Running head: Sexual Predators

# Low Base Rates Limit Expert Certainty When Current Actuarials Are Used To Identify Sexually Violent Predators: An Application of Bayes's Theorem

Richard Wollert

Vancouver, Washington

Published in *Psychology, Public Policy, and Law*, 2006, *12*, 56-85.

# Abstract

This paper applies Bayes's Theorem to age-wise sexual recidivism rates and the accuracy of high actuarial scores for predicting sexual recidivism in civil commitment cases. Recidivism rates consistently declined with age, paralleling the "age invariance" pattern found for other offenders. Furthermore, actuarials were only efficient for the youngest group, were inaccurate for identifying recidivists, and misclassified many non-recidivists as recidivists. Opinions about the accuracy of actuarials are therefore often wrong, and actuarials need to be reformulated. Finally, actuarials are useless for identifying likely sexual recidivists from populations with recidivism base rates below .25. Recommendations include seeking new trials in cases that overlooked age, focusing attention on young offenders, limiting commitment periods, and shifting resources from commitment centers to impact all offenders released to the community.

# Low Base Rates Limit Expert Certainty When Current Actuarials Are Used To Identify Sexually Violent Predators: An Application of Bayes's Theorem

Many states in the United States have enacted legislation allowing for the post-prison civil commitment of sex offenders as "sexually violent predators" (SVPs) (Covington, 1997; Doren, 2002; Miller, Amenta, & Conroy, 2005). As the first stage of this process, offenders thought to meet the commitment standards are referred by review panels to prosecutors for commitment consideration. Secondly, prosecutors decide if commitment petitions should be filed. Then, the courts determine if probable cause exists for evaluating the possibility that these offenders might be SVPs. Finally, the courts determine which respondents should be committed as SVPs because they suffer from a mental abnormality or personality disorder that makes them a) a menace to safety and b) likely to recidivate by committing new sexually violent predatory offenses.[1]

Decision making in SVP cases is fraught with uncertainty because SVP laws do not define all the terms they invoke and do not specify the time period that recidivism estimates should cover. Consequently, the judicial system relies heavily on the predictions of psychologists and other expert witnesses who have been trained in assessment techniques and scientific methods of hypothesis testing. From their graduate training these experts understand that, in connection with each SVP evaluation they undertake, they are essentially testing the "null hypothesis" that the respondent in question does not differ from non-predatory sex offenders who fall just below the commitment standard. They also understand they are justified in claiming, as the courts require, that they are reasonably certain the "alternative" or "research" hypothesis is correct, and that a candidate meets the commitment criteria, if evidence indicates that all relevant null hypotheses may be rejected at a high level of confidence.

Hypothesis testing within this context requires precision. Otherwise, many commitment candidates may be wrongly classified as predatory while many others may be wrongly classified as non-predatory (Wilkins, 1969). Attempting to avoid such “Type I” and “Type II” errors (McCall, 1975), respectively, experts often base their conclusions on “actuarial tests” (Wollert, 2002). As described in various sources, actuarial tests assign offenders to different “bins” or “risk groups” (e.g., 1, 2, 3, 4, 5, 6, 7, 8, or 9) on the basis of how they score on test items (e.g, number of previous sex offenses, marital status, etc.) that research has linked with recidivism (Beech, Fisher, & Thornton, 2003; Doren, 2002; Harris, Rice, & Quinsey, 1993; Janus & Prentky, 2004). An “experience table” has also been compiled for each test that indicates the percentage of offenders in each bin who have recidivated (e.g., 35%) during a specified period of time (e.g., 10 years).

When an expert uses an actuarial test to derive a single “average” risk estimate for an offender, she typically determines the number of points he is allocated on each item, totals the item scores to determine the bin in which the offender belongs, and consults the test’s experience table to locate the recidivism rate, or “risk,” for those in the offender’s bin. If she wishes to derive the interval between the offender’s average risk estimate and his lowest plausible estimate, she determines his lowest plausible test score by subtracting the “confidence interval” (Anastasi, 1988; Gulliksen, 1950) from his obtained test score, and then consults the experience table to locate the risk associated with his lowest plausible score. To determine the interval between an offender’s average risk estimate and his highest plausible estimate, she determines his highest plausible score by adding the confidence interval to his obtained score, and uses the experience table to locate the risk associated with this score.

Among the tests that are sometimes used in the above manner are the following:

1. Minnesota Sex Offender Screening Tool – Revised Version (MnSOST-R; Epperson, Kaul, & Hesselton, 1999).
2. Rapid Risk Assessment for Sex Offender Recidivism (RRASOR; Hanson, 1997).
3. Sex Offender Risk Appraisal Guide (SORAG; Quinsey, Harris, Rice, & Cormier, 1998).
4. Static-99 (Hanson & Thornton, 2000).
5. Violence Risk Appraisal Guide (VRAG; Harris et al., 1993).

Using actuarial tests for the prediction of sexual recidivism (ATSRs) has several advantages. One is that actuarial tests have consistently been found to be more accurate than "subjective" clinical judgment (Grove & Meehl, 1996), which is wrong from 72% to 93% of the time, for the purpose of making sexual recidivism predictions (Dix, 1976; Hall, 1988; Hanson, Morton, & Harris, 2003; Kahn & Chambers, 1991; Smith & Monastersky, 1986; Sturgeon & Taylor, 1980). Another is that the reliance of actuarials on rules for the interpretation of data holds out the possibility that the evaluation process – even for a group as stigmatized as sex offenders – will be objective and fair (Janus & Prentky, 2004). Yet another advantage of this method is that actuarials offer a way of examining the performance of the judicial system by determining whether the predicted failure rate for those who are committed is consistent with the standards of commitment, which remain un-quantified but have been discussed as falling in the range of 50 to 75 percent (Janus & Meehl, 1997; Wollert, 2002).

Unlike the typical situation described in the fourth paragraph, where the focus is on the derivation of risk estimates, an SVP evaluation requires the evaluator to determine whether the recidivism risk for a respondent is so high that the null hypothesis that he is unlikely to recidivate may be rejected to a reasonable degree of certainty. To use an actuarial for this purpose, an evaluator needs to select either a single test score or a range of test scores that she deems to be of critical importance for identifying likely recidivists. Because SVP predictions classify offenders into only two groups (will recidivate or will not recidivate), scores in the "alternate test range" below this "critical test range" are considered important for identifying likely non-recidivists. To be succinct, the letter "C" will be used in the remainder of this paper to refer to values in the critical test range and the letter "L" will be used to refer to values in the alternate test range.

Once an evaluator has selected the value of C she will use with an actuarial to select likely recidivists, she will be able to compile a 2 x 2 table from the test's full experience table that shows how many recidivists versus non-recidivists will be expected to have a score of C, and how many recidivists versus non-recidivists will be expected to have a score of L. Several measures may be calculated from this simple table that are useful for evaluating the test's performance in the sample on which it was developed and for estimating its performance in another group that has a different recidivism rate than the developmental sample. Among these measures are "P" (Meehl & Rosen, 1955), the sample-wise recidivism base rate, which is the proportion of the entire developmental sample who are recidivists; "Q" (Meehl & Rosen, 1955), the sample-wise non-recidivism rate, which is equal to 1 minus P; sensitivity (Baldessarini, Finklestein, & Arana, 1983; Metz, 1978; Rice & Harris, 1995), the hit rate (Fergusson, Fifield, & Slater, 1977; Rice & Harris, 1995), the true positive fraction (Metz, 1978; Rice & Harris, 1995; Zweig & Campbell, 1993), or "T", which equals the proportion of recidivists the test identifies for scores covered by C; 1-specificity (Rice & Harris, 1995; Zweig & Campbell, 1993), the false alarm rate (Fergusson et al., 1977; Rice & Harris, 1995), the false positive fraction (Metz, 1978; Rice & Harris, 1995; Zweig & Campbell, 1993), or "F", which equals the proportion of non-recidivists whose scores are covered by C and are thus mistakenly identified as recidivists; and, lastly, positive predictive power (Baldessarini et al., 1983; Rice & Harris, 1995), efficiency in detecting maladjustment (Metz, 1978; Meehl & Rosen, 1955), or "E", which reflects the percentage of the time that experts will be right in their predictions when they rely on a specified C. Inefficiency, or the percentage of time that experts will be wrong, is estimated by subtracting the efficiency index from 1.

Table 1 illustrates the inter-relationships among these concepts, citing figures obtained when an experience table for an ATSR, known as Static-99 was compiled from a sample of 1,086 sex offenders (Hanson & Thornton, 2000). Since P was .25 and Q was .75, the sample included 271 recidivists (1,086 x .25 = 271) and 815 non-recidivists (1,086 x .75 = 815). For a C of 6 to 9, the test was correct in identifying 67 recidivists but missed 204 other recidivists with scores below 6. Although it also correctly identified 753 non-recidivists with L scores of 0 to 5, it mistakenly flagged 62 non-recidivists as recidivists because they had scores of 6-9. The test was therefore accurate at T equal to .25 (67 / 271 = .25), and inaccurate at F equal to .08 (62/815=.08). As a result, E was equal to .52 (67 / (67 + 62) = 67 / 129 = .52).

Insert Table 1 about here

Since chance is reflected in an E of .50, an E of .52 suggests that experts will be right most of the time when they use Static-99 with a C of 6-9 to pick out likely recidivists. If a test is to be applied to a group other than the developmental sample, however, a high level of confidence is warranted only if it may be safely assumed that P for the "target" group is at least the same as the recidivism base rate for the developmental sample.

Figure 1, in which the height of the cells in Table 1 have been re-drawn to represent the number of subjects each includes, shows how and why E for Static-99 would drop if it were applied to a population with a low P. In this figure, the bar chart for Group 1 reflects the performance of Static-99 for the data presented in Table 1 - that is, a sample with P = .25 and Q = .75 . The bar chart for Group 2 shows how E for Static-99 would deteriorate in a civil commitment sample where P = .125 and Q = .875. For 1,086 commitment candidates (the same number of offenders as in the original Static-99 sample), Group 2 would include only half the recidivists in Group 1 (1,086 x .125 = 136), and many more non-recidivists (1,086 x .875 = 950). Making the usual assumption that T (.25) and F (.08) are the same across groups,[2] the absolute number of recidivists correctly classified by the test as probable re-offenders would go down dramatically (136 x .25 = 34) in Group 2. The absolute number of non-recidivists mistakenly flagged as probable recidivists would also go up (950 x .08 = 76). Since E may be calculated from frequency values by dividing the number of correct recidivism predictions by the number of both correct and incorrect recidivism predictions, E for Static-99 over the CTR would decrease, or "shrink," from 52% (67 / (67 + 62) = 67 / 129 =. 52) for Group 1 to 31% (34 / (34 + 76) = 34 / 110 = .31) for Group 2. In light of this large amount of shrinkage, a strong argument might be made that Static-99 was of limited value for assessing offenders from Group 2 because only 31% of the offenders with C scores of 6 to 9 would recidivate. In contrast, 69% of the predictions made by experts who selected this particular C for identifying likely recidivists would be wrong (1 - .31 = .69).

Insert Figure One about here

Although ATSRs minimize prediction errors and enable experts to quantify the level of confidence attached to their opinions, research on the relationship between advancing age and sexual recidivism suggests that current actuarials may be of limited value for identifying older offenders who are likely to sexually recidivate (Saari & Saari, August / September, 2002). The reason for this is that recidivism rates for offenders decline as offenders get older. Hanson (2002), for example, calculated recidivism rates for 4,763 sex offenders, of whom 94 percent had been incarcerated or confined. Finding that recidivism risk declined almost 4% per year as a function of age, he presented separate recidivism curves illustrating this decline for 3,751 rapists, molesters, and incest offenders who were subdivided into 9 age groups ranging from those who were very young to those who were very old. More recently he has presented additional evidence that "offenders older than age 50 at release" re-offended "at half the rate of … younger (less than 50) offenders (12% versus 26%, respectively, after 15 years)" (Harris & Hanson, 2004, p. 7). Parallel results for child molesters and rapists were reported by Nicholaichuk and Yates (2002).

