AFFIDAVIT IN SUPPORT OF A *DAUBERT* CHALLENGE TO THE ADMISSIBILITY OF SCIENTIFIC TESTIMONY TO SUPPORT SEXUALLY DANGEROUS PERSON COMMITMENT UNDER THE ADAM WALSH ACT

AFFIDAVIT OF DANIEL KRIEGMAN, Ph.D.

I, Daniel Kriegman do solemnly swear under the pains and penalties of perjury the following is true to the best of my knowledge.

Professional Background

1. I have been a licensed psychologist in Massachusetts since 1981. I received my B.A., with a major in Psychology, from State University of Buffalo. I received my M.A. and Ph.D. in Clinical Psychology from Boston University. I am a member of various professional associations. I have also been a designated Qualified Examiner under M.G.L. c. 123A, responsible for the evaluation of sexual dangerousness for commitment hearings in Massachusetts. My Curriculum Vitae is attached as Exhibit 1.

2. I have significant experience in the area of sex offender evaluation and treatment. Between 1977 and 1986, I was a staff Psychologist, Principle Psychologist, Chief Psychologist, and eventually a Super Chief Psychologist, at the Massachusetts Treatment Center for the Sexually Dangerous. My responsibilities included providing both individual and group psychotherapy, psychodiagnostic testing and preparing background summaries and clinical formulations used for the determination of sexual dangerousness of men who were civilly committed under the sexual dangerousness statute in Massachusetts. In 1982, I was named Unit Director for one of the two units at the Treatment Center and thus supervised all the clinical staff, treating approximately 130 sexually dangerous persons on the unit. Eventually, I became the Director of Supervision and Training responsible for the supervision and training of all the clinical staff at the Treatment

Center. I was simultaneously the Director of Intake and Treatment Planning, responsible for evaluating inmates for possible commitment and for formulating individualized treatment plans if the offender was committed.

3. After I left the Treatment Center in 1986, I was designated a Qualified Examiner. In the late 1980’s, I founded Human Services Cooperative, Inc. (HSC), which contracted with various agencies to provide psychological treatment and evaluation. Around 1989, the Commonwealth of Massachusetts placed all sexual dangerousness determinations (screenings in prison, initial evaluations during a 60-day commitment to determine if a man was sexually dangerous, and periodic reviews to determine if men who had been determined to be sexually dangerous were still so) in a contract and sought a single vendor to provide all sexual dangerousness opinions. HSC won the bid and for several years, under my supervision as President of the Corporation and Clinical Director, provided all of the sexual dangerousness determinations for the Commonwealth. HSC also provided all independent (non-state employee) psychological and psychiatric services at the Treatment Center under a contract with the Department of Mental Health (which ran the institution at the time). These latter services included the two independent psychiatrists or psychologists mandated by the statute to sit on the five member Community Access Board, which performed annual reviews of the treatment and sexual dangerousness status of all the sexually dangerous men in the Commonwealth.

4. Since leaving the Treatment Center in 1986, I have continued to provide psychotherapy to sex offenders. I have also continued to evaluate persons for sexual dangerousness in court proceedings and, since 2002, for the Sex Offender Registry Board. In the three decades (1977 to 2007) during which I have worked with and evaluated sex offenders, I have personally completed well over 400 sexual dangerousness evaluations of sex offenders, have participated on a team

involved in such evaluations on several hundred additional occasions, and have overseen well over 1,000 additional evaluations.

5. I have authored or co-authored or co-edited over 30 published articles and books in the field of psychology including articles specifically related to human sexuality and sex offender treatment. I am on the Editorial Board for *The Journal of Sexual Offender Civil Commitment*. I have been a teaching fellow, led various seminars, delivered papers at international conferences and taught at select institutions.

6. In total, I have personally treated over 100 sex offenders in the course of my career. I have also evaluated or participated in the evaluation of over 1,000 sex offenders for dangerousness in the course of my career. I have testified as an expert in court on over 350 occasions, both for the Commonwealth and for the defense.

7. I have an extremely thorough understanding of the scientific literature regarding sex offender treatment and evaluation. As an undergraduate, I was a teaching assistant leading a section for the required psychology course "Research Methods," (the only time an undergraduate ever performed such a function at SUNY Buffalo, where I graduated Magna Cum Laude with a B.A. in Psychology). I have been a teaching assistant at the graduate school level and have tutored Ph.D. candidates in statistics and research methods. I have served as a peer review editor for several scholarly journals in psychology.

8. My experience and personal knowledge are based on the Massachusetts commitment scheme for sexually dangerous persons, one of the oldest commitment laws in the nation. When relevant, I will refer to pertinent experience from Massachusetts as a way to explain the problems associated with predicting dangerous behavior and determining sexual dangerousness.

Early History About the Evaluation of Sex Offenders

9. Psychologists have long understood the limitations of predicting human behavior. Though we have been called on to opine whether someone is dangerous, and in recent years, sexually dangerous, historically, the profession had come to consider it impossible to make such predictions with a substantial degree of accuracy. Thus, the main professional associations (the American Psychological Association and the American Psychiatric Association) have determined that offering such clinical expertise in the courtroom is often unethical unless the clinician makes the very limited validity of such opinions clear. Since this rarely occurs, the main professional associations for psychiatrists and psychologists (the APA's) have questioned the ethics of offering such "expertise" in the courtroom.

10. Up until the 1980's, the predominant tool in use for making predictions of future sex offenses (or other forms of violent, dangerous behavior) was clinical judgment. Clinical judgment is based on a clinician's subjective sense of an offender when the specifics of the offender's life and the experience of interacting with the offender is "mulled over" in the mind of a clinician who has had experience working with other sex offenders. However, the reliability of clinical judgment was always in doubt and when empirically tested by seeing if clinicians could accurately predict which offenders will reoffend, the results were, on average, approximately equal to tossing a coin.

11. Professor Meehl was one of the early explorers attempting to determine the accuracy of clinical prediction. His work provided the foundation for later understanding of this issue.[1] Around the same time, an early classic study by Professor Goldberg demonstrated that novices were often able to formulate clinical opinions with equal accuracy to those formulated by seasoned

[1] A bibliography listing the articles referenced in this Affidavit and the attached Exhibits is included as Exhibit 12.

professionals. This finding has been replicated many times.

12. By 1981, Professor Monahan published his seminal work establishing that clinical predictions of dangerousness were wrong 2 out of 3 times; clinicians are often poorer predictors than lay people. I have provided a more detailed historical background in Exhibit 2 (History of Clinical Prediction Studies).

Use of Actuarials

13. The early research about clinical judgment spurred new fields of research into the effectiveness, and creation of, alternative methods of prediction. Actuarials have long been used in other contexts, such as insurance. In the 1980's, psychologists began researching and developing actuarials in the field of violence prediction.

14. Before getting into the history of actuarials, it is useful to understand how they work and what their limitations are.

15. Actuarial tests—of which life insurance actuarial methods are excellent examples—take a number of predictive factors that are each known to be independently and truly related to, for example, longevity. Age, weight, smoking behavior at time of application for insurance, among other factors, are each assessed and scored according to the magnitude of each. The different factors are *weighted* to emphasize more powerful predictors (e.g., age) over weaker predictors. And then the factors are combined according to an actuarial formula to produce a prediction. For example, all other things being equal, a sixty-year-old man has a shorter life expectancy than a thirty-year-old smoker; age would influence the risk estimate more than smoking. However, a 72-year-old smoker would have a shorter life expectancy than a 75-year-old nonsmoker; in this case, smoking would exert more influence than the age difference. The *actuarial formula* combines and

weights the various valid predictors to produce predictions that maximize accuracy in a large, real-life sample. The formula is “tweaked” and reworked until predictive accuracy is maximized. It is then cross-validated on new samples and further modifications in the actuarial formula are made to increase the predictive accuracy in a wide range of relevant samples.

16. There are several important statistical components to actuarials which allow a psychologist to apply them in the appropriate contexts. It is important to have a rudimentary understanding of the various components of an actuarial in order to appreciate its usefulness and limitations.

17. The relevance of all actuarials—indeed, all opinions regarding sexual dangerousness—begins with a base rate. No facts about an individual (person or event) can have any meaning unless they are compared to the class of similar facts about similar individuals and events. That is, it is very rare that the predictive meaning of an event can be interpreted without knowledge of base rates. Base rates are, very simply, a reference to the class of other similar events.

