sexual recidivism (ATSRs) showed that the recidivism rates for the score groups in the MnSOST-R developmental sample were inflated, and that they "shrank" by as much as 44 percentage points on cross-validation. Doren and Dow (in press) present an analysis of Wollert's (2002) research and an alternative study that they claim demonstrated a lack of shrinkage. This commentary presents a critique of Doren and Dow's response. Various considerations indicate that the original cross-validation was appropriate. Doren and Dow's alternative was also found to be unacceptably flawed. Additional analyses cast doubt on the MnSOST-R's validity coefficients and its relevance for identifying likely recidivists, and a consensus was apparent that the MnSOST-R experience table for the developmental sample should not be used in civil commitment hearings. Finally, a number of guidelines for the development of ATSRs were reviewed. In particular, it was noted that test developers could make a valuable contribution to the field by sharing data about their tests with other professionals when asked to do so.

Keywords: Actuarial tests, civil commitment, cross-validation, ethical principles for test construction, Minnesota Sex Offender Screening Tool – Revised (MnSOST-R), sexually violent predators.

Additional Flaws in the Minnesota Sex Offender Screening Tool – Revised:

A Response to Doren and Dow (2002)

The MnSOST-R (Epperson, Kaul, & Hesselton, 1999; Epperson, Kaul, & Huot, 1995; Epperson, Kaul, Huot, Hesselton, Alexander, & Gilman, 2000) has been under construction since 1991. Its developers have yet to publish a report on its psychometric properties in a peer-reviewed scientific journal, however. Nonetheless, expert witnesses have repeatedly used the recidivism rates from the original MnSOST-R experience table to estimate the chances that candidates for civil commitment as sexually violent predators (SVPs) will recidivate.

Some time ago I hypothesized that such estimates were inflated because a) the overall recidivism rate of 35 percent for the developmental group of 256 subjects, hereafter referred to as the D-256 sample, was inflated and b) the experience table for the MnSOST-R was distributed without first being verified in a cross-validation study. An opportunity to test this hypothesis arose in 2000 after Doug Epperson, lead researcher for the MnSOST-R task force, posted data for a 6-group cross-validation sample of 95 subjects (the CV-95 sample) on the internet. Comparing the experience table for the developmental sample with that of the cross-validation sample, a couple of patterns were isolated that confirmed my hypothesis (Wollert, 2002). First, a decrease, or “shrinkage”, in recidivism rates – ranging from .13 to .44 - was apparent for each of the five score groups where such a decrease might be observed. Second, whereas the recidivism rates for the two highest score groups in the D-256 sample were greater than 50%, the recidivism rate did not reach this level in the CV-95 sample for either of these groups.

A couple of recommendations were suggested by this analysis. One was that accepted principles of test construction, particularly those that have been formulated by the American Psychological Association (1985, 1999), should be followed when actuarial tests for the prediction of sexual recidivism (ATSRs) are developed. Another was that the MnSOST-R should not be used in civil commitment hearings because its recidivism estimates are inaccurate and it apparently does not identify any score groups that are associated with a recidivism rate above 50%. A study that found the MnSOST-R generates risk estimates that are 40 percentage points higher than other ATSRs (Wollert, 2000) was

discussed as further evidence of the MnSOST-R's lack of validity.

Doren and Dow (in press) argue that I should have compared two different experience tables: one from a cross validation sample of 220 subjects, or the CV-220 sample, and another that was estimated in such a way that a 21% recidivism rate was obtained. Undertaking such an analysis, they conclude that they could not find any evidence of shrinkage. They also set forth a variety of assertions, assumptions, implications, and conclusions about the MnSOST-R and cross-validated research that bear further discussion. The remaining sections of this paper constitute such a response.

Response to Doren and Dow

In many sections of their paper, Doren and Dow discount or trivialize the results of the original cross-validated analysis. To put these comments in perspective, I vigorously stress several points.