Elaborating on Hanson's approach, Barbaree, Blanchard, and Langton (2003) found that age accounted for additional variance in recidivism rates after the effects of the RRASOR (Hanson, 1997), an ATSR incorporating a dichotomous age variable (below 25 years old versus above 24), were controlled. Furthermore, in spite of this age variable, the rate of decline in recidivism was still found to be 3% per year (Barbaree, personal communication, June 6, 2004).

This second line of research indicates that sexual recidivism rates decline with age for those with the highest ATSR scores. A third body of data confirms this conclusion. Milloy (December, 2003; e-mail to B. Hampton dated July 24, 2004), for example, undertook a follow-up study of released offenders who met “the standards for civil commitment petitions, but for whom no petitions were filed” (p. 1), and reported data indicating that a significantly larger proportion of those below the age of 50 (31% out of an n of 80) committed new felony sex offenses when compared to those over 50 (0% out of an n of 9). A few months later, Hanson circulated data for 1,997 sex offenders scored on Static-99 indicating that the 5-year sexual recidivism rate for those under 50 was .154 versus .088 for those 50 to 59 years old (Hanson, May 3, 2004). Analyzing these data further, Wollert (April, 2005) determined not only that these proportions differed ($z > 1.96$) but that the expected recidivism rate for younger offenders with high Static scores was .37, while it was .23 for middle-aged offenders with the same high scores.

A fourth stream of research on criminal behavior in general also indicates that the age-crime pattern reported by Hanson (2002) for sex offenders is virtually “invariant” among more inclusive criminal populations. Hirschi and Gottfredson (1983), for example, summarized many cross-sectional studies showing that crime rates decreased with age for offender groups who lived in different centuries, came from different countries, differed with respect to age and gender, and committed different types of crimes. In 2003, Sampson & Laub published a 70-year longitudinal study of 475 “serious, persistent delinquents” that controlled for both the effects of death and incapacitation. Not only did they find that violent crimes, including sex crimes, were infrequently committed by older offenders, but that the violent crime rate for offenders with high actuarial scores converged over time with the violent crime rate for low scoring offenders. Since it may be assumed that these studies included many sex offenders, they strongly suggest that sexual recidivism declines with age and that this decline may best be conceptualized as simply an extension of Hirschi & Gottfredson’s (1983) age invariance theory.

The limits of ATSRs for predicting sexual recidivism among older offenders hold significant implications for expert opinions in SVP cases in that a large percentage of respondents are older (Washington State Senate Committee on Human Services & Corrections, February 14, 2005). From the author's personal experience and consultations with colleagues and attorneys in a number of states, he is also aware of various unpublished and sealed proceedings where fact-finders have not civilly committed respondents on the basis of their age or have ordered new commitment trials because of evaluations stressing the importance of this factor. Furthermore, in at least one appellate proceeding it was concluded that "it is undisputed among sex offender experts that age is an important factor in determining risk of re-offense …" (In re Young, 2004).

In spite of these developments and the strength and consistency of the research on age and crime, some researchers (Doren, February / March, 2002; Thornton & Doren, 2002), experts, and attorneys have continued to argue that older respondents with high scores on actuarial tests are about as likely to recidivate as younger candidates with similar scores.

Fortunately, a clarifying evaluation of the validity of this position may be obtained by applying the method of analysis illustrated in Figure 1, known as "Bayes's Theorem," to recidivism data that are broken down by age. In general, Bayes's Theorem (Bayes, 1764) is a tool for assessing the probability that a theory – for example, that a person with heart disease will die in five years - is true when considered in light of the diagnostic accuracy (T and F) of some piece of evidence such as a test score and what is known about the overall, or "base rate," probability (P) of death for those most similar to the person. Estimates pertaining to the first informational category are typically called "data" or "evidence" probabilities while estimates of the second are referred to as "prior probabilities" (Iversen, 1984).

A fair amount of information about the evidence probabilities for ATSRs may be found in documents that have either reported experience tables or "receiver-operating characteristic" (ROC) curves" for ATSRs (Barbaree, Seto, Langton, & Peacock, 2001; Doren & Dow, 2003; Epperson et al., 1999; Hanson & Thornton, 2000; Harris, Phenix, Hanson, & Thornton, 2003; Harris, Rice, Quinsey, Lalumiere, Boer, & Lang, 2003; Langton, 2003; Wollert, 2002). Although ROC plots have many advantages for presenting a great deal of information about actuarials in an unbiased way (Fergusson et al., 1977; Mossman, 1994a; Mossman, 1994b; Rice & Harris, 1995; Swets, Dawes, & Monahan, 2000) and for determining the relative accuracy of different tests designed to predict the same outcome (Hanley & McNeil, 1982; Hanley & McNeil, 1983), they do not consider the effects of the magnitude of a disorder's base rate on test performance (Fergusson et al., 1977; Hanley & McNeil, 1982). As a result, the information they provide is insufficient for the purpose of determining test efficiency for any specific C. Bayes's Theorem, in contrast, is invaluable for this purpose, because it takes base rates into account while drawing on the estimates of T and F that are included in ROC plots and other sources. Therefore, rather than being an alternative to ROC analysis, Bayes's Theorem supplements it and extends its range of application.

When applied to sexual recidivism, Bayes's Theorem enables an evaluator to determine an "average" estimate (E) of the rate with which a class of offenders with high actuarial scores – in particular, those classified as likely sexual recidivists – will re-offend (Meehl & Rosen, 1955; Janus & Meehl, 1997). Conversely, it enables an evaluator to determine how often she will be wrong when she repeatedly uses an actuarial for identifying likely recidivists, since 1 minus the average estimate for likely recidivists equals the error rate. This type of analysis holds serious implications for opinions based on actuarial tests in that the credibility of using actuarials to predict recidivism in SVP cases is undermined by error rates above 50% (Wollert, 2002), and is demolished when they are far in excess of this standard.

Only three pieces of information are needed to apply Bayes's Theorem when considering the effects of age on recidivism for a specific defendant. The first is $P$ or $Q$ in the "parent" population that covers the age interval ($A$) into which the defendant falls. The others are $T$ and $F$ for $C$. Efficiency of a test when a specific value of $C$ is selected to identify likely recidivists from age interval $A$ for which $P$ is already known may then be determined through the application of the following general formula:

(1) $E_{A\&C} = (P_A \times T_C) / ((P_A \times T_C) + (Q_A \times F_C))$.[3]

This application of Bayes's Theorem is simple to calculate. Bayes's Theorem is also recognized by statisticians and philosophers of science as "one of the most important developments in epistemology in the 20th century, and one of the most promising avenues for further progress … in the 21st" (Talbot, 2001, p. 1). Finally, the theorem holds great practical significance for making weight-of-evidence determinations in court (Dawid, 2002; Jefferys, 2003). In SVP cases, for example, it would help all parties keep sight of the fact that the odds an offender with a high actuarial score will recidivate are not the same as the odds that a recidivist will have a high score ($E \neq T$). This, in turn, would prompt recognition of the possibility, graphically depicted in Figure 1, that an offender with a high score is not always destined to be a recidivist.

In spite of the clarifying power and very important implications of Bayes's Theorem, evaluators often do not discuss their Bayesian level of certainty when they are examined in court. They are also not often asked about this topic when they are cross-examined. Perhaps this avoidance stems from the fact that many expressions of the theorem include probability terms that are either discouragingly complex (Dawid, 2002; Swets et al., 2000; Fergusson et al., 1977) or difficult to remember because of weak mnemonic connections with some of the variables of concern (Meehl & Rosen, 1955).[4] It may also be due to a lack of knowledge about Bayes's Theorem or the steps involved in its calculation. Finally, it may be tedious and time-consuming to estimate P, T , or F, from the available information.

The main obstacle to conducting a Bayesian analysis of recidivism data for different age groups is obtaining access to adequate age-wise recidivism estimates. Although a data set addressing this problem has not been placed in the public domain, one may be extrapolated from research published by Hanson (2002). This research was based on a very large pool of incarcerated sex offenders which included more Americans than any of the samples that were used to develop other actuarials. Since civil commitment laws are focused on incarcerated American sex offenders, this feature heightens the relevance of data extrapolated from Hanson's sample.

In the remainder of this paper, Hanson's data base and procedures that were used to derive an age-wise recidivism table for offenders in this pool are described. A non-algebraic worksheet for performing the operations necessary to calculate Bayes's formula is subsequently presented, and estimates of T and F are derived for the Cs of actuarial tests that have commonly been used to identify likely recidivists. The results of applying this worksheet to age-wise recidivism rates and estimates of T and F are then summarized. The final section summarizes the major findings and discusses their implications for the development of ATSRs, risk assessment practices, and policies that bear on civil commitment procedures.

## Method

### Subjects

Hanson's (2002) data base consisted of 10 follow-up studies that included 4,673 male offenders, all but 287 of whom were either incarcerated or hospitalized. Although most were Canadian or British citizens, 1,724 were from American jurisdictions. Age-wise recidivism rates were presented for 3,751 offenders classified as rapists, molesters, or incest offenders who fell in nine different age groups (18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-59, 60-69, 70 and over); 922 subjects, in other words, fell into an "unclassifiable" category. Age was measured at the time of institutional release for 4,086 subjects and at the time of sentencing for the remaining 587. On the average, offenders were 36 years old. In five studies, sexual recidivism was defined as being charged with a new offense, and in the other five as being convicted of a new offense. Since 820 of the 4,673 subjects recidivated, P for all 10 samples was 17.5%. The average follow-up period was eight years.

Regarding data patterns for classifiable offenders, Hanson (2002) reported that 45% of all rapists were younger than 30 and that the recidivism rates for molesters, rapists, and incest offenders were 19%, 17%, and 8%, respectively. Regarding data patterns for both classifiable and unclassifiable subjects, he indicated that 5 of 131 offenders older than 60 committed new sex offenses, yielding a P of 3.8% for this group.

## Estimation of Prior Probabilities

Age-wise recidivism rates (P) for Hanson's classifiable offenders were determined by extrapolations from Figures 1 and 2 of his 2002 article. These results were then adjusted for the discrepancy in recidivism rates between classifiable and unclassifiable offenders (.152 versus .271).

A colleague of the authors who was unaware of the initial results also estimated age-wise recidivism rates using the author's procedures. Rater differences in estimates for specific age groups ranged from .001 to .007, with the average difference being .003. The correlation between ratings was .999 ($p$<.001), indicating that the results of the estimation procedure were highly reliable.[5]

Calculation of Bayes's Formula

The calculation of Bayes's Formula involves the completion of three steps. The first consists of compiling information required by the formula. The following operations achieve this step:

1. The test or procedure being evaluated is identified.
2. The "target population" to which the test or procedure is to be applied is specified.
3. The "critical test range" ($C$) used to identify likely recidivists is specified.
4. $P$ for the target population is recorded.
5. $P$ is subtracted from 1 to determine the non-recidivism rate ($Q$) for the target population.
6. The test's true positive fraction (the proportion of all recidivists it captures, referred to as $T$) for $C$ is recorded.
7. The test's false positive fraction (the proportion of non-recidivists it mistakenly flags as recidivists, referred to as $F$) for $C$ is recorded.

The second step is directed towards estimating the proportion of subjects in the target population that the test will flag as likely recidivists. This step is achieved by performing the following operations:

8. $P$ for the population is multiplied by $T$ to discover the proportion of the population the test will correctly identify as likely recidivists; as a result of this operation the area of a

rectangle is obtained, with P being the length of one side and T being the length of the other.

9. Q for the population is multiplied by F to discover the proportion of the population the test will incorrectly identify as likely recidivists; the area of a second rectangle is obtained through the application of this operation, with Q being the length of one side and F being the length of the other.
10. The results of the two foregoing operations are added together to discover the overall proportion of the population that, both correctly and incorrectly, will be identified by the test as likely recidivists; this sum equals the total area of the two rectangles referenced above, PT and QF.