18. Base rates are important for two reasons. First, the lower the base rate, the harder it is to be accurate when concluding that the target behavior is likely (in this case, future sex offending) and the more likely that false positives (false conclusions of dangerousness) will be produced than accurate “hits.” If the percentage of convicted sex offenders released after attaining the age of fifty who recidivate is 5%, it will be extremely hard to predict accurately which persons over the age of fifty will recidivate (correct predictions of dangerousness) without including a larger number of non-recidivators (false positives); on the other hand, if the percentage of persons who have multiple convictions for sex offenses and are under the age of 25 who recidivate is over 50%, it will be far easier to be right more often than not when making the prediction that a specific offender from this group will reoffend.

19. Additionally, base rates are important because they help determine when and how to apply a specific risk assessment to a given case. For example, if a base rate were obtained by studying offenders with a prior conviction, over a certain age, and with a male victim, such a base rate would be expected to produce accurate results when applied to offenders with the same characteristics. To the degree that an offender does not match the group on which a base rate was determined, that base rate would be misleading if applied to the offender in question.

20. Actuarials are able to calculate base rates by combining and calculating a number of significant variables in a manner that increases the accuracy that could be obtained by using any one factor. For a more complete discussion and explanation about base rates, see Exhibit 3 (Base Rates).

21. The highest recidivism estimate ever produced based on empirical observations of what actually occurs when sex offenders are released and followed for 25 years (approximately 50% for a certain category of offenders) comes from Prentky, et al. (1997). The Prentky study was based on men who had been adjudicated sexually dangerous in Massachusetts and subsequently released from prison. See Exhibit 4 (Prentky Study). This study, however, is likely to be an overestimate of recidivism for a number of reasons.

22. It is more commonly believed that the base rate for sex offender recidivism as a whole lies between 15 and 35%. This base rate fluctuates depending on various factors. For example, studies consistently confirm that older offenders have a significantly lower base rate; on the other hand, offenders under 25 have a higher base rate. Extra-familial child molesters have a higher rate of recidivism than men whose victims were adults. Both of these groups have a higher rate of recidivism than incest offenders, who have the lowest rate of recidivism in this simple

classification schema. In all, sex offenses have been shown to have the *lowest* rate of recidivism of all common offense types (only murderers recidivate at a substantially lower rate).

23. There are other aspects of actuarials that are important in understanding their limitations. Actuarials are able to tell us how well a given set of traits or characteristics correlate with the studied outcome, in our case, recidivism. The correlation can be measured by a correlation coefficient, (denoted by *r*). Consider that the absolute value of *r* can range from 0.0 to 1.0, with 1.0 being a "perfect" relationship and 0.0 being no relationship at all, i.e., 0.0 is the equivalent of what one would get by tossing a coin and saying "Heads, he'll reoffend; Tails, he won't." That is, if the correlation between two variables yields an $r = 1$, then knowing how an individual scores on one variable tells you perfectly how they will score on the other.

24. Correlation coefficients between 0.0 and 1.0 have different predictive values. A correlation coefficient less than .l0, i.e., an *r* close to zero, has little predictive value and virtually no utility in practical applications. A correlation coefficient less than .20 has such a small amount of predictive value that, in a real life application in a small number of cases, it would be hard to see a difference between tossing a coin and using such a predictor. Between .20 and .40 there is an increase in accuracy (over chance or coin tossing) that would be perceptible. From .40 to .70 there are clear advantages to using the predictor. Above .70 we have a fairly strong predictor that is far better than random guessing.

25. By combining predictive factors with correlations between .10 and .25, it is possible to create actuarial tests that produce correlations between .20 and .40 with sex offense recidivism. Experts agree that this type of actuarial prediction has been conclusively shown to be the essence of the most valid predictive methods currently available; nevertheless, this actuarial prediction is not very strong (valid). Given the aforementioned problem of making predictions about

phenomena with low base rates, it becomes vital to understand the real but limited validity of our best predictive method in order for a court to be able to decide if this is sufficiently valid to be used in a civil commitment proceeding. For a more detailed discussion of correlation coefficients, see Exhibit 5 (Correlation Coefficients, Validity, and Variance).

26. Given large enough samples, it is possible to show that weak correlations with no practical utility are real, and that therefore the small correlation is unlikely to be just a random deviation from chance. This means that if we examine a different large sample, we can expect to find the correlation again and again. However, when comparing small groups of individuals or single individuals with or without the weakly correlated factor, there will be no discernable difference between the presence or the absence of the factor. That is, with small groups (a small sample), we won't be able to perceive the increase in accuracy obtained by using a weakly correlated factor (e.g., $r = .10$) over the random variations we will produce when tossing a coin ($r = 0.0$) a small number of times. Thus, simply knowing that a given trait has a significant (or real) correlation does not, in and of itself, tell us how much more accurate a prediction will be when the presence (or absence) of that trait is utilized in making a prediction.

27. However, given the correlation, there is a way of determining how *valid* a prediction is. Validity means how much we can count on knowledge of one variable (clinical or statistical assessment of risk) to tell us what in fact will happen when we look at another variable, in this case actual recidivism. This is known as the percent of variance (variation in outcome) in an outcome predicted by a variable predictor . For a more detailed discussion of validity and variance, see Exhibit 5.

28. The leading researcher in the field of sex offender recidivism is Karl Hanson. In 1998, he published a meta-analysis of all known sex offender recidivism studies. That meta-analysis (and a

follow-up published in 2004) became the foundation for modern risk assessment in the field of sex offense recidivism. In those analyses, factors that had been empirically demonstrated to be correlated with recidivism were identified. Numerous factors that had been *thought* to be correlated but which are not predictive were also identified. Based on this information, Dr. Hanson and a colleague developed an actuarial tool called the Static-99, which was based on formulas taking into account various characteristics (factors), each of which had been shown to demonstrate some predictive value.

29. Since the development of the Static-99, Dr. Hanson has also led the way in studying the effects of age and recidivism, demonstrating that age seriously decreases the rate at which sex offenders reoffend.

30. The development of actuarials has changed the way we think about risk prediction. In its increased accuracy over clinical opinion, the actuarials highlight how unreliable clinical judgment is; yet, the actuarials produce only limited validity of their own. Actuarials help demonstrate that in the real world of sex offense recidivism prediction, the best predictive method has resulted in only a small degree of validity. Furthermore, this greater validity over clinical opinion highlights how very little correlation there is between an expert's clinical opinion that someone is sexually dangerous and whether that person will actually recidivate. For a more detailed discussion of the specific application of the broad statistical concepts to these hearings, see Exhibit 6 (Specific Application of Statistical Concepts to Findings of Sexual Dangerous).

31. Actuarials also reveal the extent of their own limitations. Although it is true that the actuarial method produces results that are *much* more accurate than clinical judgment (indeed, the best actuarials may be more than ten times more accurate than clinical judgment), the predictive power is far from perfect. A perfect predictor would account for 100% of the "variance" in an

outcome (in this case, the “variation” is reoffend or does not reoffend). A useless predictor, e.g., tossing a coin, will account for 0% of the variance. Clinical judgment accounts for less than ½ of 1% of the variance, making it almost useless. If the actuarials can predict 10 to 20 times more accurately than clinical judgment (which they can), this does *not* make them highly accurate. In fact, they have been shown to have a correlation of between .3 and .35, thus predicting just 9 - 12% of the variance, leaving about 90% unaccounted for. For a more detailed example, see Exhibit 7 (Application of Actuarial Correlations).

32. Today, variations of clinical judgment and the use of pure actuarials have evolved. Some professionals use a method called guided clinical, which calls for the clinician to evaluate a list of factors with valid (real) small correlations with recidivism. The clinician then uses clinical judgment to combine the information from the list of factors into a single risk estimate. This method has demonstrated validity between the use of pure, unstructured (unguided) clinical judgment and the use of actuarial prediction. Others use a method called the adjusted actuarial method, which uses actuarials to establish a base rate for a typical group of offenders similar to the offender in question. The clinician is then free to adjust this risk estimate by using clinical judgment to raise or lower the risk estimate when unvalidated factors (factors believed to be correlated with recidivism that have not been researched sufficiently to determine if they are so correlated) or other validated factors not taken into account by the actuarial formula are present. However, none of these have proven to be substantially more accurate than pure clinical judgment and limited research indicates that they are not as accurate as the pure actuarial method. Since the actuarial method has limited accuracy, these weaker methods cannot produce the level of accuracy needed for confident conclusions and the use of these alternate methods has not gained acceptance in the psychological field. For a more detailed explanation of these variations, see Exhibit 8

(Clinical Variations). It still remains the case that simply using an actuarial upon which to base an opinion of dangerousness is the most accurate of the various methods of prediction, none of which is highly accurate.