A statistically significant reduction in group-wise recidivism rates was found between the D-256 and CV-95 samples in the original cross-validation study. By posing the question "What Shrinkage?" as the title of their comment, by noting that statistical tests were not reported in my original paper, and by objecting that it was "statistically inappropriate" to evaluate the MnSOST-R by comparing only high risk groups, Doren and Dow imply that the experience table for the D-256 sample did not differ from the experience table for the CV-95 sample.

Additional statistical analyses show that such a view is untenable. For example, a test of proportional differences was calculated that compared the recidivism rates for the two highest score groups from each sample (see Tables 1 and 3 from the original Wollert study). This comparison was particularly appropriate because only the top two score groups in the D-256 experience table, associated with recidivism rates above 50%, were relevant for the topic of concern (SVP hearings). The rates for the CV-95 groups were significantly less than those for the D-256 groups ($z=2.02$, $p<.05$, two-tailed test). Although Doren and Dow argue that a z -test was inappropriate because of the categorical nature of the data and claim the test no longer recommended, these objections are groundless inasmuch as the selected test is used specifically for analyzing "data on a logically dichotomous variable" (Bruning & Kintz, 1977, p. 222).

“Shrinkage” was also apparent for each of the five score groups in the CV-95 sample where this phenomenon might be observed ($p < .05$). The unreliable nature of the original MnSOST-R experience table was therefore evident for particularly relevant groups and for all groups on the scale that were associated with any risk of recidivism.

Rather than being trivial, the degree of shrinkage reported in the original cross-validated analysis holds important implications for the validity and utility of the MnSOST-R. Doren and Dow suggest that the findings reported in my article were trivial because they merely confirmed the “statistical expectation” that “if one very significantly . . . decreases the overall proportion of recidivists within a group of offenders, that the recidivism percentages in subcategories (i.e., risk categories) might also . . . go down”. This line of criticism is unjustified in that it wrongly assumes that professionals are keenly aware that the magnitude of recidivism rates in a sample can have a huge effect on a test’s predictive power. On the contrary, sensitizing professionals to this issue has been an ongoing intellectual campaign (Meehl & Rosen, 1955; Janus & Meehl, 1997; Simon, 1971; Stott, 1960; Walters, 1955-56). The results of the original cross-validation are therefore not trivial. This, in turn, means that the major implication of the cross-validation – that risk predictions based on the MnSOST-R are too unreliable to be useful in civil commitment cases – bears serious consideration.

Other sections of the target response attempt to put research on ATSRs in perspective and conceptualize cross-validated theory as it relates to the initial study. The following observations are relevant to these issues.

ATSRs are psychological tests and must satisfy the standards pertaining to such tests. Throughout their article, Doren and Dow evaluate the MnSOST-R by applying concepts (e.g., cross-validation, inter-rater reliability, predictive validity, concurrent validity) that are used to evaluate psychological tests (APA, 1985, 1999). Doren has previously argued, however, that ATSRs are not psychological tests and “psychological test standards do not apply” (Doren, August/September 2000, p. 66). By evaluating the MnSOST-R and any other tests that they truly believe are not psychological tests, Doren and Dow agree – or apparently agree more than they used to – with the thesis that ATSRs are

psychological tests (e.g., see Anastasi, 1988; Cronbach, 1970; Turner, DeMers, Fox, & Reed, 2001). Perhaps their views might change even more in this direction if they were to consider that the prediction of recidivistic behavior is a psychological concept and that psychologists who are expert witnesses may be ethically obligated to avoid the use of the MnSOST-R if they truly believe it is not a psychological test.