The third step is directed towards estimating what proportion of offenders identified by the test as likely recidivists will actually recidivate. This is done by dividing the area of the rectangle calculated in the eighth step by the area of the two rectangles calculated in the tenth. The result (E) not only indicates the percentage of the time that experts who use C to identify recidivists will be right, but also the recidivism rate for subjects with test scores falling in C. If this result is subtracted from 1, the percentage of the time that experts will be wrong when they use C to identify recidivists will be obtained. This figure also represents the non-recidivism rate for subjects with test scores falling in C.[6]

Table 2 presents a worksheet for calculating Bayes's formula that users may find efficient and comprehensible, because it organizes calculations into an easy to follow flowchart that provides an explanation of the values resulting from each operation.[7]

Insert Table 2 about here

Estimation of the Data Probabilities

As the introduction and the foregoing section have indicated, estimates of a test's T (true positive fraction) and F (false positive fraction) for the C selected to identify likely recidivists are necessary to calculate Bayes's formula. Since much attention in SVP cases has been focused on test scores associated with subgroups whose recidivism rates exceed 50%, the Cs for this study included only those test intervals which identified risk groups with scores above 50%. The following subsections describe how T and F were estimated for each actuarial.

SORAG. Two experience tables for the SORAG have been published, one that predicts violent recidivism (VR) for 7 years, and another that predicts VR for 10 years (Quinsey et al., 1998). Equivalent tables on sexually violent recidivism (SVR) have not been disseminated by the test's developers based on the reasoning that "we don't want to encourage people to predict sexual recidivism specifically because we don't think it is sound policy" (e-mail from V. Quinsey to the author dated February 2, 2003), and that "practitioners' desire for the material … cannot be sufficient grounds to provide information that we believe will lead to confusion and misunderstanding" (e-mail from G. Harris to the author dated January 26, 2004). In spite of this decision, a table reporting the percent of subjects who fall in each of the 9 bins identified by the SORAG is available (Quinsey et al., 1998). Cross-validation studies indicating that the ratio of SVR to VR among sex offenders falls somewhere in the range of 54% (Harris et al., 2003) to 60% (Rice & Harris, 1997) have also been published.

This information was combined in several steps to derive estimates of T and F for predicting sexual recidivism with the SORAG over a 10-year period. First, the SVR rate for each bin was estimated by multiplying the 10-year VR recidivism rate for each bin (Quinsey et al., 1998, p. 244) by the most favorable SVR/VR ratio of .6 (Rice & Harris, 1997). This operation also indicated that the most appropriate C for using the SORAG in SVP cases includes scores (26-45) that fall in bins 8 and 9. Second, the SVR rate for each bin was multiplied by the percent of the total number of subjects in each (Quinsey et al., 1998, p. 245) to determine the percent of all subjects in each bin who were sexual recidivists. By summing the percentages for each bin, the sample-wise recidivism rate was obtained. Third, the sexual non-recidivism rate for each bin was multiplied by the percent of the total number of subjects in each to determine the percent of all subjects in each bin who were non-recidivists. By summing the percentages for each bin, the sample-wise non-recidivism rate was obtained. To estimate T, the percent of all subjects in bin 8 who were recidivists was added to the equivalent percent for bin 9, and this sum was divided by the sample-wise recidivism rate. To estimate F, the percent of all subjects in bin 8 who were non-recidivists was added to the equivalent percent for bin 9, and this sum was divided by the sample-wise non-recidivism rate. For the data reported by the developers of the SORAG, these operations indicated that $T \approx .127$ and $F \approx .055$ for 10-year predictions. For 7-year predictions, $T \approx .152$ and $F \approx .055$.

These results were consistent with estimates derived from cross-validational data that Barbaree et al. (2001) and Langton (2003) collected on the SORAG for a 4.5 year period. Like the Quinsey group, they reported the percentage of all subjects in each SORAG bin (Barbaree et al., 2001, p. 499) and the VR rates for each (Langton, 2003, p. 112). When the procedures described in the above paragraph were applied to their figures, it was found that $T \approx .136$ and $F \approx .042$. Averaging the 10-year estimates with these figures produced a final T estimate of .132 and a final F estimate of .049.

VRAG. On the basis of research published by the VRAG's developers (Harris et al., 2003; Quinsey, 1998), the most appropriate C for using the VRAG in SVP cases was determined to include scores (21-36) that fall in bins 8 and 9 of the test. Applying the same methods that were used to estimate T and F for 10-year SORAG predictions, it was found that T ≈ .145 and F ≈ .045. Although these accuracy measures were selected for analysis because of their long-term relevance, the 7-year predictions for the SORAG were more accurate in that T ≈ .184 and F ≈ .043.

MnSOST-R. When accuracy indicia were calculated from developmental data for the MnSOST-R (Epperson et al., 1999) over a 6 year risk period, it was found that T ≈ .44 and F ≈ .10 for a C of 8 and above. The MnSOST-R has been criticized by Wollert (2002; 2003), however, because the original sample was small (N=256) and non-representative in that about 18% of the 90 recidivists included came from outside sources, while 30% of all subjects were excluded because they were familial offenders. Wollert (2002) also reported cross-validational data, for a cohort of 95 subjects that did not include an excessive number of recidivists, which suggested that T ≈ .64 and F ≈ .21. In a second cohort with 125 subjects, T ≈.36 and F ≈.05 (Doren, 2002). In a third, with 150 subjects, a T of .39 and a F of .20 were obtained when a) the percentage of subjects in each MnSOST-R risk group (Barbaree et al., 2001) was multiplied by the corresponding group-wise recidivism rate (Langton, 2003); b) the percentage of subjects in each risk group was multiplied by the corresponding non-recidivism rate; c) the percentage of all subjects who were recidivists with scores falling in C was divided by the percentage of subjects who were recidivists, regardless of test scores; and d) the percentage of subjects who were non-recidivists with scores falling in C were divided by the percentage of subjects who were non-recidivists, regardless of scores. Averaging the results of these cross-validation studies on samples that did not include an excessive number of recidivists, the estimates of $T$ and $F$ that were adopted for the MnSOST-R in the present study were .46 and .15, respectively.

*Static-99*. When accuracy indicia are calculated from developmental data for Static-99 (Hanson & Thornton, 2001) over a 5-year risk period, $T \approx .27$ and $F \approx .09$ for a $C$ of 6 and above. The equivalent estimates for both 10- and 15-year follow-up periods are .25 and .08. Data from one 5-year cross-validational study indicated, however, that $T \approx .16$ and $F \approx .10$ (Harris et al., 2003), while data from another suggested that $T \approx .25$ and $F \approx .14$ (Barbaree et al., 2001; Langton, 2003). The original sample did, however, include a cross-validation sample and other confirmatory studies of Static-99 have been reported (Harris et al., 2003). In light of these advantages, it seemed reasonable to adopt the 15-year $T$ and $F$ estimates from the original report, but to add a cautionary note that they probably represent the most favorable set of assumptions as far as the performance of Static-99 is concerned.

*RRASOR*. When accuracy indicia were calculated from developmental data for the RRASOR (Hanson, 1997) over a 5-year risk period, $T \approx .08$ and $F \approx .01$ for a $C$ of 5 and above. Over a 10-year risk period for this $C$, $T \approx .07$ and $F \approx .007$. Estimates of $T$ and $F$ for the 10-year risk period are of doubtful accuracy, however, because regression to the mean effects were overlooked when 10-year recidivism estimates for each score group on this test were generated by multiplying 5-year estimates by a constant. Furthermore, data from one 5-year cross-validational study suggested that $T \approx .038$ and $F \approx .014$ (Harris et al., 2003), while another settled on $T \approx .08$ and $F \approx .03$ (Barbaree et al., 2001). Considering these cross-validation results in light of the facts that all RRASOR items are on Static-99 and the two tests are highly correlated (Roberts, Doren, & Thornton, 2002), it seemed reasonable to assume that the $F$ to $T$ ratio for RRASOR is about the same as that for Static-99, which is about .33. Setting long-term

T for RRASOR at .075 on the basis of the convergent estimates of Hanson (1997) and Barbaree and his colleagues (2001), the corresponding F was estimated to be about .025.

## Results

### Age-wise Recidivism Rates

Table 3 presents the recidivism rate for each age category based on the extrapolation procedures listed above. The standard error of the proportion for each rate (Gerstman, 2003), which suggests that the "true rate" is very close to the corresponding "observed rate," is also included. The data patterns in Table 3 for classifiable subjects are comparable to those reported by Hanson and summarized in the Subjects section: 46% of all rapists were younger than 30 years old and the recidivism rates for molesters, rapists, and incest offenders were 20%, 17%, and 8%, respectively. Comparability is also apparent in the data patterns for both classifiable and unclassifiable subjects: 4 of 116 offenders older than 60 were estimated to have committed new sex offenses, a recidivism rate of 3.4%. Furthermore, Hanson has confirmed that the 8.7% recidivism rate for those in the "50-59 year-old age category is a reasonable reading" (letter to S. Sappington from K. Hanson dated 10.17.04). Therefore, although the extrapolation at hand rests on a number of assumptions, it provides a good approximation of the data analyzed by Hanson (2002).

The entries for the pooled data in Table 3 show a consistent decline in recidivism as a function of age across all age categories, one that is even more orderly than the results broken-down by Hanson (2002) for sub-types of offenders. No doubt exists as to the presence of a very strong effect of age on sexual recidivism as the correlation between these variables is -.99 (df = 7; $p < .001$). A comparison of linear and logistic regression analyses also indicated that the linear model estimated the obtained data more accurately ($\underline{R}^2 = .98$; $\underline{F}$ (1,7) = 329, $\underline{p} < .001$) than any of the other models.

Insert Table 3 about here

Test Efficiency and Recidivism Rates Across Age Groups for Offenders Classified as Likely Recidivists

Recidivism rates for each age group with high scores for each test were calculated according to formula (1) and the calculation steps presented in Table 2. Since the study included 8 age groups and five tests, forty calculations were performed.

As an example of one such calculation, the following version of formula (1) was applied to determine the recidivism risk of offenders in the 18 to 24 year-old group when the accuracy indicia of Static-99 were considered:

(2) $\underline{E}_{\underline{A}\text{: 18-24 \& }\underline{C}\text{: 6+}} = (\underline{P}_{\underline{A}\text{: 18-24}} \text{ x } \underline{T}_{\underline{C}\text{: 6+}}) / ((\underline{P}_{\underline{A}\text{: 18-24}} \text{ x } \underline{T}_{\underline{C}\text{: 6+}}) + (\underline{Q}_{\underline{A}\text{: 18-24}} \text{ x } \underline{F}_{\underline{C}\text{: 6+}}))$.

The specific values required for solving this formula are found in the second sentence of the subsection on Static-99 in the Method section, where it is indicated that $\underline{T} \approx .25$ and $\underline{F} \approx .08$ for a $\underline{C}$ of 6+, and in the second to last column of the first row of Table 3, where it is indicated that $\underline{P}_{\underline{A}\text{: 18-24}} \approx .271$. Since $\underline{P}_{\underline{A}\text{: 18-24}} \approx .271$, $Q_{\underline{A}\text{: 18-24}} \approx .729$ ($1 - .271 \approx .729$).

Inserting these values into formula (2), the following solution is obtained:

(2a) $\underline{E}_{\underline{A}\text{: 18-24 \& }\underline{C}\text{: 6+}} = (.271 \text{ x } .25) / ((.271 \text{ x } .25) + (.729 \text{ x } .08))$,

$= (.068) / ((.068) + (.058))$,

$= .068 / .126$,

$= .54$.

Figure 2 is a plot of the results of applying the worksheet presented as Table 2 to age-wise recidivism data and accuracy indicia for the actuarial tests set forth in the "Data Probabilities" subsection of the Method section. Remarkably similar efficiency levels were obtained for all of the tests evaluated in this study. Setting the commitment standard at 50%, however, none of the tests were efficient for subjects over 24 years old. These results also indicate that experts who rely on actuarial tests for predicting likely recidivists for all but the youngest age group, will be wrong most of the time. For a population similar to Hanson's (2002) sample, this error rate will vary from about 52% for offenders in the 25-29 age range, to almost 90% for those in the 60-69 range.