33. Since the accuracy of our most accurate methods of prediction is poor, we can rarely have a sufficient degree of professional certainty that future offenses are likely to occur when: 1) the base rate is unknown but likely to be fairly low, between 15 and 35%, 2) we know that the base rate does not predict very reliably in many cases, 3) the factors we can use to modify the base rate are weak predictors with fairly low validity, and 4) many years have passed since the last sexual offense. See Exhibit 9 (The Psychological Limitations of Opining Sexual Dangerousness) and Exhibit 10 (Why an Actuarial Prediction Can Never Be a Sufficient Basis for a Finding of Sexual Dangerousness).

Jimmy Ryce Civil Commitment Program

34. Under the Program, a person declared sexually dangerous may be released if the "Director of the facility in which a person is placed pursuant to subsection (d) determines that the person's condition is such that he is no longer sexually dangerous to others, or will not be sexually dangerous to others if released under a prescribed regimen of medical, psychiatric, or psychological care or treatment." This presents several problems already addressed above. For one, if psychologists themselves cannot accurately predict whether someone will recidivate, and indeed over-predict recidivism more often than not, there is no reason to believe a Director has an accurate predictive ability.

35. My experience with the Massachusetts commitment scheme is that the Department of Corrections has never filed a petition for release and opposes an inmate's petition for release

approximately 99% of the time, even after offenders have been civilly committed and treated for over a decade. Therefore, the constraints inherent in a psychological prediction are more apparent when the Government is responsible for declaring someone is no longer dangerous (as opposed to a state expert declaring someone is). For a more detailed discussion about an inmate's release in Massachusetts, see Exhibit 11 (Release from Commitment in Massachusetts).

Conclusion

36. In conclusion, in three decades of working with sex offenders, evaluating them for sexual dangerousness, and studying the research—the quantity of which has exploded in the last 15 years—it has become clear that we simply have no ability to accurately predict sex offense recidivism for individuals.

Date: May 16, 2007

/s/ Dr. Daniel Kriegman_

Dr. Daniel Kriegman

[Curriculum Vitae of Dr. Kriegman is the first of the exhibits followed by Exhibits 2 through 12. Dr. Kriegman's CV is at the end of this document.]

History of Clinical Prediction (Exhibit 2)

Up until the 1980's, the predominant tool in use for making predictions of future sex offenses (or other forms of violent, dangerous behavior) was clinical judgment. Though there was a mounting body of research that challenged this approach, there were no well established, generally accepted reasons to question this methodology; it certainly seemed reasonable to turn to experts who had treated and studied sex offenders when it came time for making release decisions. Surely, they were more likely to know whether an offender was dangerous than someone whose first contact with an offender was as a juror. After years of experience—sometimes with hundreds of sex offenders—and careful study of the literature that built upon their years of formal training, such experts were also expected to have more pertinent wisdom regarding this matter than a judge who presided over a relatively small number of such cases.

This was still the prevailing view when—despite the testimony that was presented to challenge the prevailing wisdom—the Supreme Court issued its *Barefoot* decision. The challenge failed to convince the SC of the unscientific and invalid nature of the predictions it was accepting even though, years earlier, researchers had begun to show that assumptions of such clinical expertise were usually *invalid*. The assumed expertise was *almost always* found to be an illusion whenever it was subjected to close examination. For example, in an early classic study examining clinical expertise, Goldberg (1959) asked novices and highly experienced, Ph.D. psychologists to look at MMPI[2] scores from individuals tested prior to their discharge from a mental hospital. The clinicians were asked to use the MMPI results to determine the severity of the discharge diagnosis (neurotic vs. psychotic, a much simpler distinction than will vs. won't commit a new sexual offense). Goldberg had previously constructed an actuarial rule based on his studies of the MMPI profiles of patients whose discharge diagnoses he already knew. The formula—which since has come to be known as "Goldberg's Rule" (Dawes, Faust, and Meehl, 1989)—was to add up the scores on three MMPI scales and subtract the scores on two others. According to Goldberg's (simple) Rule, if the result was 45 or more, it was predicted that the individual would be discharged with a psychotic diagnosis. The vast amount of information available in the rest of the MMPI and the complex patterns formed by the many scales were ignored in this simple actuarial measure.

The result? The novices and the experts using "clinical judgment" performed similarly averaging 62% correct diagnoses:

[2] The MMPI is *the* most researched and studied psychological test; there have been thousands of carefully controlled studies using the MMPI and it has been administered to many hundreds of thousands of people. It consists of approximately 560 true-false questions about a wide variety of behaviors, feelings, beliefs, values, experiences, etc. The pattern of answers an individual gives is then compared to see if it matches patterns obtained when the test was given to carefully selected groups, e.g., paranoid schizophrenics, severely depressed people, individuals with obsessive-compulsive disorder, etc. Scales have been developed that measure syndromes or character patterns like "depression"; for example, if a person answered "true" to a question like "I often wake up dreading the day ahead," he would get one point for a depression scale. If the answer was "false" to something like "Every now and then I have a really good day," one point would be scored on the depression scale. 20 or 30 items would be selected for such a scale based on the pattern of actual responses that had been given by people who were known to be suffering from depression. The higher the score on the depression scale, the more depression it indicates. Hundreds of scales have been developed including paranoia, hypochondriasis, phobias, social conformity, etc. Patterns of high scores on the different scales scored are then used to make more specific conclusions about individuals.

> One surprising finding—that amount of professional training and experience of the judge does not relate to judgmental accuracy—has appeared in a number of studies.[3] (Goldberg, 1959)

The single most accurate judge made 67% correct decisions. The Goldberg Rule achieved an accuracy rate of 70%. Goldberg then trained the judges by giving them 300 MMPI profiles to practice with and providing immediate feedback on accuracy. Even after 4,000 practice judments, *none* of the judges equaled the Goldberg Rule. Finally, Goldberg gave some of the judges (including all of the experts) the information from the Goldberg Rule on each case and allowed them to use the extra information (that the rule in its simplicity could not make use of) and their clinical judgment to use or modify the prediction made by the rule. Though there were some gains in accuracy, no judge did as well as the rule: Every judge would have been more accurate if they avoided using clinical judgment and always used the rule alone.

Doubt about clinical judgment built upon the work of Meehl (1954) who had started the wave of controversy when he first questioned the expertise of experts in predicting human behavior. Meehl's book served the same function as the child who dared to question the emperor's new "clothing." Slowly, the research questioning the clinical expert's, legal clothes picked up momentum so that, by the 1970's the tide among *scholars* (i.e., scientists and researchers who were not practitioners who made a substantial part of their livelihood from making clinical predictions) was beginning to turn.

> Many clinicians have been making unreliable and invalid judgments based on invalid premises, illogical assumptions, unproven relationships, inappropriate applications of unproven theories and other types of error. (Thorne, 1972)

By the late 70's, the waves of mounting evidence began to take effect resulting in the "Report of the Task Force on the Role of Psychology in the Criminal Justice System."

> [T]he validity of psychological predictions of violent behavior [is] . . . so poor that . . . one could [say] . . . that psychologists are not professionally competent to make such judgments.

But it wasn't until the early 80's that the tide began to turn decisively. Monahan (1981), in an influential work, reported that clinicians were wrong in two out of three predictions of violence. The *amicus* brief filed by the American Psychiatric Association in 1983 (*Barefoot*) was based, in part, upon Monahan's work. This is where the APA concluded that "psychiatric predictions of long-term future dangerousness are wrong in two out of every three cases."