A systematic bias is not introduced when findings from a sample with an inflated base rate are compared with findings from a sample with a naturalistic base rate. Doren and Dow advance the views that a “systematic bias” is introduced when samples such as the D-256 are used in cross-validated analyses and that “proper cross-validation analyses assume that the development sample and the cross-validation sample are drawn in a similar fashion”. Both views run counter to accepted scientific beliefs and practices. Numerous publications by test experts, for example, were referenced in my article that described cross-validation. These publications consistently indicated that it is proper in cross-validated analysis to compare data from a developmental sample, artificially inflated or not, with “fresh” data from a representative sample. Furthermore, although not discussed in my article, comparing developmental and follow-up data in cross-validated research has long been recognized as a valuable method of correcting, not introducing, biases in developmental studies with inflated recidivism rates (Stott, 1960; Walters, 1955-56; Simon, 1971). Finally, I have not come across a paper published by a test expert that claims developmental data should be excluded from cross-validated analyses when the developmental and cross-validation groups are not drawn in the same way or portrays developmental versus follow-up comparisons as a source of spurious or systematically biased research results. Doren and Dow do not offer a logical argument, an example, or a citation that contradicts this foundation or supports their assertions. Their views therefore seem idiosyncratic and ad hoc in nature. They are also misleading in that they do not accurately portray cross-validated theory or the history of cross-validated practices.

Including the D-256 sample in the initial cross-validated analysis was entirely correct and proper. Doren and Dow argue that data for the D-256 sample should not have been included in the initial cross-validated analysis. This assertion represents a specific application of the theoretical position

evaluated in the foregoing section. Since this position is untenable, so are its applications. Furthermore, as noted in the previous section, test experts have consistently indicated that original data, such as that for the D-256 sample, should be compared with “fresh” data. Why? Because this type of comparison is invaluable for determining what Meehl and Rosen (1955) have referred to as “shrinkage ... in the cross validation of cutting lines” (p. 205). So inclusion of the D-256 sample was entirely correct and proper.

Still other sections describe the approach that Doren and Dow adopted for carrying out their analysis. Four criticisms point up flaws in their methodology and interpretation of their results.

An estimated experience table should not have been used in lieu of real developmental data for the purpose of clarifying shrinkage questions. Doren and Dow propose that the most adequate test for shrinkage revolves around a comparison of cross-validated data with “the risk percentages from the statistically extrapolated 21% recidivism base rate sample”. Described so euphemistically, the “21% recidivism base rate sample” sounds, at first glance, like a plausible comparison group. According to the MnSOST-R developers, however, this data set was derived by taking the MnSOST-R developmental sample and simply “multiplying the number of non-recidivists at each MnSOST-R score (group) by a factor of 2” (Epperson, Kaul, & Hesselton, 1999, p. 18). These procedures indicate that what Doren and Dow refer to as a “21% recidivism base rate sample” does not constitute an actual sample in that it was derived from unconfirmed hypothetical data rather than a specified number of real subjects. Furthermore, in describing their strategy for deriving this table, hereafter referred to as the MnSOST-21, the test developers warned that relying on their estimated data would be “inappropriate for inferential statistics” (Epperson, Kaul, & Hesselton, 1999, p. 18). Consequently, the results of the statistical analyses reported by Doren and Dow are invalid.

The results obtained by Doren and Dow do not replicate the estimated experience table for the MnSOST-21. Doren and Dow disclose that the artificial nature of the MnSOST-21 was not an “issue” for them because their results “replicated” the table. Furthermore, they claim that this was their “sole point”. If so, by suggesting that test users should equate a recidivism rate of .78 with one of either .44 or .54 (the last two paragraphs of section two refer to the tables where these numbers can be found), their sole point

challenges everyday commonsense. It also challenges statistical commonsense. Doren & Dow's notion that their alternative analysis amounted to a successful replication study, for example, is untenable because they should not have applied statistical tests to the figures from the estimated experience table in the first place (see the foregoing subsection). In addition, the "power" (Pagano, 1986, pp. 199-214) of the tests they did apply to the highest risk groups – the ones showing the largest differences in recidivism rates - was very low because of their nonparametric nature and the small number of subjects for whom data were analyzed. Under such circumstances, it is misleading to use statistical tests for the purpose of claiming that the distributions of two groups with grossly different results are somehow the same. This type of mistake, which is widely discussed in relation to hypothesis testing, is known as a "Type II error" (McCall, 1975; Pagano, 1986).