Insert Figure 2 about here

## Discussion

### Major Findings

This paper has reported the results of applying Bayes's Theorem to a) age-wise sexual recidivism rates and b) accuracy indicia (T and F) for ATSR scores that are often used to identify civil commitment candidates as sexually violent predators. Five major findings stand out. First, a great deal of variation exists in the recidivism rates for sex offenders from different age groups, ranging from .27 for those who are youngest to .03 for those who are over 60. Second, recidivism rates consistently decline with advancing age. Third, the pattern of the decline in sexual recidivism with age parallels the pattern reported for more diverse offender samples, indicating that the age invariance theory (Hirschi & Gottfredson, 1983; Sampson & Laub, 2004) applies to sex offenders. Fourth, the ratio of F to T, which is critical for determining which actuarials are most efficient for making positive identifications, regardless of the condition being identified (Biggerstaff, 2000), is about the same for all tests. Because of this, they attain similar levels of efficiency. Fifth, all tests appear to be somewhat efficient when applied to the youngest group, which was characterized by a relatively high recidivism rate, but lose this efficiency when they are applied to older groups with lower recidivism rates.

## Actuarial Limitations and Implications for the Future Development of ATSRs

By indicating an error rate in excess of 50% for those older than the 18-24 group, the results of the study at hand raise an important practice question – does test efficiency for current actuarials deteriorate so rapidly with age that they are useful only for the very youngest group of adult offenders? There are two sides to consider regarding this issue. On the "con" side, the average period of risk for Hanson's (2002) samples covered 8 years. Deriving a 15-year estimate from the 8-year rate on the basis of long-term recidivism curves (Harris & Hanson, 2004), it might be argued that the long-term rate for sex offenders could be as much as 33% higher than what was estimated in the present study, or about 23%. This, in turn, would mean that efficiency rates for the 18-24, 25-29, and 30-34 year-old groups would increase to 64%, 58%, and 52%, respectively.

On the "pro" side, a recent U.S. Justice Department study of a comprehensive sample of 10,000 sex offenders released from prisons in 15 states found that only 5.3% of the releasees sexually recidivated in 3 years (Langan, Schmitt, & Durose, 2003). Other studies of sex offenders in Iowa (Adkins, Huff, & Stageberg, December, 2000), Washington (Barnowski, July, 2004), and Arizona (Bartosh, Garby, Lewis, & Gray, 2000) have obtained similar findings. If rates are this low in the United States, it is almost certain that the 15-year recidivism rate will not exceed the overall rate of 25% reported by Hanson and Thornton (2000), and that the levels of E reported in Figure 2 are overestimated.

Regardless of which of these scenarios is eventually shown to be most accurate, the results of the present study suggest that current actuarials are of limited value, at best, for SVP determinations. They may not even be useful at all. These possibilities underscore the importance of searching for ways to improve the performance of ATSRs.

In the short run, three steps would seem to be of potential value for addressing this issue. By applying the Bayesian methods described above to both low and high test scores, it would be possible to estimate the sexual recidivism rate for each possible age and test score combination.[8] These combinations could then be sorted into new risk groups that might, compared to their current counterparts, offer more in the way of efficiency, internal consistency, and coverage of the risk continuum. If these procedures were applied to Static-99, for example, the highest risk group would include offenders who were 18-24 years old with scores of 6. The next highest group would include offenders who were 25 to 29 years old with scores of 6, and offenders who were 18-24 with scores of 5. In contrast, offenders who were 60-69, but had scores of 6 would fall in a very low risk group. As the last step, the performance of tests reformulated along these lines could be compared against their original counterparts by analyzing data sets that have been compiled for the purpose of cross-validation. Hopefully, the new tests would achieve better results while adding to what is known about the effects of age on recidivism.

In the longer run, actuarials will need to be developed that are more effective than those that are now in use. In the interest of enhancing generalizability, a database should be compiled for a large, contemporaneous, and representative sample of U.S. offenders released from prisons in many different states, such as the one analyzed by Langan and his colleagues (2003) of the Justice Department. Drawing on what has been learned from the MacArthur Study of Mental Disorder and Violence (Monahan, Steadman, Silver, Appelbaum, Robbins, Mulvey, Roth, Grisso, and Banks, 2001), a range of outcomes (from convictions to self-reported and suspected sexual misconduct) and the potential value of compiling multiple actuarials (because different clusters of risk factors may be relevant for different age groups) should be studied. Finally, to prevent estimates from becoming outdated, information from a new cohort of offenders should be added every couple of years, accompanied by the elimination of data for the oldest cohort, so that actuarials may be re-normed.

These long-run recommendations would be costly to implement. However, the expenses they entail would surely be justified in light of the limitations of current actuarials, the huge amount of money now being spent on civil commitment centers (LaFond, 2003; Washington State Institute for Public Policy, March, 2005), and the importance of minimizing the number of unjust detentions that occur in conjunction with SVP commitment proceedings.

Accuracy Indicia for Actuarials and Their Implications for Risk Assessment In SVP Cases

In addition to considering the effects of recidivism and accuracy on test efficiency, the research at hand pointed to three important conclusions pertaining to measures of accuracy. These were as follows:

A. The highest T and F levels for any of the ATSRs were .46 and .15, respectively.

B. The average ratio of F to T was about .33.

C. The ratio of F to T for long-term predictions did not equal or exceed the ratio for predictions that covered a shorter risk period.

These findings hold at least four noteworthy implications. First, by documenting the restricted range of T and the considerable variability of F over high test scores, finding “A” suggests that experts, attorneys, and journal editors alike should avoid making optimistic assumptions about T or F, and that opinions advanced on such assumptions should be carefully scrutinized. Doren (February/March, 2002), for example, speculated that the 7-year recidivism rate for older offenders with high actuarial scores might exceed 38% on the grounds that a) 13 of the 131 offenders over 59 in Hanson's (2002) sample probably had high scores of 4 or 5 on the RRASOR; b) 5 offenders in the over-59 group recidivated; and c) 100% of these 5 re-offenders *might conceivably* have had high scores on the RRASOR. The last scenario could have occurred under only one condition - with a T of 100% for a C of 4-5 on the RRASOR. Had existing developmental data (Hanson, 1997) been analyzed, however, it would have been found that the best estimate of T for a C of 4-5 on the RRASOR was .20 and that there was therefore only 1 chance in 3,000 that all five re-offenders could have had high RRASOR scores ($.20^5 = .0003$). In this case, an analysis of T estimates might have averted the spread of misinformation occasioned by the publication of Doren's highly improbable speculations.

Overly optimistic assumptions about T also underpin the view, sometimes voiced in commitment cases, that the violent recidivism rates from the SORAG should be regarded as the best estimates of sexual recidivism for those who are seen as "specializing" in sexual misconduct. The first assumption on which this assertion rests is that, since the SORAG's C for the identification of likely violent recidivists includes bins 4 through 9, these bins also make up the SORAG's C for likely sexual recidivists. The second is that the T level for identifying likely sexual recidivists, when the SORAG is applied to sexual recidivists, is the same as that for identifying likely violent recidivists, which is .88. Going back to finding "A," however, a T of this magnitude is almost twice as large as the T level for any instrument yet developed for the prediction of sexual recidivism. Therefore, until an efficient experience table is compiled that focuses specifically on sexual recidivism, equating sexually violent recidivism with violent recidivism is simply untenable. This, in turn, indicates that the SORAG should only be used to predict violent recidivism rather than sexual recidivism in SVP evaluations, a view consistent with statements the test's developers have expressed against using "the VRAG or SORAG to make a numerical estimate of the lifetime likelihood of a person being arrested for a new sex offense" (e-mail to the author from V. Quinsey dated February 7, 2003).

The second implication of the accuracy analysis bears on the limits placed on actuarial efficiency by $T$, $F$, and $P$. As Meehl and Rosen (1955) have indicated, the ratio of $P / Q$ must be greater than the ratio of $F / T$ "in order for a positive diagnostic assertion" - in this case, sexual recidivism - "to be 'more likely true than false'" (p. 200). Finding "B," taken together with this inequality, indicates that ATSR's will be useless for predicting likely sexual recidivists unless those who are evaluated are drawn from populations with recidivism rates greater than .25 (the following steps solve this inequality: $P / Q > F / T$; $P / (1 - P) > F / T$; $P / (1 - P) > .33$ because finding B indicated that the average $F / T$ ratio for the actuarials in this study was .33; $P > .33\ (1 - P)$; $P > .33 - .33P$; $P + .33P > .33$; $1.33\ P > .33$; $P > .33 / 1.33$; $P > .25$).

Finding "B" also holds implications for evaluating the credibility of other risk assessment approaches in that it provides a quantified standard of accuracy. Non-actuarial approaches - in particular, clinical judgment and adjusting actuarial estimates on the basis of extra-test "risk factors" (Campbell, 2004; Hanson, 1998) - have been unable to match this standard. Where a reasonable body of evidence is available, as in the case of clinical judgment, the average $F / T$ fraction ($.36 / .42 \approx .86$) has been so large for the prediction of sexual recidivism among adults (Dix, 1976; Hall, 1988; Sturgeon & Taylor, 1980), that this approach will be useless for identifying recidivists unless they are drawn from populations with an unrealistically high $P$ of .46. Therefore, in the absence of new comparative evidence, claims that actuarial methods are matched or outperformed by other risk assessment methods are unjustified. Furthermore, when experts wish to advance such claims as part of their testimony, they should be prepared to disclose the decision rules they selected to identify recidivists (the equivalent of C) and to provide the court with evidence as to the T, F, P and E levels associated with their methods, as this will quantify both their level of uncertainty and how often they expect their predictions will be correct. Needless to say, the same expectations apply to those who use ATSRs. Finally, when information about decision rules, test accuracy, and efficiency is not elicited during direct examination, it would be helpful to the fact-finder if it were elicited on cross. It is hoped that the definitions and examples that have been presented in this paper will aid attorneys and experts addressing these issues.

The last implication of the accuracy analysis speaks to the efficiency of actuarials over time. Since experts in SVP cases are often charged with predicting whether respondents will recidivate over a long period, it may be tempting to assume that long-term estimates will necessarily be more accurate than those for intermediate periods. Finding "C", taken together with evidence that ATSRs are most accurate for very short follow-up periods (Sjostedt & Grann, 2002), suggests that this assumption is probably wrong. On the contrary, it is likely that ATSRs become *less* accurate over a long period as offenders with low test scores recidivate, consequently driving down T levels. Taken together, these considerations suggest that researchers should track the performance of ATSR's across time in order to determine the length of the risk periods over which they will be most accurate for making predictions.

## Should Actuarials Be Revised On the Assumption They Underestimate Sexual Recidivism?

Having been involved in many SVP cases, the author anticipates that a number of attorneys, researchers, and experts will respond to the results in Figure 2 by asserting that the practice of classifying older offenders as likely recidivists is justified in light of several arguments indicating that ATSRs underestimate sexual recidivism. These arguments, and sources that are sometimes cited in support of them, are listed below:

- The "prevalence discrepancy" argument asserts that the "real" recidivism rate for sex offenders exceeds the officially-recorded rate that is factored into actuarial experience tables because more sex offenses are committed in our society than are reported (Koss, Gidycz, & Winiewski, 1987; Lisak & Miller, 2002).
- The "systemic impact" argument asserts that the recidivism rate for sex offenders exceeds the officially-recorded rate that is factored into actuarial experience tables because not all who are suspected of committing sex crimes are convicted of them due to acquittals, charging decisions, and plea bargains (Harris et al., 2003).
- The "inadequate timeframe" argument asserts that ATSRs underestimate lifetime recidivism risk, which should be of primary concern in SVP proceedings, because the longest risk period they table spans only 15 years, whereas high rates of sexual recidivism have been reported for samples followed for 20 or more years (Doren, 1998; Hanson, Scott, & Steffy, 1995; Langevin, Curnoe, Fedoroff, Bennett, Langevin, Peever, Pettica, & Sandhu, 2004; Prentky, Lee, Knight, & Cerce, 1997).
- The "self-admission" argument asserts that ATSRs underestimate recidivism risk because sex offenders recidivate by committing undetected crimes, the occurrence of which is reflected in the fact that they consistently report engaging in more instances of sexual misconduct than the crimes listed on their records (Abel, Becker, Mittelman, Cunningham-Rathner, Rouleau, & Murphy, 1987; Baker, Tabacoff, Tornusciolo, & Eisenstadt, 2001; Groth, Longo, & McFadin, 1982; Weinrott & Saylor, 1991; Zolondek, Abel, Northey, & Jordan, 2001).