By the mid 1980's, the new understanding had begun to reach general acceptance among the scholars and professional associations. At this time, Meehl (the fellow who started all the trouble) could look back and, comparing actuarial versus clinical prediction, conclude:

[3] And it has since been confirmed in many studies (Dawes, et al., 1989; Goldberg, 1959; Menzies, Webster, & Sepejak, 1985; Monahan, 1981; Faust & Ziskin, 1988; Mossman, 1994; Menzies, Webster, McMain, Staley, & Scaglione, 1994; Rice & Harris, 1995).

> There is no controversy in social science that shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one. (1986, pp. 373-374)

There is a tendency now to talk of Pre-Monahan (1981)—again, Monahan was the researcher that established that clinical predictions of dangerousness are wrong 2 out of 3 times and that clinicians are often poorer predictors than lay people—and Post-Monahan research. By the end of the 1980's, Dawes, et al. (1989) captured the prevailing view that had finally been established—general acceptance had become clear—among scientists in the field. They reviewed and summarized the studies comparing clinical and actuarial judgment. They concluded that there were “nearly 100 comparative studies in the social sciences. In virtually every one of these studies, the actuarial method has equaled or surpassed the clinical method” (p. 1669). By the early 1990's, this was no longer controversial: All of the experts who study methods of predicting this type of human behavior have concluded that actuarial/statistical prediction is more accurate than clinical prediction. This is so well established, I have been unable to unearth a single violence prediction study carried out in the last decade that even bothers to study or use the old-style clinical prediction. The standard goal of research has now become improving statistical prediction.

Many of the professionals *who make their living* from testifying about future dangerousness do make use of clinical prediction of the type that has been discredited (see Thorne, 1972; Epperson, Kaul, and Hesselton, 1995: Rice & Harris, 1995; Borum, 1996; Rice, 1997), though they now tend to cloak this clinical prediction under the rubric of the “guided clinical” or even the “adjusted actuarial” method. This should not be taken to mean that there has been *any* acceptance in the general psychiatric or psychological community of the validity of such clinical methods. In fact, the reverse is true. Clinicians who engage in this work comprise a tiny fraction of practicing clinicians. And their professional associations have not changed their positions from the highly skeptical and ethics questioning *amicus* brief (1983) submitted by the American Psychiatric Association in *Barefoot* and the 1978 “Report of the Task Force on the Role of Psychology in the Criminal Justice System” published in the American Psychological Association’s primary organizational journal (*American Psychologist*, 1978), which stated:

> the validity of psychological predictions of violent behavior, at least in sentencing and release situations we are considering, is extremely poor, so poor that one could oppose their use on the strictly empirical grounds that psychologists are not professionally competent to make such judgments.

More recently, one finds these positions affirmed again and again in the same organs of these associations. For example, in the *American Psychologist* we find:

> Assessments of dangerousness made by clinicians continue to ignore the research on the prediction of violence. (Borum, 1996)

> Most assessments of dangerousness were (and still are) based exclusively on unaided clinical judgment. In almost every situation in which they have been studied, actuarial predictions have outperformed unaided human judgement

> (Meehl, 1954, 1986, 1996). . . . Mossman (1994) recently showed again the superioritv of actuarial methods over clinical methods for long-term predictions in a meta-analysis of studies of predictions of violence specifically." (Rice, 1997, p. 416)

The American Psychiatric Association Task Force Report On Sexually Dangerous Offenders (Zonana, et al,, 1998) and a Brief of *Amicus Curiae* American Psychological Association (DeBruin, 2005) also demonstrate a fundamental lack of change between their more current professional associations' stance and the psychiatric association's 1983 *amicus* brief and the psychological association's 1978 task force report.

Furthermore, just as Meehl noted the unequivocal uniformity of the empirical results, there is virtual unanimity among those who study the accuracy of behavioral prediction, in general, and the prediction of violence and/or sexual offenses, in specific. Actuarial/statistical prediction beats clinical prediction every time.

> Indeed, a large body of literature over the last several decades has consistently demonstrated the general superiority of actuarial prediction over clinical prediction in virtually every decision-making situation for which the issue has been studied. (Harris, Rice, & Quinsey, 1993)

> In virtually every decision-making situation for which the issue has been studied, it has been found that statistically developed prediction devices outperform human judgments. (Gottfredson, S. 1987, p. 36)

> [T]he solution to improved violence prediction is the same as for the improvement of clinical predictions in general—the use of actuarial methods. (Epperson, et al., 1995)

Based on a meta-analysis (52 studies of 16,191 persons) of predictors of general and violent recidivism, Bonta & Hanson (1998) found criminal history was the best predictor and clinical factors were the worst. (Also see Andrews & Bonta, 1998; Dawes, Faust, & Meehl, 1993; Gottfredson & Gottfredson, 1994; Grove & Meehl, 1996; Grubin, 1999; Hanson, 1998; Hanson & Bussiere, 1998; Hanson & Harris, 2000; Hanson & Thornton, 1999; Howe, 1994; Janus & Meehl; 1997; Milner & Campbell, 1995; Monahan, 1984, 1992, 1995, 1996; Webster, 1994; Quinsey & Maguire, 1986; Quinsey, et al., 1998).

In a more recent meta-analysis of 136 studies—i.e., *all* of the studies that the authors were able to find that compared clinical versus mechanical prediction and met minimum standards that would enable a comparison to be made, Grove, Zald, Lebow, Snitz, & Nelson (2000) concluded that "mechanical predictions of human behaviors are equal to or superior to clinical prediction for a wide range of circumstances" (p. 19). Note that:

> The only design variable [of the studies] that substantially influenced the relative efficacy of the mechanical- and clinical-prediction methods was whether the clinicians had access to a clinical interview. *Alas, clinical predictions were*

> *outperformed by a substantially greater margin when such data was available to the clinician.* (p. 25, emphasis added)

The 136 studies examined attempts to predict phenomena as diverse as the diagnosis of small bowel disease, business startup success, homosexuality, career satisfaction, and performance in medical school. The authors categorized the studies as falling into one of the following six categories: educational, financial, forensic, medical, clinical-personality, and other. Note that the single category that showed *the largest difference between clinical and mechanical prediction*, i.e., the *greatest* superiority of mechanical prediction, was in the forensic category. Of the 136 studies, there were ten studies that compared clinical and mechanical prediction in forensic settings and two looked at issues that might have forensic implications. These 12 studies were described as attempts to predict probation success, criminal behavior, parole success or failure, assaultive behavior, lie detection, juvenile delinquency, assault by psychiatric inpatients, juvenile criminal recidivism, and 4 studies of criminal recidivism. In 11 out of the 12 studies, mechanical prediction was superior to clinical judgment. In one study of probation success, clinical judgment yielded a 96% hit rate beating mechanical prediction, which yielded a 95% hit rate. Given a choice, the empirical data is unequivocally clear: (1) in forensic settings, you are almost certain to do better with mechanical prediction; (2) in the rare situation in which this is not so, you will do as well; and (3) any reliance on clinical interviews makes the clinical method even less accurate.

And there is evidence that when using non-actuarial methods—i.e., clinical, subjective judgment—that clinicians are no better than and often are *worse than lay persons*. For example, given enough information, laypersons making clinical predictions are as accurate as clinicians, and when there is a difference, lay people usually make better predictions.

> In the particular case of predicting violence, it is well documented that mental health professionals possess no special expertise in the prediction of violence, and that reliance on clinical judgements alone results in numerous inaccurate predictions of violent recidivism. (Rice and Harris, 1995)

Menzies, Webster, McMain, Staley, & Scaglione (1994) studied the accuracy of clinician and layperson predictions of dangerousness among Metropolitan Toronto Forensic Service patients using the Dangerous Behavior Rating Scale (DBRS). Three outcome measures were used: violent behaviour, criminal behaviour and general incidents. They found that clinicians were no better than laypersons at assessing risk; in fact, laypersons were better at using the DBRS than clinicians. (Also see Dawes, et al., 1989; Goldberg, 1959; Jackson, 2004; Menzies, Webster, & Sepejak, 1985; Monahan, 1981; Faust & Ziskin, 1988; Mossman, 1994).