The CV-220 sample analyzed by Doren and Dow does not provide a superior estimate of the recidivism rate for subjects with the highest risk scores. Doren and Dow acknowledge that I contacted Dr. Epperson before my paper was published to obtain cross-validated data for the CV-220 sample, but that the data proved to be elusive. Nonetheless, they imply that a more accurate estimate of the recidivism rate for those with the highest test scores could have been calculated by pooling the data for the CV-95 sample with data from a second cross-validation sample of 125 subjects, or the CV-125 group. This would have been appropriate only if similar patterns of recidivism were apparent across samples. This was not the case in that 75% of those in the top group of the CV-125 sample recidivated, compared to only 44% from the same group in the CV-95 sample.

Although it is argued or implied in several places in their paper that the distributions for the CV-95 and CV-125 samples are the same, this position is unjustified: as the previous section indicates, the power of the tests that Doren and Dow applied to the highest score groups was inadequate to determine whether the samples were truly different. Furthermore, the original cross-validation study was published before the CV-125 data set were publicly distributed. Dissemination of the cross-validation study may therefore have been followed by a change in the way data were coded for the CV-125 subjects compared to the CV-95 subjects. In light of this possibility, the most defensible research approach consists of

treating these data sets as though they come from different distributions.

The statistical analyses conducted by Doren and Dow do not justify their conclusion that a “lack of shrinkage” was evident in their alternative study. At several points, Doren and Dow claim that a series of nonsignificant chi-square tests were sufficient to conclude that “no ‘shrinkage’ was found” in their alternative study. This conclusion is unjustified because, as has been pointed out, it was inappropriate to apply statistical tests to the MnSOST-21 and also because the chi-square tests that were applied were too insensitive to identify group differences.

After presenting their analysis, Doren and Dow discuss a number of issues that relate to the validity of the MnSOST-R. The first of the following comments focuses on the topic of convergent validity, while the last three explain how subject deselection and compilation of a nonrepresentative sample have undermined research on the MnSOST-R so that its generalizability is low and its validity coefficients are uninterpretable.

The fact that the MnSOST-R generates estimates that are 114% higher than the estimates of other actuarial tests that are used to predict sexual recidivism is further evidence that it is invalid. Doren and Dow offer a couple of explanations as to why risk estimates for the MnSOST-R were found to be much higher than other ATSRs (Wollert, 2000). Their explanations lack merit, however, because: a) the study with which they were concerned included two ATSRs that, like the MnSOST-R, predicted sex offense re-arrest rates; and b) test theory holds that tests that predict the same construct should converge on similar results (American Psychological Association, 1999, p. 14). Even Doren and Dow, in the fourth paragraph of their fifth section, recommend the use of a “multidimensional model ... that essentially compares the results from any single risk assessment instrument with other measures of the same dimension of risk, for the purpose of bolstering the confidence the evaluator can have”. From this perspective, when only the predictions of the MnSOST-R are high in a battery of ATSRs, as in the Wollert (2000) study, it is most likely that the MnSOST-R is inaccurate and stands in need of recalibration. Doren and Dow’s explanations may offer some recalibration directions, but the MnSOST-R’s apparent lack of convergent validity is undeniable.

The D-256 sample was nonrandom and unrepresentative. Doren and Dow state that the MnSOST-R developmental sample included those who were re-arrested for a new sex offense during a 6-year follow-up period and a random selection of releasees who were not. Other sampling procedures are also worth keeping in mind. According to the test's developers (Epperson, Kaul, & Hesselton, 1999), the initial sample was made up of 371 releasees. Sixteen recidivists were added to this sample. Then 113 offenders who sexually abused family members in a way that "did not involve penetration or high degrees of force" (p. 6), were eliminated. Another 18 subjects were eliminated because of missing data. These procedures created a nonrandom sample that was unrepresentative of the population from which it was selected.