- The “undetected recidivism” argument asserts that ATSRs underestimate recidivism risk because follow-up studies of sex offenders based on sources of information other than self-report or official records have indicated that released sex offenders commit sex crimes that are undetected in the sense of never having been adjudicated (Falshaw, Bates, Patel, Corbett, & Friendship, 2003; Langevin et al., 2004; Marshall & Barbaree, 1988).

If these arguments are correct, it would be reasonable to consider whether existing actuarial tables might be adjusted for underestimation effects and, if this were possible, to determine the range in test performance that would be expected under different assumptions about the magnitude of these effects. The feasibility of pursuing these options depends, however, on the validity of the underestimation hypothesis and the extent to which it is capable of unbiased quantification.

If one analyzes each of the above arguments thoroughly, reviews the articles offered as evidence in their support, and considers other relevant documents, it is clear that the underestimation hypothesis does not satisfy either of the foregoing conditions. Regarding the “prevalence discrepancy” argument, for example, it is undeniable that more sex offenses are committed than reported. No evidence exists, however, that this discrepancy is attributable primarily to sex offenders. On the contrary, the great majority of unreported sex crimes are probably committed by men who have never been convicted of a sex offense (Johnson, 1980; Koss et al., 1987; Lisak & Miller, 2002). Furthermore, the effect of this discrepancy on the accuracy of ATSRs has never been estimated in a peer-reviewed publication, so it is unclear as to how it might be used to adjust actuarial tables.

It is also undeniable that not all of those who are suspected of committing a sex crime are convicted of doing so. The impact of “systemic” factors associated with underestimation would seem to be countered to some extent, however, by the impact of other systemic factors that inflate recidivism rates. In particular, it has been shown that a large number of false allegations of sexual misconduct are made under some conditions (Kanin, 1994) and that a substantial number of defendants charged with sex offenses were falsely convicted prior to the advent of DNA identification testing (Gross, Jacoby, Matheson, Montgomery, & Patil, 2005). In addition, it is the case that a) the number of reported molestation cases and forcible rapes have dropped substantially (Koch, 2005, August 24; Washington Sentencing Guidelines Commission, 2004); b) recidivism rates for rapists have dropped (Beck & Shipley, 1989; Langan et al., 2003); and c) the percentage of recidivists a test is capable of identifying (E) decreases when the rate of sexual recidivism decreases (Janus & Meehl, 1997; Saari & Saari, 2002). Taken together, these considerations raise the possibility that decreases in the reported number of victimizations and in the sexual recidivism rate may be so large that new actuarials that included both "convictions" and "suspicions" of sexual misconduct as outcome measures would be irrelevant to SVP cases, because the recidivism rate for even those with high test scores might not approach the commitment standard. A final reservation is that a formula for estimating the magnitude of systemic impacts has never been published in a peer-reviewed source.

With respect to evaluating the inadequate timeframe argument, it is helpful to keep in mind that the studies cited in its support have some serious limitations. Two of them (Langevin et al., 2004; Prentky et al., 1997) are of limited relevance for SVP hearings, for example, because they monitored subjects with 3 to 4 times as many convictions as the "prison-release" population to which SVP laws are applied (Janus & Meehl, 1997), 90% of whom have been convicted of only one sex crime (Song & Lieb, 1995). The third (Hanson et al., 1995) reports an inflated recidivism rate, because data for elderly non-recidivists were destroyed (Wollert, 2001).

Regarding the issue of inflation, it is probably the case that the cited studies, and current actuarials as well, are inflated to an unknown extent because they are based on old data that do not take into account such factors as a) the increase in exonerations attributable to improved methods of investigation (Gross et al., 2005; Kanin, 1994); b) apparent decreases in the sexual recidivism rate that are reported in recent documents (Adkins et al., 2000; Barnowski, July, 2004; Bartosh et al., 2003; Beck & Shipley, 1989; Langan et al., 2003); and c) the discrepancy between the rate of sexual recidivism in general, which is typically studied, versus the rate of *predatory* sexual recidivism, which has not been studied but is the predicted outcome with which most SVP laws are concerned (Janus & Prentky, 2003; Wollert, 2001). Finally, the cited studies all used small samples, ranging from 247 to 361 offenders, and selected offenders from a single source. Lower long-term recidivism rates are reported for large samples that are drawn from many different sources. Harris and Hanson (2004), for example, performed a survival analysis on a pool of 4,724 offenders drawn from 10 different sources and reported a failure rate of 24% over a 20-year period. Even in this study, the actual recidivism rate could be as low as 16% because the failure rate for a sample may be as much as one and half times larger than the recidivism rate for the same sample (Prentky et al., 1997).

The self-admission argument derives much of its promise from the assumptions that a) it is possible to calculate the undetected sexual recidivism index (URI), which is the ratio of the number of detected and undetected recidivists to the number of detected recidivists; b) that the URI is large; and c) that it is possible to adjust actuarials on the basis of the URI once it has been estimated. The promise of this argument remains undetermined, however, because all studies cited in its support asked offenders how many crimes they committed in their past. Not one, in other words, asked about the number of sex offenses that were committed *after each conviction*.

As a result, none of these studies reported a URI or data that could be used to calculate a meaningful URI.

Like the self-admission argument, the status of the undetected recidivism argument turns on the calculation of the URI. It also runs afoul of the same problems besetting the self-admission argument in that none of the cited research reported URIs. Furthermore, no evidence exists that incidents of undetected recidivism may be identified at acceptable levels of reliability. In contrast, a quantified analysis published in the journal at hand has argued that "the unofficially measured 're-offense' rate, may not be far off from the officially measured 'reconviction' rate" (Janus & Meehl, 1997, p. 52). Finally, about the only recent data that might be used for calculating a URI was found by the author in a study (Falshaw et al., 2003) in which 10 offenders were reportedly re-convicted of committing a new sex offense while 12 were classified as having been involved in "the perpetration of another illegal sexual act, whether caught or not" (p. 211). Taken together, these figures would suggest a URI of about 1.2 (i.e., (10 + 2/10) = 1.2). The stability and generalizability of this estimate is open to question, however, because it was derived from a very small number of British subjects, and the reliability with which incidents of undetected recidivism could be identified by the researchers was not determined.

Perhaps the underestimation hypothesis will eventually be confirmed. Presently, however, a very large number of considerations must be taken into account to insure estimation procedures that are not biased in favor of one side or the other. Further, the size of the effect of almost all of these variables has never been quantified. As a result it is virtually impossible to derive a defensible formula for adjusting actuarials for the effect of undetected recidivism or any other factor associated with the underestimation hypothesis. In the absence of such a formula, which is the cornerstone of the actuarial method (Dawes, Faust, & Meehl, 1989), the most accurate and unbiased approach for experts, attorneys, and fact-finders is to resist the temptation to speculate and to rely instead on actuarial formulas that are informed by solid empirical research. This would include the Bayesian formula that was used above to adjust actuarials for the clearly defined impact of age on recidivism and that, as a result, constitutes a meaningful addition to other scientific tools that inform the prediction of sexual recidivism.

## Implications for Commitment Issues and Policies

Since the effects of age on recidivism were apparently overlooked when many older sex offenders were committed, it would be in the interests of justice to seek new trials for these individuals to determine whether they actually qualify as SVPs. A corollary of this position is that end of sentence review committees that refer prisoners for commitment consideration could "do a more thorough job of screening potential SVP cases" (LaFond, 2003) by focusing their attention primarily on young adults who were fully competent at the time they offended and reducing the number of older offenders that are identified as probable SVPs. .

The terms of commitment for members of this younger group of offenders should not be regarded as indefinite, however. The reason for this is that the best available risk assessment method (i.e., actuarial testing) will eventually point to the conclusion that the recidivism rate for each detainee - given the P for his age group, his test score, and the effect of measurement error (Anastasi, 1988; Gulliksen, 1950) [9] - does not meet the commitment standard.

These policies, if adopted, might free up resources that could be allocated to other interventions to combat sexual recidivism, such as outpatient sex offender treatment and improved sex offender supervision and education on sex offending issues for all offenders released from prison. The importance of a much broader allocation of societal resources, which has been recognized by other researchers (LaFond, 2003; Janus & Prentky, 2003; Janus, 2005), is underscored by a recent Justice Department report indicating that 517 of 9,691 sex offenders released from prison in 1994 committed new sex offenses in a three-year period, compared to 3,328 sex offenses that were committed by 269,174 other released offenders (Langan et al., 2003). If all sex offenders in this cohort had been screened as possible civil commitment cases using Static-99, a relatively small number of sex crimes would have been averted due to the detention of 129 likely recidivists (Static-99 T of .25 x 517 sexual recidivists = 129) while 734 offenders with high test scores would have been unjustly detained (Static-99 F of .08 x 9,174 non-recidivists = 734). The Static-99 screen would miss 383 (517 – 129 = 383) recidivists, however. All of the sex crimes committed by the other released offenders would also be missed, setting the stage for the commission of 3,711 (383 + 3,328 = 3,711) sex crimes within a relatively brief span of time. Overall, only 3% (129 / (3,711 + 129) = 3%) of those who committed new sex crimes would be incapacitated under these conditions.

From the author's perspective, a risk management scheme that identifies 3% of all sexual recidivists is not cost-effective. Furthermore, in the above scenario, 383 recidivists would be mistakenly released while 734 non-recidivists would be unjustly detained. This means that the result of dividing the first quantity by the latter, also known as "R" (Lloyd & Grove, 2001), approximates .5. In other words, only one dangerous respondent would be mistakenly released for every two non-dangerous respondents who were unjustly detained.

R is a useful measure for understanding trends in criminal justice policies because it reflects the restraint a society is willing to place on punishment, through its jurisprudence system, in the interest of protecting individual liberty. A large R, for example, indicates that the release of many potentially dangerous respondents is tolerated so that the limits of fairness are not breeched by the unjust commitment of large numbers of non-dangerous respondents. A small R, in contrast, is indicative of an emphasis on incarcerating as many potentially dangerous respondents as possible at the expense of incarcerating a large number of non-dangerous respondents as well. Overall, it seems reasonable to assume that the value of R decreases as communities throughout a society become more punitive because of the spread of fear and frustration.

Volokh (1997) summarized and analyzed R values that have been espoused by jurists, legal theorists, biblical figures, teachers, American patriots, Mafiosi, talk show hosts, politicians, Roman emperors, English kings, police commissioners, novelists, religious leaders, philosophers, and military commanders from different countries and different eras. Although wide variations were evident, almost all Rs (except those attributed to Bismarck and Stalin) were greater than 1. Furthermore, the most widely-endorsed value of R was equal to 10, a figure cited by the British jurist William Blackstone (1767/1979) that has come to be known as the "Blackstone Ratio".

Against this historical backdrop, the .5 R value associated with the use of actuarials in SVP cases flies in the face of both Anglo-American and international legal traditions. In some SVP states, the true value of R may not be quite this low due to the additional requirement of proving the existence of a mental abnormality. In others, R would probably be *even lower* because of decisions that allow "likely recidivism" to be defined as less than a 50% level of risk (Massachusetts v. Boucher, 2002). Regardless of which estimates are most accurate, however, Rs in the 0 to 1 range raise troubling questions about the implications of SVP commitment procedures for the status of individual liberty that citizens from all walks of our society would do well to consider.

In light of these questions, legislators, policy makers, and opinion-leaders are encouraged to study Bayes's theorem for three reasons. The first is that this would enhance their understanding of why it is very difficult to predict alarming but infrequent sex crimes with any reasonable degree of certainty, no matter how much money is spent on doing so. The second is that it would help them discharge their leadership duties by explaining this unpleasant reality to anxious constituents. The third is that it would emphasize the importance of evaluating every piece of proposed legislation directed at the goal of averting shocking but rarely predictable crimes in terms of the magnitude of the financial and liberty costs entailed by each option. Such advances might, in turn, increase their motivation to develop practical and non-draconian options for containing sexual violence while conceptualizing programs that impact many potential sex offenders rather than just a few.

## References

Abel, G., Mittelman, M., Becker, J., Cunningham-Rathner, J, Rouleau, J., & Murphy, W. (1987). Self-reported sex crimes of non-incarcerated paraphiliacs. Journal of Interpersonal Violence, 2, 3-25.