Hanson (1998), a pioneer of and outspoken advocate of the statistical model, noted that guided clinical assessment cannot be ruled out because there were two (2) studies that did almost as well as the actuarial studies. In guided clinical assessment—in contrast to pure clinical method—only factors that have received empirical support are assessed. How the empirically validated factors are then weighted to reach a decision is left to clinical judgment. However, the larger of the two “guided clinical” studies Hanson cited was Epperson, et al.’s development study of the Minnesota Sex Offender Screening Tool (MnSOST). The MnSOST was actually an actuarial measure in which the evaluators reviewed the files in order to score the items in the actuarial tool. When they were done and had scored the MnSOST—i.e., after they had worked through each of the variables assessed in the actuarial tool and knew exactly what the actuarial measure predicted—they were asked to use clinical judgment to modify the actuarial prediction if they thought it should be adjusted. This was therefore an adjusted actuarial tool, not a guided clinical measure, using Hanson's own definitions.[4] The correlation between the actuarial MnSOST score and recidivism was .27. In accord with the view I am presenting, when the evaluators were allowed to use their judgment to adjust the actuarial prediction, accuracy fell to .20.

So, the larger of the two studies supporting the guided clinical method was actually an adjusted actuarial study that, once again, showed the superiority of actuarial over clinical input. Furthermore, the MnSOST has been replaced by the MnSOST-R. The details of the revision are quite telling in this regard:

> The second element of this revision was a change to empirical methods for item selection and scoring. The previous MnSOST produced a total score that was used in an actuarial manner, but the scoring of individual items was clinically based. In contrast, the MnSOST-Revised (MnSOST-R) utilized empirical methods for item selection and scoring. Given the strong support for the general superiority of systematically derived empirical risk assessments over intuitive or even trained clinical predictions, as summarized above, it was assumed that the predictive validity of the MnSOST-R would be significantly improved by using empirically based, rather than clinically based, item selection and scoring. (Epperson, et al., 1999)

This proved to be the case. So, the single major "guided clinical" assessment tool has been *improved by eliminating clinical judgment entirely*. And the accuracy of the only other study to show validity for guided clinical judgments was accidentally exaggerated.[5]

[4] When I brought this to Hanson's attention, he agreed:

> On reflection, I agree with your characterisation of the 1995 MnSOST study as an "adjusted actuarial" approach. The evaluators had scored the MnSOST, and they were allowed to form their own overall judgement. It was not clear to me on reading the original report whether the MnSOST scores were calculated by the raters or whether they only rated the items. Epperson's explanation seems reasonable and I will change my characterisation of the study accordingly. (Hanson, personal communication, emailed 5/21/01)

[5] A calculation error led Hanson to the conclusion that the accuracy of that study was .29 when in actuality it was .12 (the error accidentally multiplied the validity of that study by a factor of 6, from 1½% of the variance accounted for to 9%):

> I had originally calculated the accuracy of the Smith and Monastersky study as about .27, but I made a calculation error (pointed out to me by Grant Harris). The correlation is closer to .17 - I would have to go back and check the exact number, but it was less than .20. (Hanson, personal communication, via email, 5/14/01).

> I checked my notes on Smith & Monastersky. The original (erroneous) correlation was .29. It should be .12. This was a simple calculation error that was only pointed out to me after the 1998 article was published. *The actual correlation was pretty much what you would expect*

One reason why psychologists predict poorly is the result of using factors that have been shown empirically to be *unrelated* to future events; clinical experts typically use their theory of what they *believe* (without corroborating empirical evidence) predicts future behavior. For example, experts frequently opine that *denial* of the offenses or details of the offenses indicates sexual dangerousness. Or that clinicians' evaluations of whether or not the offender shows *empathy* for his victims can be used to make accurate predictions. However, when one looks at all the studies that have evaluated denial and empathy prior to release and then looked at whether there was any difference between the recidivism rates of deniers (vs. admitters) or empathizers (vs. those that lack empathy), one finds no correlation.[6] The reality is that clinicians' evaluations of such things have no predictive utility whatsoever despite their clear beliefs to the contrary:

> [M]any of the frequently used risk assessment procedures have questionable validity . . . [T]he accuracy of [clinical] risk assessments has been unimpressive . . . The clinical prediction of sexual offender recidivism is no exception. Across 10 studies (N = 1,453) that examined the predictive accuracy of clinical judgments concerning sexual offender recidivism risk, the average correlation was only .10[7] . . . Evaluators assessing the long-term risk for recidivism can be

> *from clinical assessments.* (Hanson, personal communication, via email, 5/23/01, emphasis added).

Taken with his conclusions that the MnSOST study was adjusted actuarial and not guided clinical (see discussion above), there are no findings in Hanson & Bussiere's 1998 meta-analysis (or at that time, anywhere else) that offered any support for guided clinical assessments (and, again, the states' experts typically use *unguided* clinical assessments that are even less accurate, though they now often call their method the "guided clinical method").

[6] Hanson, R. K. and Bussiere, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology*, 66, 348-362. (Used statistical techniques to combine all known studies on factors related to sex offender recidivism to determine what the empirical evidence tells us about what factors are predictive and how reliable they are. The 61 studies included 23,393 offenders. This overview is thus the "state of the art" summary of what we know about predicting sex offender recidivism.)

[7] "Correlations less than .10 would have little practical utility in most settings" (Hanson and Bussiere, 1998, p. 351). In actuality, Hanson and Bussiere found a correlation of .07 with clinical judgment when the two studies using "guided clinical judgment" were removed. Those studies had an average *r* = .23. However, the larger of the two "guided clinical" studies was Epperson, Kaul, and Huot's (1995) MnSOST development research. For technical reasons (because of oversampling of recidivists in their development sample), the correlations between measure and recidivism can be assumed to be inflated. Furthermore, Epperson (personal communication, May, 2001) noted that the "clinical judgment" was tightly controlled and "highly guided": The clinical judges were the scorers of the MnSOST and were forced to focus on the empirically validated factors, to carefully assess them, and to derive the actuarial score *prior* to using any clinical judgment. Because clinical judgment *lowered* accuracy (.27 without and .20 with clinical judgment), Epperson, et al. removed it when they developed the revised version, the MnSOST-R, a pure actuarial tool (Epperson, Kaul, and Hesselton, 1999) was not really a "guided clinical" assessment as Hanson (1998) defined the term. The MnSOST was actually an adjusted actuarial tool that was improved, in part, by *removing* all clinical judgments (eliminating the adjustment) when it was updated to become the MnSOST-R. When I brought this to Hanson's attention, he agreed with my classification of the studies.

While adjusting actuarial assessments may at times be necessary and can be reasonably defended in certain circumstances, there is no empirical evidence that this improves accuracy and considerable evidence that it decreases accuracy (Goldberg, 1959; Quinsey & Maguire, 1986; Dawes, Faust, and Meehl, 1989; Webster, Harris, Rice, Cormier, & Quinsey, 1994; Quinsey, Harris, Rice, & Cormier, 1998). For another example, consider the best predictor of non-sexual violence, the Violence Risk Appraisal Guide (VRAG). The VRAG was also intended to be an adjusted actuarial tool, when Webster, et al. (1994) created it. In less than five years, they

> reasonably sure that a factor should not be considered when the average correlation is near 0 . . . Included among the unrelated factors are measures of general psychological adjustment . . . (average *r* = -.01) . . . Additionally, those offenders who denied their offense were at no higher risk for recidivism than other sexual offenders (average *r* =.02) . . . Many of the factors that clinicians intuitively believe to be related to sexual offense recidivism, such as denial of the offense and verbal statements of treatment motivation, have not been found to predict sexual offense recidivism[8]

A negative clinical presentation (e.g., low remorse, denial, low victim empathy) was unrelated to sexual recidivism.[9]

In summary, the clinical method for predicting human behavior is no longer taken seriously among among prediction scientists.

recanted and eliminated all clinical adjustment to their measure (Quinsey, Harris, Rice, & Cormier, 1998). This will be discussed further. At this point, it is important to simply note that the type of clinical judgment typically used by State examiners has a correlation estimated to be .07 with recidivism (Hanson & Bussiere, 1998). This is so close to zero it can often be surpassed by tossing a coin, a fact that is consistent with the fact that lay people are usually as good as or better predictors than experts using clinical judgment.

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

[9] Hanson and Bussiere, 1998, p. 357.