Elimination of familial offenders from the MnSOST-R sample reduces its relevance for civil commitment cases. Doren and Dow argue that the MnSOST-R is relevant for assessing rapists and molesters who may be involved in civil commitment hearings, but not incesters and other familial offenders. In actuality, the procedures that were used to select the MnSOST-R sample decrease the test's relevance for all categories of potential commitment candidates in some states. In Washington, for example, the statute that specifies the characteristics of sexually violent predators (RCW 71.09.020) excludes only a small subset of incarcerated familial offenders - those convicted of incest against a person over the age of 14 - from the population of potential commitment candidates (PCCs). In contrast, over 30% (i.e., $113/371 = .30$) of the subjects in the MnSOST-R developmental sample were eliminated because they were familial offenders. Undoubtedly, many offenders were eliminated who were convicted of crimes - such as rape of a child or second degree child molestation - that qualify offenders as PCCs in Washington. Furthermore, offenders who committed crimes that fall outside the scope of Washington's SVP laws - such as third degree rape or child molestation - were probably added. These eliminations and additions virtually guarantee that the spectrum of sex offenders in the MnSOST-R developmental sample differs substantially from the spectrum of Washington PCCs. In view of the dissimilarity between the "local" (i.e., Washington) population and the "developmental" (i.e., Minnesota) sample, it is unlikely that the tables for either the MnSOST-R or the MnSOST-21 accurately identify the recidivism rates for

rapists, molesters, or incesters who are subject to Washington's SVP laws. This situation exemplifies the reasons why testing guidelines hold that tests are invalid unless their scores are interpretable on the basis of "evidence ... for populations similar to those representative of the examinee" (American Psychological Association, 1999, p. 134). The most direct way to overcome the MnSOST-R's lack of generalizability in cases such as this one is to collect local recidivism data on the population to which the MnSOST-R is to be applied. This has not been done in Washington, however, or other states with SVP laws.

Validity coefficients for the MnSOST-R are uninterpretable because they do not take into account the effects of measurement error associated with the elimination of familial offenders. Doren and Dow cite several studies they claim attest to the MnSOST-R's validity. This evidence is weak, however, in that only three of the studies have been published or accepted for publication, and two of these (Barbaree, Seto, Langton, & Peacock, 2000; Bartosh, Garby, Lewis, & Gray, in press) reported negative validity findings. Some of the statistically significant correlations and AUROC coefficients they discuss extensively as replication evidence, such as those reported by Langton (2002), also display a large amount of shrinkage when compared to the developmental sample coefficients.

Another flaw is that all validity coefficients for the MnSOST-R are open to question because previous studies have eliminated familial offenders. This type of deselection raises many difficult sampling questions. Is someone a familial offender if he has sexual contact with his 14 year-old daughter? His four year-old niece? The 16-year-old daughter of his best friend? Suppose the victim was his best friend's 16-year-old son? What about those who had sex with their teenage daughters, but have a prior history of convictions for sex abuse outside the family or rape? Different researchers at different sites will answer these questions differently. This is a source of unreliability and measurement error that has never been controlled. Until this step is taken, none of the validity coefficients discussed by Doren and Dow can be accepted as definitive. Nor can their view that high recidivism rates are attributable to the exclusion of only incest offenders (eighth paragraph of section three) – the MnSOST-R obviously excludes many other types of offenders as well.

Elsewhere, Doren and Dow launch a “strawman argument” over this issue by wrongly stating that I believe the MnSOST-R’s validity coefficients are open to question because the test is applicable only to “extrafamilial offenders”. Then they claim my supposed opinion is unjustified because the studies they reviewed were “conducted with this limitation in mind”. Both points need correction. My overarching concern does not revolve around the envisioned scope of the MnSOST-R, but the extent to which the test falls short of achieving this goal because of inadequate subject selection procedures. Also, in contrast to what Doren and Dow imply, research studies that take this source of measurement error into account in the calculation of validity coefficients have not been undertaken. When they are, confidence intervals for the MnSOST-R may be found to be much wider than what has been reported. This would further underscore the test’s inadequacy for identifying a lowest plausible recidivism estimate for civil commitment candidates that exceeds the commitment standard.