Adkins, G., Huff, D., & Stageberg, P. (December, 2000). The Iowa Sex Offender Registry and recidivism. Iowa City, IO: Iowa Department of Human Rights.

Anastasi, A. (1988). Psychological testing. New York: Macmillan.

Baker, A. J., Tabacoff, R., Tornusciolo, G., & Eisenstadt, M. (2001). Calculating number of offenses and victims of juvenile sexual offending: The role of post-treatment disclosures. Sexual Abuse, 13, 79-90.

Barbaree, H. E., Blanchard, R., & Langton, C. (2003) The development of sexual aggression through the lifespan. In R. A. Prentky, E. Janus, & M. Seto (Eds.), Annals of the New York Academcy of Sciences, Vol. 989. Sexually coercive behavior (pp. 59-71). New York: New York Academy of Sciences.

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.

Baldessarini, R. J., Finklestein, S, & Arana, G. (1983). The predictive power of diagnostic tests and the effect of prevalence of illness. Archives of General Psychiatry, 40, 569-573.

Barnoski, R. (July, 2004). Sentences for adult felons in Washington: Options to address prison overcrowding. Olympia, WA: Washington State Institute for Public Policy.

Bartosh, D. L., Garby, T., Lewis, D., & Gray, S. (2003). Differences in the predictive validity of actuarial risk assessments in relation to sex offender type. International Journal of Offender Therapy and Comparative Criminology, 47(4), 422-438.

Bayes, T. (1764). An essay toward solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, 53, 370-418.

Beck, A. J., & Shipley, B. (1989). Recidivism of prisoners released in 1983. Washington, D.C.: U.S. Department of Justice.

Beech, A. R., Fisher, D., & Thornton, D. (2003). Risk assessment of sex offenders. Professional Psychology: Research and Practice, 34, 339-352.

Biggerstaff, B. J. (2000). Comparing diagnostic tests: A simple graphic using likelihood ratios. Statistics in Medicine, 19, 649-663.

Blackstone, W. (1767/1979). Commentaries on the laws of England, Vol. 4. Chicago: University of Chicago Press.

Campbell, T. W. (2004). Assessing sex offenders. Springfield, IL: Charles C Thomas.

Covington, J. R. (1997). Preventive detention for sex offenders. Illinois Bar Journal, 85, 493-498.

Dawes, R. M., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. Science, 243, 1668-1674.

Dawid, A. P. (2002). Bayes's Theorem and weighing evidence by juries. In R. Swinburne (Ed.), Proceedings of the British Academy: Vol. 113. Bayes's Theorem (pp. 71-90). London: Oxford University Press.

Dix, G. E. (1976). Differential processing of abnormal sex offenders: Utilization of California's Mentally Disordered Sex Offender Program. The Journal of Criminal Law & Criminology, 67, 233-243.

Doren, D. M. (2002). Evaluating sex offenders. Thousand Oaks, CA: Sage.

Doren, D. (February/March, 2002). To what extent does aging negate historically assessed recidivism risk? Sex Offender Law Report, 3(2), 17-27.

Doren, D. M. (1998). Recidivism base rates, predictions of sex offender recidivism, and the "sexual predator" commitment laws. Behavioral Sciences and the Law, 16, 97-114.

Doren, D. M. & Dow, E. (2003). What shrinkage of the MnSOST-R? A reply to Wollert (2002). Journal of Threat Assessment, 2, 49-63.

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, Minnesota Department of Corrections at St. Paul.

Falshaw, L., Bates, A., Patel, V., Corbett, C., & Friendship, C. (2003). Assessing reconviction, re-offending, and recidivism in a sample of UK sexual offenders. Legal and Criminological Psychology, 8, 207-215.

Fergusson, D. M., Fifield, J., & Slater, S. (1977). Signal detectability theory and the evaluation of prediction tables. Journal of Research in Crime and Delinquency, 14, 237-246.

Gerstman, B. B. (2003). Epidemiology kept simple. Hoboken, NJ: Wiley-Liss.

Gross, S. R., Jacoby, K., Matheson, D., Montgomery, N., & Patil, S. (2005). Exonerations in the United States 1989 through 2003. The Journal of Criminal Law & Criminology, 95, 523-561.

Groth, A. N., Longo, R., & McFadin, J. Undetected recidivism among rapists and child molesters. Crime and Delinquency, 23, 450-458.

Grove, W. M., & Meehl, P. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures. Psychology, Public Policy, and Law, 2(2), 293-323.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Hall, G.C.N. (1988). Criminal behavior as a function of clinical and actuarial variables in a sex offender population. Journal of Consulting and Clinical Psychology, 56, 773-775.

Hanley, J. A. & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143, 29-36.

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

Hanson, R.K. (1997). The development of a brief actuarial risk scale for sexual offense recidivism. Unpublished manuscript, Public Works and Govt. Services Canada at Ottawa.

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

Hanson, R. K. (2002). Recidivism and age. Journal of Interpersonal Violence, 17, 1046-1062.

Hanson, R. K. (May 3, 2004). The applicability of Static-99 to older offenders. Unpublished manuscript.

Hanson, R.K., Morton, K., & Harris, A. (2003). Sex offender recidivism risk. Annals of the New York Academy of Sciences, 989, 154-166.

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.

Hanson, R. K., Scott, H., & Steffy, R. (1995). A comparison of child molesters and nonsexual criminals. Journal of Research in Crime and Delinquency, 32, 325-337.

Harris, A. J. R. & Hanson, R. K. (2004). Sex offender recidivism: A simple question. Unpublished manuscript, Public Safety and Emergency Preparedness Canada at Ottawa.

Harris, A., Phenix, A., Hanson, R. K., & Thornton, D. (2003). STATIC-99 coding rules revised-2003. Unpublished manuscript, Solicitor General of Canada at Ottawa.

Harris, G.T., Rice, M., Quinsey, V., Lalumiere, M., Boer, D., & Lang, C. (2003). A multisite comparison of actuarial risk instruments for sex offenders. Psychological Assessment, 15(3), 413-425.

Harris, G. T., Rice, M., & Quinsey, V. (1993). Violent recidivism of mentally disordered offenders. Criminal Justice and Behavior, 20, 315-335.

Hirschi, T. & Gottfredson, M. (1983). Age and the explanation of crime. American Journal of Sociology, 89, 552-584.

In re Young, 120 Wn. App. 753, 761-762, 86P.3d 810 (2004).

Iversen, G. R. (1984). Bayesian statistical inference. Beverly Hills: Sage.

Janus, E. S. (2004). Closing Pandora's Box: Sexual predators and the politics of sexual violence. Seton Hall Law Review, 34, 1233-1253.

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

Janus, E. S. & Prentky, R. (2003). Forensic use of actuarial risk assessment with sex offenders: Accuracy, admissibility, and accountability. American Criminal Law Review, 40 (4), 1443-1499.

Jefferys, W. H. (2003). Review of Bayes's Theorem (Proceedings of the British Academy, Vol. 113). Journal of Scientific Exploration, 17 (3), 537-542.

Johnson, A. G. (1980). On the prevalence of rape in the United States. Signs: Journal of Women in Culture and Society, 6, 136-146.

Kahn, T. J. & Chambers, H. (1991). Assessing reoffense risk with juvenile sex offenders. Child Welfare, 70, 333-345.

Kanin, E. J. (1994). False rape allegations. Archives of Sexual Behavior, 23, 81-92.

Koch, W. (2005, August 24). Despite high-profile cases, sex offense crimes decline. USA Today. Retrieved August 25, 2005 from http://www.usatoday.com/news/nation/2005-08-24-sex-crimes-cover_x.htm.

Koss, M. P., Gidycz, C., & Winiewski, N. (1987). The scope of rape: Incidence and prevalence of sexual aggression and victimization in a national sample of higher education students. Journal of Consulting and Clinical Psychology, 55, 162-170.

LaFond, J. Q. (2003). The costs of enacting a sexual predator law and recommendations for keeping them from skyrocketing. In B. Winick and J. LaFond (Eds.), Protecting society from sexually dangerous offenders (pp. 283-299). Washington, D.C.: American Psychological Association.

Langan, P. A., Schmitt, E., & Durose, M. (2003). Recidivism of sex offenders released from prison in 1994. Washington, DC: U.S. Department of Justice.

Langevin, R., Curnoe, S., Fedoroff, Bennett, R., Langevin, M., Peever, C., Pettica, R., & Sandhu, S. (2004). Lifetime sex offender recidivism: A 25-year follow-up study. Canadian Journal of Criminology and Criminal Justice, 46, 531-552.

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

Lisak, D. & Miller, P. (2002). Repeat rape and multiple offending among undetected rapists. Violence and Victims, 17, 73-84.

Lloyd, M. D. & Grove, W. (2001). The uselessness of the Minnesota Sex Offender Screening Tool-Revised (MnSOST-R) in commitment decisions. Unpublished manuscript, University of Minnesota at Minneapolis.

Marshall, W. & Barbaree, H. (1988). The long-term evaluation of a behavioral treatment program for child molesters. Behavior Research and Therapy, 26, 499-511.

Massachusetts v. Boucher, 438 Mass. 274; 780 N.E.2d 47; 2002 Mass. LEXIS 876.

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.

Metz, C. E. (1978). Basic principles of ROC analysis. Seminars in Nuclear Medicine, 8, 283-298.

Miller, H. A., Amenta, A., & Conroy, M. (2005). Sexually violent predator evaluations: Empirical evidence, strategies for professionals, and research directions. Law and Human Behavior, 29, 29-54.

Milloy, C. (December, 2003). Six-year follow-up of released sex offenders recommended for commitment under Washington's Sexually Violent Predator Law, where no petition was filed. Unpublished manuscript, Washington State Institute for Public Policy at Olympia.

Monahan, J., Steadman, H., Silver, E., Appelbaum, P., Robbins, P., Mulvey, E., Roth, L., Grisso, T., & Banks, S. (2001). Rethinking risk assessment: The MacArthur Study of Mental Disorder and Violence. New York: Oxford University Press.

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

Mossman, D. (1994b). Further comments on portraying the accuracy of violence predictions. Law and Human Behavior, 18, 587-593.

Nicholaichuk, T. & Yates, P. (2002). Treatment efficacy – outcomes of the Fullwater Sex Offender Program. In B. Schwartz (Ed.), The sex offender: Current treatment modalities and system issues (pp. 25 – 38). Kingston, NJ: Civic Research Institute.

Prentky, R., Lee, A., Knight, R., & Cerce, D. (1997). Recidivism rates among child molesters and rapists: A methodological analysis. Law and Human Behavior, 21, 635-659.

Quinsey, V.L., Harris, G., Rice, M., & Cormier, C. (1998). Violent offenders. Washington, D.C.: American Psychological Association.

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

Rice, M. E. & Harris, G. (1997). Cross-validation and extension of the Violence RiskAppraisal Guide for child molesters and rapists. Law and Human Behavior, 21, 231-241.

Roberts, C. F., Doren, D. & Thornton, D. (2002). Dimensions associated with assessments of sex offender recidivism risk. Criminal Justice and Behavior, 29, 569-589.

Saari, R. J. & Saari, L. (August/September, 2002). Actuarial risk assessment with elderly sex offenders: Should it be abandoned? Sex Offender Law Report, 3(5), 68, 73-76.

Sampson, R. J. & Laub, J. (2003). Life-course desisters? Trajectories of crime among delinquent boys followed to age 70. Criminology, 41, 301-339.

Sjostedt, G. & Grann, M. (2002). Risk assessment: What is being predicted by actuarial prediction instruments? International Journal of Forensic Mental Health, 1, 179-182.

Song, L. & Lieb, R. (1995). Washington State sex offenders: Overview of recidivism studies. Unpublished manuscript, Washington State Institute for Public Policy at Olympia.

Smith, W. R. & Monastersky, C. (1986). Assessing juvenile sexual offenders' risk for reoffending. Criminal Justice and Behavior, 13, 115-140.

Sturgeon, V. H. & Taylor, J. (1980). Report of a five-year follow-up study of mentally disordered sex offenders released from Atascadero State Hospital in 1973. Criminal Justice Journal, 4(3), 31-63.

Swets, J. A., Dawes, R., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. Psychological Science in the Public Interest, 1, 1-26.