Base Rates (Exhibit 3)

No facts about an individual (person or event) can have any meaning unless they are compared to the class of similar facts about similar individuals and events. That is, it is very rare that the predictive meaning of an event can be interpreted without knowledge of base rates.[10]

So, let's turn to that principle: *in no case can any event be interpreted or predicted without reference to the class of other similar events (base rates).* Here are three thought experiments that illustrate the relationship between fact, meaningful prediction, and comparison.

To set the groundwork for these experiments, consider that meaning itself is impossible without prior experience with similar objects or events. As in the classic example of the impossibility of explaining what "red" means to someone who has been blind from birth, it is impossible to even understand what words refer to without reference to prior experience with other objects, events, or feelings similar to that to which the speaker is referring. Likewise, an expert can only find meaning in a symptom, sign, pattern or behavior, etc. in an offender by reference to prior experience with similar phenomena in other offenders. In the first experiment, we will see that prediction is impossible without prior experience with similar phenomena. In the second experiment, we will see that knowledge of the comparison group on which the "expert" bases his opinion becomes crucial in determining how much weight to give his opinion. In the third experiment, we will see that, if the fact finder cannot be told about the methodology used to apply information about similar offenders in order to make a prediction about a specific offender, then they cannot know if the expert is presenting the type of clinical impression that has been shown to be wrong more often than it is right, or presenting true expertise, i.e., basing his opinion on careful, controlled systematic study (science).

Let's start with a simple example illustrating the importance of prior experience with other similar objects in order to predict what will happen with a new object. If we were to ask an adult of ordinary intelligence to predict the number of heads and tails we would obtain if we tossed a coin 100 times, most people would accurately predict something close to 50/50. However, imagine an Amazonian aborigine from the deepest parts of the rainforest, someone who has never seen a coin or any two-sided object so regularly shaped. He might very well be expected to study it and, based on his impressions of the images on the coin (an image of a "god" on one side and a toad on the other), make a prediction that would be far off the mark. He would have no way of knowing. However, with a little experience tossing that and other coins, he can be expected to improve the accuracy of his prediction and soon will be able to generalize to other coins he has never seen. The point is that prior experience with *other* coins or

[10] Can we ever predict accurately without base rates? Yes, of course. For example, if we are trying to determine if Johnny Baseball will ever pitch a no-hitter again, without knowledge of base rates, we can be fairly certain that he will not if we know that he lost his pitching arm in an auto accident. But these are unusual circumstances for which base rates will never be available (the phenomena are too rare) and *for which we need no expert opinion*. In situations in which expert opinion is needed to make predictions, relying on clinical judgment without knowledge of base rates will almost always produce erroneous conclusions.

other regular two-sided objects is necessary to make reasonably accurate predictions about a particular coin we have never seen.

In the second thought experiment, let's complicate the matter by involving an event determined by human behavior and interpretation of human constructs. If we took our Amazonian and dropped him into downtown Boston and showed him a parked car—coming from a dense jungle where they don't even utilize the wheel—he wouldn't have any way to recognize it as a mechanical means of transportation or what we mean by "a car." He would have no ability to make any predictions about what that car would be likely to be used for in the future, e.g., that if one observed it steadily for a month, it is extremely likely (not certain, but extremely likely) that it would be entered and would accelerate carrying at least one person off and out of sight. Indeed, if he only saw people sitting and leaning on the car as they talked and later saw another parked car, he would reasonably predict that the car would *remain where it is* and that *people would gather around and on it* for conversation. While that prediction may turn out to be true, it clearly is (1) based on his prior experience of a *different* car, (2) it is a prediction that is likely to have limited validity, and (3) the validity of the prediction can only be evaluated by knowing what experience with similar objects (cars) the Amazonian has had, i.e., what the comparison group was and how extensive and systematic the observations of that group were.

Indeed, we can easily imagine that if, upon first seeing a car, another Amazonian was told by the first that cars are decorative devices used by people for leaning on and talking, he would be likely to question the first one as to how he knows that. If told that the first Amazonian saw that occur on one occasion after observing a different car for a couple of minutes, the second might very well glance at the car and say, "Hmmn. Maybe. But it looks like it might be used for something else." If however, the first Amazonian said that he had spent years studying many, many cars and had never seen anything else happen with them, then the second might conclude that the first was an "expert" on predicting what would happen if you observe parked cars for an extended period and accept the first one's opinion as likely to be accurate. But the second Amazonian would be expected to doubt the first's opinion about an individual (person, event, or object) without knowing the quality and extent of prior experience with *similar*—note, that *neither* has had *any* prior experience with the *particular* car in question—objects that formed the basis for the opinion.

It has been incorrectly argued that, unlike our Amazonians, psychologists do have prior experience with the defendant in the form of the offender's record and the history. However, this misrepresents the notion of what is being interpreted in order to make a prediction. Everybody knows that the defendant in these cases has a history of committing sexual offenses. We don't need experts to tell us that a person who committed a sexual offense (or any other crime) is, on average, more likely to commit another similar offense than a person who never committed such a crime. Everybody knows that.

We need experts to take the specific facts, such as the nature of the crimes, the extent of the criminal history, the age at which the crimes were committed, the mental condition of the defendant at the time of the offenses, etc., and tell us what they mean (interpret them) in terms of the likely future behavior of this individual. Let's take a closer look at just one of these phenomena, the defendant's offense history.[11]

We do not need to know if a future sex offense is more likely for someone with an offense history than for someone with no offense history. We already know this without the expert; this is the only valid prediction that can be made without comparing the offender to other offenders; this is not something that the law requires the fact finder to decide. We need to know whether, given the particular defendant (among the larger group of offenders) with his particular history, another offense is likely enough to conclude on the basis of clear and convincing evidence that he is in the small group of truly sexually dangerous sex offenders.

The first conclusion ("more likely than a non-offender to commit a future offense") can be based on the offender's history without comparison to other offense histories from other sex offenders. In this sense only, i.e., only in comparing the offender to non-offenders in order to reach a relative to non-offenders risk conclusion, can a sex offense history be said to give us risk information without base rates based on other offenders. But this is meaningless for SD cases since the fact that there is greater relative risk compared with non-offenders is already known; it is assumed and provides the only legitimate basis for evaluating sex offenders for commitment in the first place. It is only in this sense that the history alone gives us something that could be called "prior experience of the offender" of the type necessary for making predictions, i.e., without base rates, the offender's history of sex offending enables us to conclude that they are, indeed, more likely to commit a future offense than someone who has never offended.

But for SD evaluations, this is a meaningless use of an expert because we start with that premise; that premise is the basis for evaluating sex offenders to determine if they are SD. We are trying to assess what the actual risk is; this is the only concern at hand, i.e., whether the particular offender is so much more likely than the typical sex offender to commit a new offense that he meets the SD standard and qualifies for commitment beyond his criminal sentence. We need an expert to help us know the meaning (in terms of making a specific prediction, i.e., is another offense significantly more likely than the average sex offender, and if so, how much more likely) of the defendant's particular history. Given the fact that sex offenders are more likely to commit sex offenses in the future than non-offenders, i.e., given the only risk conclusion we can come to without comparison to other offenders (without base rates)—which is what justifies the SD law and the hearing to determine sexual dangerousness in the first place, but doesn't tell us whether a future offense is likely enough to justify a finding of SD in the specific case at hand—is this particular offender like those other offenders we have known who go on to commit new offenses when released?

[11] The exact same analysis could be applied to any other feature of the defendant, e.g., whether they are continuing to act out, whether they are cooperating in therapy, etc. The following also applies to any combination of features used to make a prediction. By focusing only on offense histories, I am merely trying to simplify a discussion of the logic of prediction based on data, to show that comparisons to prior knowledge of other offenders, i.e., base rates, on this, as well as all other variables or combinations of variables, is essential to making accurate predictions.

Interpretations of what offense histories predict—like our Amazonian expert's conclusions about what could be expected of "car phenomena" in the future—can only be constructed through study of similar "cars," i.e., similar offense histories in other similar individuals and how well they predicted future offenses with those similar individuals. While we can know the defendant's history, we cannot have prior experience of the particular history, itself, i.e., we do not know from this offender's history whether such a history indicates high or low risk. Without prior experience with phenomena similar to the one before us (other offense histories), we cannot interpret the meaning of the particular history in question. Since we must interpret the particular individual's offense history, any meaning we attach to such a history (again, as in the example of the parked car) must come from prior experience of other similar histories in other individuals. It is such similar histories themselves that an expert must have prior experience with in order to interpret their meaning accurately. Because the history of the individual cannot be used to make any SD relevant prediction whatsoever without comparison to other similar histories, the defendant's history does not give us the type of "prior experience" necessary for making a prediction and cannot be used in isolation, i.e., cannot be used without comparison to other sex offender's histories. And again, that comparison can be either scientific and improve the validity of the interpretation, or idiosyncratic (i.e., using the clinical method) and lead to a very high degree of erroneous prediction.