Finally, Doren and Dow make a number of suggestions regarding the use of the MnSOST-R in the future. My reactions to these suggestions are as follows.

Experts should not use the original MnSOST-R prediction table based on the developmental sample having a 35% recidivism base. Doren and Dow register their agreement with my conclusion that the experience table based on the D-256 sample should not be used to predict sexual recidivism risk for civil commitment candidates. Against the backdrop of our other differences, this would seem to represent the initiation of some movement towards a commonly shared point of view.

Experts who follow Doren and Dow’s advice to use “risk percentages based on the 21% sample” to predict sexual recidivism do so at their own risk. As noted earlier, the term “sample” refers to an existing set of data compiled for a specified number of real subjects. This is not the case for the table of numbers that make up the MnSOST-21. Experts who are thinking about following Doren and Dow’s advice might first give some thought to the fact that adequate validity information with respect to these numbers is unavailable. They might also gauge their enthusiasm for explaining to the court the inconsistency of purporting to render an actuarial prediction by using a table based on systematic guessing rather than empirical research. A wide range of latitude is available when it comes to the

structure of evaluations, but it seems to me that those who continue to use any existing version of the MnSOST-R are just asking for trouble.

Doren and Dow's observation that "there can be reasonable concern about using the MnSOST-R outside of Minnesota" is empirically sound. No cross-validation studies have been completed which show that the MnSOST-R prediction tables are generalizable to states outside of Minnesota. Langton (2003), in fact, recently found that the shrinkage for the two highest MnSOST-R score groups of a large sample of sex offenders from Ontario (Canada) was even greater, by 18 to 28 percentage points, than what I reported. Furthermore, the idiosyncratic nature of the MnSOST-R sample and the lack of control over subject selection effects makes it virtually impossible to accurately estimate the generalizability of almost any of the test's validity coefficients. Doren and Dow's doubt about the use of the MnSOST-R outside Minnesota is therefore completely realistic. The only amendment I would add is that the MnSOST-R's flaws should arouse reasonable concerns about using the MnSOST-R within the state as well.

Discussion

This paper has critiqued Doren and Dow's (in press) response to an article which showed that initial risk estimates for the MnSOST-R were greatly inflated (Wollert, 2002). The points raised in the critique support the view that the comparison described in the original article constituted the most adequate cross-validated study that could have been performed under the circumstances. Doren and Dow's alternative study was also unacceptably flawed, and their assertion that the MnSOST-R is characterized by a lack of shrinkage is contradicted by studies of both Minnesota and non-Minnesota sex offenders. Finally, additional analyses cast further doubt on the test's relevance and validity coefficients. In light of these conclusions, Doren and Dow's claims that the results of research "have been consistent ... in support of the instrument's validity" (last paragraph, fourth section) and that a consensus exists that the MnSOST-R "consistently works for the purpose it was designed" (last paragraph, last section) are unwarranted and misleading.

Agreement was apparent, however, on at least two issues. First, guidelines that are used for

evaluating the validity of psychological tests should be used to evaluate the validity of actuarial tests. Second, an important limitation of the MnSOST-R was evident in the shared observation that reasonable concern/uncertainty exists over using the MnSOST-R outside Minnesota. Third, convergent recommendations that the original experience table based on the D-256 sample should not be used in civil commitment cases underscored the advisability of placing restrictions on the use of the MnSOST-R.

Other restrictions should be put in place. As has been noted, no justification exists for using the MnSOST-21 to predict sexual recidivism. Furthermore, troubling patterns in the data for the CV-125 sample leaves the prediction table for the CV-95 sample as the most reasonable predictive device. No risk group in this table, however, has a recidivism rate in excess of 50%. Even the highest rate in this table may be seriously distorted as a result of inadequate sampling methods. Taken together, these considerations point to the conclusion that the MnSOST-R should be completely withdrawn from use in civil commitment cases.