Talbot, W. (2001). Bayesian epistemology. Stanford encyclopedia of philosophy. Retrieved January 15, 2005, from http://plato.stanford.edu/archives/fall2001/entries/epistemology-bayesian/.

Thornton, D. & Doren, D. (2002). How much safer are older offenders? Annual Conference of the Association for the Treatment of Sex Abusers.

Volokh, A. (1997). n guilty men. University of Pennsylvania Law Review, 146, 173-216.

Washington State Institute for Public Policy (March, 2005). Involuntary commitment of sexually violent predators: Comparing State Laws. Retrieved June 15, 2005, from http://www.wsipp.wa.gov/.

Washington State Senate Committee on Human Services & Corrections (February 14,2005). Senate bill report on SB 5582. Retrieved February 25, 2005, from http://www.leg.wa.gov/wsladm/billinf1/dspBillSummary.cfm?billnumber=5582&year=2005.

Washington State Sentencing Guidelines Commission (2004). Sex offender sentencing. Online. Retrieved June 15, 2005, from http://www.sgc.wa.gov.

Weinrott, M. R. & Saylor, M. (1991). Self-report of crimes committed by sex offenders. Journal of Interpersonal Violence, 6, 286-300.

Wilkins, L. T. (1969). Evaluation of penal measures. New York: Random House.

Wollert, R. (2001). An analysis of the argument that clinicians under-predict sexual violence in civil commitment cases. Behavioral Sciences and the Law, 18, 171-184.

Wollert, R. (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.

Wollert, R.(2003). Additional flaws in the Minnesota Sex Offender Screening Tool-Revised: A response to Doren and Dow. Journal of Threat Assessment, 2 (4), 65-78.

Wollert, R. (April, 2005). An application of Bayes's Theorem to age-wise sexual recidivism rates. Paper presented at the meeting of the Western Psychological Association, Portland, OR.

Zolondek, S. C., Abel, G., Northey, W., & Jordan, A. (2001). The self-reported behaviors of juvenile sexual offenders. Journal of Interpersonal Violence, 16, 73-85.

Zweig, M. H. & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental tool in clinical medicine. Clinical Chemistry, 39, 561-577.

## Author's Note

Portions of this paper were presented at the 2005 meeting of the Western Psychological Association. The author is indebted to Jacqueline Waggoner, Kevin Krieg, Robert Halon, and Ted Donaldson for their comments on earlier versions of this manuscript. Requests for reprints of this article and correspondence should be sent to the author at P.O. Box 61849, Vancouver, WA 98666-1849. The author may also be contacted at 360-737-7712 or rwwollert@aol.com. The address of his website is richardwollert.com.

# Footnotes

[1] See Miller et al., 2005, for a summary of how the key elements of SVP laws are defined in each state.

[2] This is a critical assumption in that a test will not be generalizable from one population to another, and hence worthless, unless $\underline{T}$ and $\underline{F}$ are stable.

[3] In this application of Bayes's Theorem the value of $\underline{E}$ is numerically (but not conceptually) identical to what has been referred to elsewhere as "positive predictive power" (Quinsey et al., 1998). If expert confidence in predictions of non-recidivism were of primary concern, it would be appropriate to use another version of Bayes's Theorem (i.e., $\underline{E} = (\underline{Q} \text{ x } (1-\underline{F})) / (((\underline{Q} \text{ x } (1-\underline{F})) + (\underline{P} \text{ x } (1-\underline{T})))$ that yields a value identical to a measure called "negative predictive power." Per the null hypothesis, however, commitment candidates are regarded as non-predatory offenders until evidence proves otherwise. One implication of this principle is that an expert who hints, suggests, or argues that a respondent meets the recidivism prong of a commitment standard is essentially espousing the view that she is reasonably certain that the chance of the respondent's recidivating is greater than that for the "riskiest" group of non-predatory offenders. In doing so, she is concerned with the prediction of recidivism rather than non-recidivism. Experts are also unlikely to focus primarily on the prediction of non-recidivism because, analogous to the principle of innocent until proven guilty, it is unnecessary for a respondent to prove his recidivism risk is lower than the commitment standard in order to be released. Taken together, these considerations point to the conclusions that the appropriate focus of the analysis at hand falls on certainty about recidivism predictions and that the appropriate formula to use for this analysis is the one given in the text.

[4] The author has sought to address these particular issues by emphasizing arithmetic descriptions and examples of the variables that are fed into Bayes's Theorem (see paragraphs 8 through 11 in the introduction) and by adopting symbols in the formal statement of the theorem (see the fifth paragraph from the end of the introduction) that have some mnemonic connection with these variables ($E$ = predictive efficiency; $A$ = age interval; $C$ = critical test range; $L$ = left-over or alternate test range; $P$ = recidivism base rate; $Q$ = non-recidivism base rate; $T$ = true positive fraction; $F$ = false positive fraction).

[5] The author is indebted to Diane Lytton for undertaking this time-consuming task. The specific estimation steps she followed may be obtained from the author or downloaded from his website.

[6] Thinking of Bayes's theorem as the division of the area of one rectangle by the sum of the area of that rectangle plus another is useful in a number of respects. For one thing, it decreases the "intimidation factor" that dogs algebraic notation. For another, it makes it easier to remember the terms in the theorem. For still another, it enhances an intuitive understanding as to why, all other things being equal, $E$ decreases when $P$ decreases – that is, the area of the $PT$ rectangle becomes smaller in relation to the sum of the areas of the $PT$ and $QF$ rectangles.

[7] An Excel © program for calculating Bayes's Theorem, formatted like Table 2, may be obtained from the author or downloaded from his website. Operation of the program requires that only three values ($P$, $T$, $F$) be inserted in the first two rows.

[8] For those with high test scores, this elaboration would include the same prior probabilities and data probabilities used in the analysis at hand. Prior probabilities and data probabilities would need to be determined for those with lower scores, however, for each age and score group combination. In general, the prior probabilities for a group with a specified $P_A$ and

$C$ would be obtained by multiplying $P_A$ by the result of dividing the recidivism rate for offenders with scores of $C$ by the recidivism rate for all offenders. For example, the age-wise recidivism rate for those in their mid-fifties is .087, the long-term recidivism rate for those with Static-99 scores of 4 or less is .191 (Harris et al., 2003), and the long-term recidivism rate for all sex offenders in the Static-99 developmental sample is .25 (Harris et al., 2003). The prior probability for those in their mid-fifties with Static-99 scores of 4 is therefore about .066 (.087 x (.191/.25) = .066). Data probabilities for a specified $C$ would be obtained by calculating $T$ and $F$ after eliminating data for offenders with scores greater than $C$. To obtain $T$ and $F$ for a $C$ = 5 on Static-99, for example, only data for those with scores of 5 and less would be analyzed. To obtain $T$ and $F$ for $C$ = 4, only data for those with scores of 4 and less would be considered. An 8-year risk table, consisting of 56 cells (8 age groups by 7 score groups), would be obtained by applying Bayes's Theorem to the foregoing prior probabilities and data probabilities. A 15-year table could be derived by multiplying each cell by 1.33, which is the result of dividing the 15-year sexual recidivism rate for all offenders by the 8-year rate (see the first paragraph of this subsection for further discussion of this ratio).

[9] In order for an expert to be reasonably certain in rejecting the null hypothesis that the recidivism risk for a respondent is not the same as the risk for non-SVPs, she must be reasonably certain that the lowest plausible estimate of his risk level (see the fourth paragraph of the introduction) exceeds the non-SVP standard. The lowest plausible estimate for a respondent will always be less than his obtained test score, however. This difference, and the width of the corresponding confidence interval (or what might also be called the "region of doubt") is due to measurement error, which arises because experts sometimes disagree when they score the same group of subjects on the same test. If the measurement error for a test is small, the region of

possible detection error will be narrow and the lowest plausible estimate will not fall too far below the level suggested by the offender's obtained test score. If it is large, the region of possible error will be wide and the lowest plausible estimate will be substantially less than the level suggested by the offender's obtained test score. It is therefore important for experts to consider measurement error when deriving predictions because the chances of rejecting the null hypothesis decrease as measurement error increases.

Table 1

Key Elements of the Basic Model for Actuarial Prediction of Sexual Recidivism,

Illustrated with Data For Static-99

| Predictions &<br>Decision Rules | Outcomes | | Efficiency &<br>Base Rate |
|---|---|---|---|
| | Recidivated | Didn’t Recidivate | |
| Will Reoffend:<br>C: 6-9 | 67 | 62 | E (efficiency):<br>67/129 = .52 |
| Will Not Reoffend:<br>L: 0-5 | 204 | 753 | |
| | Sum = 67+204 = 271 | Sum = 62+753 = 815 | All = 271+815 = 1086 |
| | T (sensitivity):<br>67/271 = .25 | F (1-specificity):<br>62/815 = .08 | P (base rate):<br>271/1086 = .25 |

Note. The number in the upper left cell of the 2 x 2 matrix in the center of the table is the sum of recidivists with Static-99 scores of 6 and above (Harris, Phenix, Hanson, & Thornton, 2003, pp. 72 & 79). The number in the upper right cell is the sum of non-recidivists with scores of 6 and above. The number in the lower left is the sum of recidivists with scores of 5 and below. The number in the lower right is the sum of non-recidivists recidivists with scores of 5 and below.

Table 2

Worksheet for Calculating Bayes's Formula

Step 1: Record the Information Needed for the Calculations

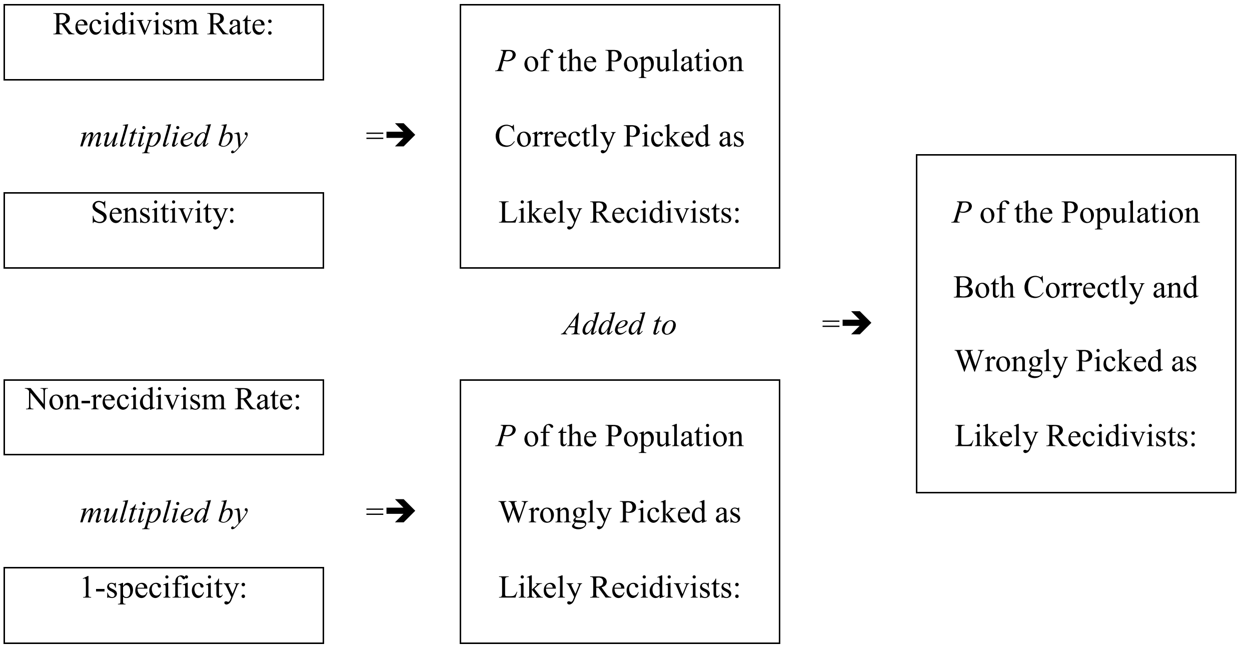
Recidivism Rate (P):

Non-recidivism Rate (Q):

Sensitivity for C (T):

1-specificity for C (F):

Step 2: Determine the Proportion (*P*) of the Population the Test Will Flag as "Likely Recidivists"