If one does not describe the comparison group—i.e., what the cumulative experience of observation of similar events/objects one is basing the opinion on—there is absolutely no way for the fact finder to judge the validity of the opinion. If the cumulative experience used for comparison is clinical observation/impression, we now have overwhelming evidence that that is an extremely poor basis for predictive comparisons—e.g., it is, in effect, equivalent to predicting what will happen with the parked car after watching a different car for five minutes;[12] we should expect it to lead us to a wrong conclusion far more often than to a correct one. If the basis for predictive comparisons is scientific study, then and only then can the expert be expected to be of aid in helping the fact finder come to the truth about the likelihood of future offenses. Thus, like the second Amazonian who is seeing a car for the first time, fact finders are often trying to determine the meaning of a sex offender's history and current condition for the first time and will naturally tend to rely on anyone presented as an "expert." When the expert is using the equivalent (in terms of reliability) of five minutes of observation of a parked car—i.e., when the expert is using clinical theory and clinical impressions and findings—their opinion is worse than worthless in that it is more likely to lead to false conclusions.

Without the knowledge of (1) what the basis is for predictive comparisons, i.e., what the source of the base rates is if there are any such base rates, (2) what method was

[12] The vast majority of expert witnesses have never worked with sex offenders long enough to see which ones actually reoffend and which ones do not. Indeed, the majority of experts have never had significant treatment experience with sex offenders until after they become known as an "expert" in this area (due to their being an expert witness in these cases) and then some of these "experts" do start to get referrals of such offenders. That is, even when employing the invalid clinical method, most state "experts" don't even have the clinical experience with sex offenders that would enable them to have any basis for comparison.

used to establish the base rates (quantified scientific observation and study vs. subjective clinical observation and theory) and (3) that conclusions based on theoretical clinical notions and clinical judgment have been conclusively shown to be the wrong methodology for these purposes, there is absolutely no way the fact finder can intelligently weigh the expert's opinions.

A third and last example, this one (like predicting sex offenses) involving the prediction of behavior that is relatively rare in the general population and more common among a select few. Imagine a new major league ball player who, in his first at bat, hits a home run. We are asked to then predict how likely it is that he will hit a home run in his second at bat. Also imagine that we have just two pieces of information: He is one-for-one in hitting home runs in the majors and his high school record shows that he often hit home runs. As past behavior is the single best predictor of future behavior and—in addition to fairly frequent home runs in high school—since he hit a home run the only time he engaged in "batting behavior" in the majors, we should predict a home run in his second at bat. Of course, if we regularly used such logic for betting, we would lose a lot of money. Obviously, the high school performance—where the pitching and fielding is rarely, if ever, of major league quality—of the world's best batters (i.e., those who get to play in the majors) is going to be a very poor predictor of their major league performance where they will be batting against the world's greatest pitchers. So, even though we have the evidence of a first home run in the majors and many earlier home runs, we would predict that a home run is unlikely in his second at bat.

But how do we know what weight to place on his first at bat in the majors and how to utilize the information from his high school history? By looking only at his history, we have no prior experience with what such a history means, no way of interpreting it. Instead, if we look at other major league players and compare their batting averages and home run history (in the majors vs. high school), we immediately find that they go down, dramatically, when they enter the majors.

Furthermore, among the thousands of players who have entered the majors, there ought to be dozens who hit a home run in their first at bat. By looking at their second at bats—and by specifically seeing if there is any predictive relationship between the second at bats of that small group that hit home runs in their first at bat along with their high school batting record—we can refine our prediction by looking at truly relevant information. Thus, by looking at what other new players have done (specifically, other new players who fall into the same category) we can utilize our prior experience with similar events (those other players) to predict this particular player's next at bat performance. In this example, to make our prediction, we can use the past rate of home runs by all new major leaguers in their second at bat. We can then adjust that estimate if we found a significant difference between the average new player and those in the group who hit a home run in their first at bat. And we can adjust it further if we found that knowledge of high school performance is statistically related to actual second at bat performance in this select group.

Only by reference to prior experience (or knowledge of base-rates) with other similar phenomena can any event or phenomenon be interpreted and predicted. There is simply no other valid basis (other than prior experience with similar phenomena) to make meaningful predictions. To some degree, all experts in these cases are making predictions in this way. However, that degree varies and the methodology used to apply prior knowledge varies. As noted, one can apply prior knowledge in two ways: scientifically (through controlled study) or through subjective clinical judgment and impression. The former is an accurate, useful method, the latter has been proven to be worse than useless.

Prentky Study (Exhibit 4)

While the under reporting of sex offenses is typically taken to indicate that any empirically obtained rates seriously underestimate the actual offense rate, this is unlikely to be so in the case of the Prentky study. The study tracked 251 designated "sexually dangerous" persons who were released from the Massachusetts Treatment Center for Sexually Dangerous Persons between 1964 and 1988 over a 25 year period. Its results are likely to be overestimates of recidivism for eight reasons.

First, Prentky, et al. were not counting offenses, they were counting recidivists. If a man was caught once, he was counted as a failure (a recidivist) and, even if he had committed other offenses that went undiscovered, he couldn't have been counted as more than one failure. So even if there is a very high rate of unreported offenses, if the men under question are repetitive recidivists, then even if most of their offenses went unreported, over a 25 year period they would be likely to be caught once. Thus, even if every one of a repetitive recidivist's unreported offenses had been known, this knowledge would *not* raise this man's contribution to the estimated rate of overall recidivism in this study.

Second, it is reasonable to assume that some sex offenders are very good at escaping detection and that they account for a significant part of the unreported offenses. It is also fairly safe to assume that these stealth offenders were *not*, by and large, in the offender group in this study; remember Prentky, et al. were dealing with only repetitive and compulsive offenders who had been committed as sexually dangerous persons, i.e., they were all men who had already been caught more than once (and in most cases several times or more).

Third, even if some offenses were not reported, if an offender was released despite still being SD—i.e., if he was still a compulsive and/or repetitive sex offender—then given enough time, he was likely to either be caught or, because of the last reason (below), to be counted as caught. While there may be some men who are compulsive, repetitive offenders who would only commit one additional offense over a 25 year period, these are likely to comprise a small minority of truly sexually dangerous men who are falsely considered not dangerous and released. We can expect that most of the false negatives (i.e., truly sexually dangerous men mistakenly released into the community) would commit more than 1 or 2 offenses if they remained at large for 25 years. When this factor is combined with the previous factor (poor stealth among our sample) and the fifth factor (heightened surveillance for our sample), it becomes less likely that unreported offenses will lead to an underestimate of recidivism in this sample.

Fourth, the Prentky, et al. study used a very long follow up period of up to 25 years for some men. The recidivism rate for the others were estimates based on the group of men who were released early enough that they could have been at large for the entire 25 years.

Fifth, in my experience at the Treatment Center (1977-1986), when my patients were out in the community (and a federally court ordered community access program made this a common experience) and a sex offense occurred in a geographical area that they could physically have been in, they were almost always investigated. Thus, unlike other offenders who could get away with numerous offenses without being caught, current and former TC patients were high profile suspects for the police long before Megan’s laws. Thus, in this study, repetitive, compulsive offenders who were, on average, not very good at escaping detection were under heightened surveillance. Clearly, under reporting of offenses that would enable recidivists to escape detection is less likely to have been as big a problem for this group than for sex offenders in general.