Although the MnSOST-R's performance does not recommend it, the history of the MnSOST-R's construction points to a number of practices (e.g., representative sampling, local validation, defining sexually violent recidivism precisely and relevantly, developing manuals that feature prediction tables for specific samples as well as pooled data) that should be adopted as new tests are developed. Requesting data that remained elusive was a singularly disappointing experience, however, in the case at hand. Having encountered similar experiences while attempting to examine other ATSRs, it seems that it might be helpful to gently, but with great concern, suggest that test developers bear a special ethical responsibility – given the momentous application of their tests - for sharing the data they've collected with their colleagues on request. A more open stance, in turn, would assist the professional community in determining which tests are valid and give experts the information they need in order to use them with confidence.

References

- American Psychological Association (1985). Standards for educational and psychological testing. Washington, DC: Author.
- American Psychological Association (1992). Ethical principles of psychologists and code of conduct. American Psychologist, *47*, 1597-1611.
- American Psychological Association (1999). Standards for educational and psychological testing. Washington, DC: Author.
- Anastasi, A. (1988). Psychological testing. New York: Macmillan.
- Barbaree, H. E., Seto, M., Langton, C., & Peacock, E. (2001). Evaluating the accuracy of six risk assessment instruments for adult sex offenders. Criminal Justice and Behavior, *28*, 490-521.
- Bartosh, D. L., Garby, T. Lewis, D., & Gray, S. (in press). Differences in the predictive validity of actuarial risk assessments in relation to sex offender type. International Journal of Offender Therapy and Comparative Criminology.
- Bruning, J. L., & Kintz, B. (1977). Computational handbook of statistics. Glenview, IL: Scott, Foresman and Company.
- Cronbach, L. J. (1970). Essentials of psychological testing. New York: Harper & Row.
- Doren, D. (August/September 2000). Evidentiary issues, actuarial scales, and sex offender civil commitments. Sex Offender Law Report, *1*(5), 65-66, 78-79.
- Doren, D. M. & Dow, E. (2002). What shrinkage of the MnSOST-R? A response to Wollert (in press). Journal of Threat Assessment.
- Epperson, D. (2000). Description of the MnSOST-R cross-validation sample. Available: <http://psych-server.iastate.edu/faculty/epperson/crossvalidationupdate>.
- 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 at the annual ATSA conference, San Diego, CA.

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.

Langton, C. M. (2003). Contrasting risk approaches to risk assessment with adult male sexual offenders. Doctoral dissertation: University of Toronto.

McCall, R. B. (1975). Fundamental statistics for psychology. New York: Harcourt Brace Jovanovich.

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

Pagano, R. R. (1986). Understanding statistics in the behavioral sciences. St. Paul, MN: West Publishing Company.

Simon, F. H. (1971). Prediction methods in criminology. London: Her Majesty's Stationery Office.

Stott, D. H. (1960). The prediction of delinquency from non-delinquent behavior. British Journal of Delinquency, 10, 195-210.

Turner, S. M., DeMers, S., Fox, H., & Reed, G. (2001). APA's guidelines for test user qualifications. American Psychologist, 56, 1099-1113.

Walters, A. A. (1955-56). A note on statistical methods of predicting delinquency. British Journal of Delinquency, 6, 297-301.

Wollert, R. (2000). The value of archiving actuarial data from commitment cases for analyzing test validity, expert judgment, and average sexual recidivism risk. Unpublished manuscript.

Wollert, R. W. (2002). The importance of cross-validation in actuarial test construction: Shrinkage in the risk estimates for the Minnesota Sex Offender Screening Tool – Revised. Journal of Threat Assessment, 2(1), 87 – 102.

RECEIVED: December 30, 2002

REVISED: April 14, 2003

ACCEPTED: April 17, 2003