Recidivism Rate:

*multiplied by*

Sensitivity:

=➔

*P* of the Population Correctly Picked as Likely Recidivists:

*Added to*

Non-recidivism Rate:

*multiplied by*

1-specificity:

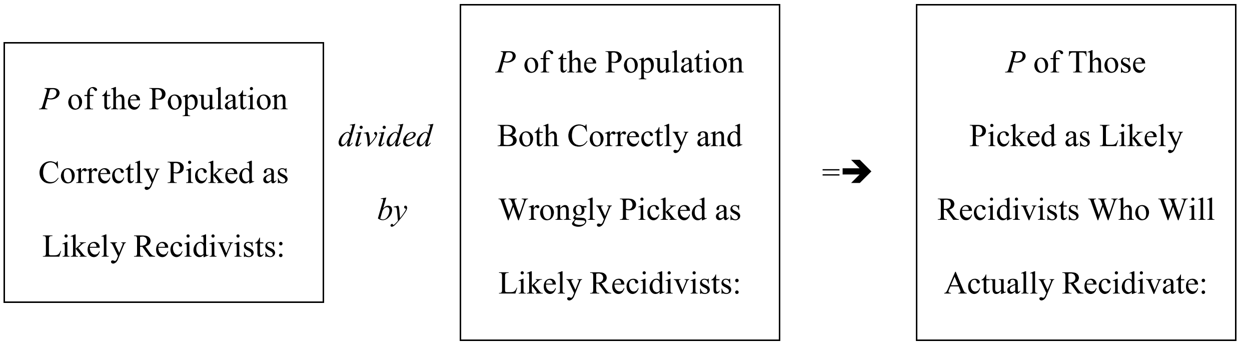
=➔

*P* of the Population Wrongly Picked as Likely Recidivists:

=➔

*P* of the Population Both Correctly and Wrongly Picked as Likely Recidivists:

Step 3: Determine the Proportion of Likely Recidivists in the Population Who Will Recidivate



*P* of the Population Correctly Picked as Likely Recidivists:

*divided by*

*P* of the Population Both Correctly and Wrongly Picked as Likely Recidivists:

=➔

*P* of Those Picked as Likely Recidivists Who Will Actually Recidivate:

Table 3

Recidivism and Age Among Rapists, Molesters, Incesters, and Unclassified Sex Offenders (Extrapolated from Hanson, 2002)

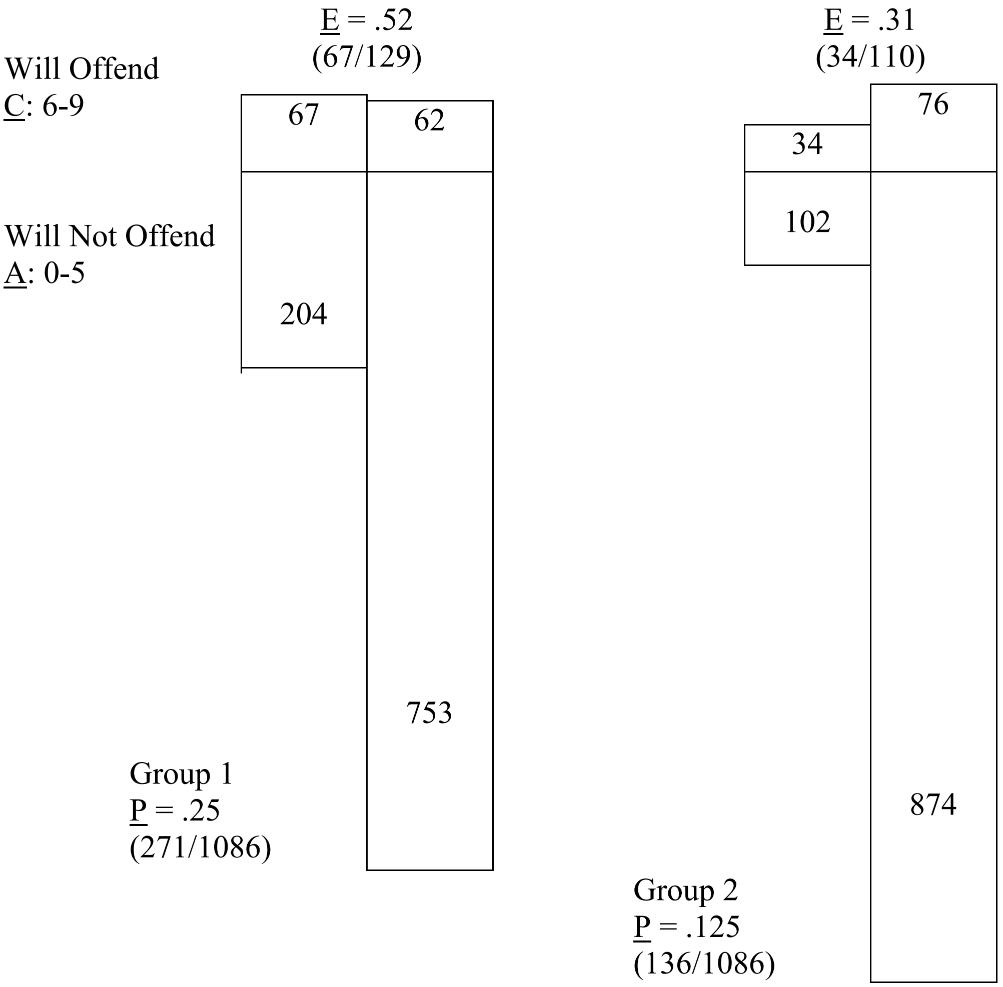
| Age Group | Rapists | | Molesters | | Incesters | | Classifieds | | Initial P[1] | Unclassifieds | | All Subjects | | Final P | SP[2] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | n | r[3] | n | r | n | r | n | r | | n | r | n | r | | |
| 18-24 | 230 | 53 | 195 | 41 | 80 | 25 | 505 | 119 | .235 | 125 | 52 | 630 | 171 | .271 | .018 |
| 25-29 | 290 | 58 | 225 | 58 | 130 | 12 | 645 | 129 | .200 | 160 | 57 | 805 | 186 | .231 | .015 |
| 30-34 | 248 | 42 | 270 | 65 | 220 | 19 | 738 | 126 | .171 | 182 | 55 | 920 | 181 | .197 | .013 |
| 35-39 | 170 | 19 | 215 | 43 | 283 | 20 | 668 | 82 | .123 | 165 | 36 | 833 | 118 | .142 | .012 |
| 40-44 | 105 | 15 | 153 | 29 | 210 | 13 | 468 | 56 | .120 | 115 | 25 | 583 | 81 | .139 | .014 |
| 45-49 | 50 | 6 | 130 | 22 | 132 | 6 | 312 | 34 | .109 | 77 | 15 | 389 | 49 | .126 | .017 |
| 50-59 | 30 | 3 | 157 | 14 | 117 | 6 | 304 | 23 | .076 | 75 | 10 | 379 | 33 | .087 | .014 |
| 60-69 | 14 | 1 | 49 | 2 | 30 | 0 | 93 | 3 | .032 | 23 | 1 | 116 | 4 | .034 | .019 |
| 70 + | 1 | 0 | 5 | 0 | 0 | 0 | 6 | 0 | .000 | 2 | 0 | 8 | 0 | .000 | .000 |
| All[4] | 1,138 | 197 | 1,399 | 274 | 1,202 | 101 | 3,739 | 570 | .153 | 924 | 251 | 4,663 | 823 | .176 | .006 |

[1] P stands for recidivism rate.

[2] SP stands for standard error of the proportion.

[3] r stands for number of recidivists.

[4] Some inconsistencies in cell totals are due to rounding error.



Note. The number in the top left cell for each group indicates the number of recidivists correctly identified by the test. The number of non-recidivists mistakenly classed as recidivists is in the top right cell. The number of recidivists the test misses is in the bottom left cell. The number of non-recidivists that are correctly identified is in the bottom right cell.

Figure Caption

Figure 1. An example showing how the efficiency of Static-99 deteriorates when it is used to identify likely recidivists in a group of sex offenders having a low sexual recidivism rate.

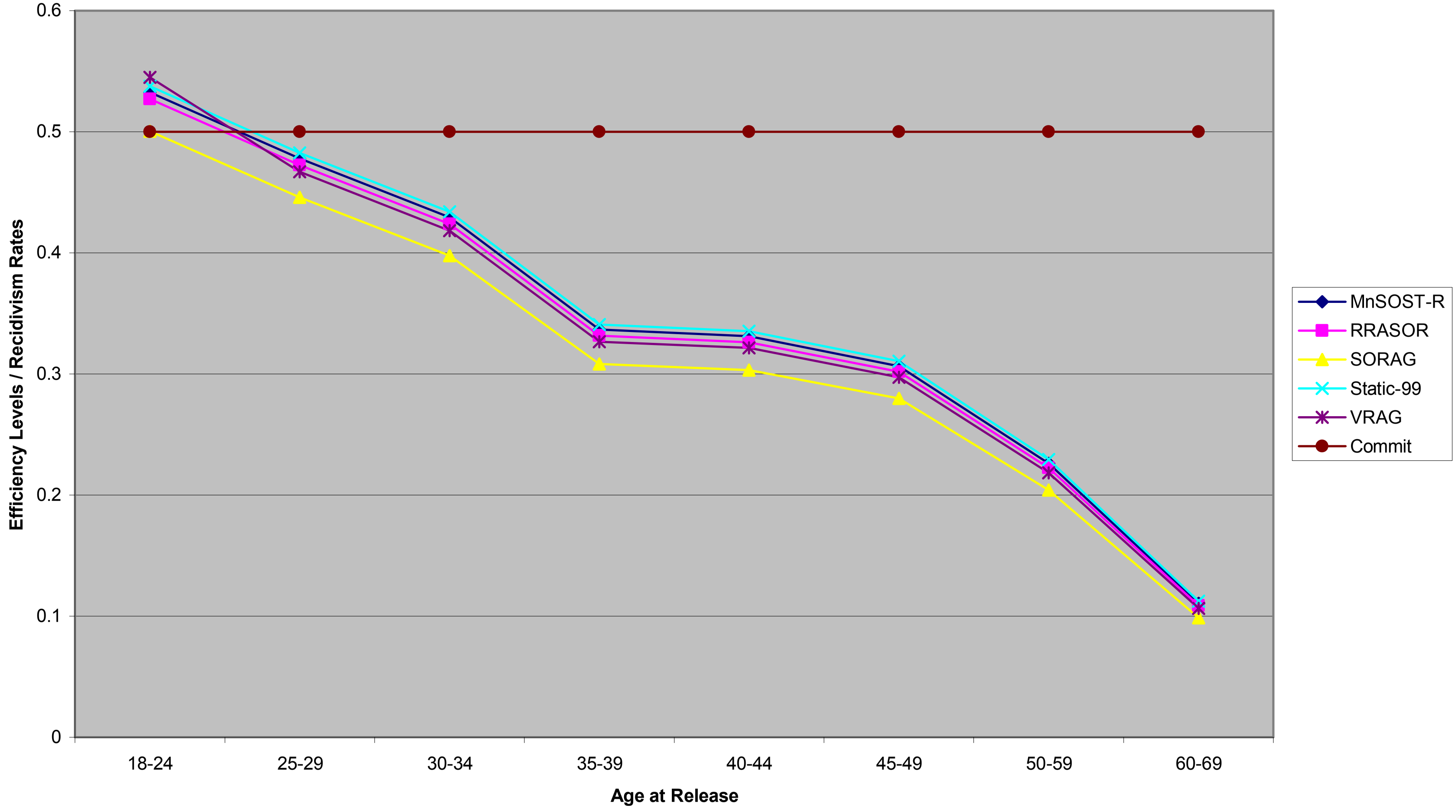

0.6
0.5
0.4
0.3
0.2
0.1
0
Efficiency Levels / Recidivism Rates
18-24
25-29
30-34
35-39
40-44
45-49
50-59
60-69
Age at Release
MnSOST-R
RRASOR
SORAG
Static-99
VRAG
Commit

Figure Caption

Figure 2. Efficiency levels and recidivism rates for subjects classified as likely recidivists from different age groups.