Sixth, the estimates of recidivism for the full 25 years for the majority of the men who were not released early enough to have been at large for 25 years, were based on a few men who were released early in the Treatment Center’s history. Only this small group of men who were released in the Treatment Center’s first five years could have been out for 20 - 25 years and thus the long term recidivism rate had to be based on this group of releasees who left shortly after the Treatment Center first opened. Since this group could only have been in the Treatment Center (which had only existed for five years at that point) for a short stay of one to five years, they were also likely to be significantly younger than the men who were considered for release from the Treatment Center later in the study, most of whom had been in for more than 10 years. Thus, long term recidivism estimates were based on offenders who left after relatively short stays, often after having received little or poorly formulated treatment (if any). Thus, the 25 year risk was estimated based on the few men who had the opportunity to be out for the full 25 years and who were released at a younger than average age (a known risk factor) after having received much less risk-reducing treatment.

Seventh, for the first one-and-one-half decades of the existence of the Treatment Center, there were two ways for a man to leave. The courts could adjudicate him no longer sexually dangerous, or he could be released on parole while still carrying the SDP label. This latter mechanism accounted for a significant numbers of releases (R. Prentky, personal communication, May 2001). When the men had two shots at release—with both being a genuine avenue of release—it was easier to get out of the Treatment Center. Since the mid 70's, the parole option has been shut down, making release harder. If the men used to estimate long term recidivism had an easier time getting out, they would have been younger, on average, when they entered the community and higher risk men who were still deemed SD would have been released. This established risk factor (age) and the presumed risk factor of still being SD suggests that their recidivism rate should be higher than for the older crop of men who came up for release later in the study. Thus, we can assume that younger men with shorter periods of incarceration and less treatment—thus more dangerous men—were released in the early part of Prentky’s study, i.e., on average, the group of men that were used to estimate long term recidivism for the men released later in the study (as only the early releasees could actually have stayed out 15 to 25 years) were likely to be more dangerous than the average releasee in the study.

And eighth, the high estimates of recidivism Prentky, et al. reported were based on the most inclusive measure of failure and surely contained some cases where no recidivism existed. For example, a former SD person is charged or arrested because an offense similar to his historical pattern was reported. Upon further investigation, the victim says he was not the perpetrator or other evidence exonerates him. Though he was not convicted—and if the charges were dropped, he may not even have been arrested—even if someone else was convicted of the offense, this counted as recidivism in Prentky's study. (Prentky, et al. counted such incidents precisely to attempt to offset the under reporting of offenses.) Keeping in mind that former TC patients were under heightened surveillance and were considered prime suspects for any sex offenses that were committed in the area they were known to live or work in, there were more than a few charges and/or arrests for offenses that they did not commit.

Thus there are eight reasons to conclude that the figures from the Prentky, et al. study do not underestimate recidivism and may very well be an overestimate:

1. that, for each offender, being caught once eliminates the impact of any unreported offenses;
2. that the offenders under study were not very good at remaining undetected and
3. if they were truly compulsive, repetitive offenders (as they had been adjudicated to be) they were likely to commit more than 1 or 2 offenses over a 25 year period and thus likely (given factors 2 and 5) to come to the attention of the authorities (given enough time);
4. that they were given enough time (adjusted to simulate 25 years in the community for all of the men);
5. that this group of compulsive repetitive offenders who were below average for their ability to remain undetected were under heightened surveillance over that long time period;
6. that the 25 year estimates of recidivism that were used for the majority of men in the study who were not out for the full 25 years were based on the few men who had the opportunity to be out for the full 25 years and who were released at a younger than average age (a known risk factor) after having received much less treatment (an assumed risk factor);
7. that it was easier to obtain release during the first half of the studied period and thus, on average, younger and more dangerous men who received less treatment, punishment, and had aged less were used to estimate recidivism for all of the longer periods of exposure
8. Prentky, et al. used the most inclusive estimate of recidivism, "charge," which included some false recidivists in this large study over such a long period.

To summarize: It is unlikely that many true SDP's, i.e., compulsive or repetitive offenders, could survive a twenty-five-year period in the community under heightened surveillance without being counted as a failure if we used a very broad, inclusive definition of recidivism. Thus the Prentky 25 year estimates of sex offender recidivism of men who were the highest risk sex offenders, all of whom had been adjudicated SD, provides an upper limit of recidivism, despite the often noted problem of unreported sex offending.

The purpose of the Prentky, et al. study was not to provide a base-rate that could be used to estimate the risk of recidivism for men leaving the Treatment Center. Rather, their goal was to utilize a method that would provide an upper estimate of recidivism (counting charges and arrests as survival failures, i.e., recidivists) and then to compare it with the estimate obtained on the same sample when using the different methods (using sex offense conviction or incarceration for another sex offense as indicating recidivism) that had been used to estimate recidivism in other research studies reported in the literature. Their goal was to understand the wide, discrepant range of results produced as estimates of sex offender recidivism. Rather than being a study of the recidivism rate (an attempt to establish a base rate), it was a study of one aspect of the problem of accurately determining recidivism rates. One long range goal of such a study was to eventually enable us to use a more thorough understanding of the problems associated with measuring such phenomena.

Given these qualifiers, Prentky's 25 year recidivism estimates of 52% for extra familial child molesters (actual number that recidivated during the study period was 32%), 39% for rapists (actual 26%), and for all the offenders combined, 45% (actual 29%) should both be adjusted downward. Since lengthy involvement with therapy, documented behavioral change in the institution, and the passage of a one-and-one-half decades since the last sex offense is the norm for these cases, the most reasonable adjustment to an overall base-rate between 29% (actual) and 45% (upper estimate) is downward, making it highly likely that the majority of the men at SD commitment review hearings are not sexually dangerous.[13]

[13] Recently, the men I have been evaluating have been much older. Their recidivism rates should not be those of an average Treatment Center patient of the 1960's who was released at the age of 28 and followed for 25 years until they were 53 (the numbers are illustrative only as the actual ages are unknown, but Prentky, personal communication, May 2001, acknowledges that this argument is reasonable). The men evaluated during the decade from 1995 to 2005—during a period in which no new commitments were being made because the new commitment section of the Massachusetts SDP law had been eliminated—averaged 40 to 45 years (with a significant percentage in their fifties). These men would be entering or in *old age* before 25 years pass. We know that aging is a significant factor. If the average age of men who were followed for 20 to 25 years in the Prentky study was 30 at release (again, the actual ages are unknown), then they had a decade of exposure in the community in their thirties, a decade in their forties, and a few years in their fifties. The majority of men evaluated in the decade from 95 to 05 are probably a decade older and thus their first decade would, on average, be in their forties, then their fifties, followed by their sixties. This means that SD men whose commitments were reevaluated during most of the nineties up through 2005—which is the period during which the statistics were gathered for the bias analysis presented in Attachment #9—were almost certainly less dangerous (less likely to reoffend) than the men in the Prentky, et al. sample. Thus, it seems almost certain that the men evaluated during the period of the bias analysis had a true base rate of recidivism of less than 45% while the State's experts opined sexually dangerous at a rate more than 200% greater (greater than 95% of the time). As we will see, this rate of over-predicting recidivism indicates gross bias way beyond *any* possibility of doubt.

Correlation Coefficients, Validity, and Variance (Exhibit 5)

In order to understand the problems inherent with expert testimony predictions and the Jimmy Ryce Civil Commitment Program, we need to understand how we can measure the degree of relationship between a predictor and the predicted variable, the latter being whether or not a sex offender will reoffend. For example, we need to know what a correlation coefficient (denoted by *r*) indicates, for this is crucial to understanding how important the scientific findings are for our enterprise.

To illustrate the meaning of *r*, I will present a hypothetical experiment that could actually be carried out. But first, consider that the absolute value of *r* can range from 0.0 to 1.0, with 1.0 being a "perfect" relationship, i.e., if the correlation between two variables yields $|r| = 1.0$, then knowing how an individual scores on one variable tells you perfectly how they will score on the other. For example, the correlation between the blood test for HIV and the presence of the virus and the disease, AIDS, might yield an *r* of .97 (the test isn't perfect or it would yield an *r* of 1.0).

If we looked at only people between the ages of 35 and 90 and tried to measure the correlation between age and AIDS infection, we might arrive at an *r* of -0.3. The negative correlation means that the relationship is inverse; as one variable (age) goes up the likelihood of the other being positive decreases.[14] Such a negative correlation could be explained by the fact that as we look at older and older people, they are less and less likely to be carriers of the AIDS virus (both because older people are less likely to be as promiscuous and as sexually active as younger people *and* because people who have contracted AIDS are less likely to make it to old age).

On the other hand, trying to determine if someone is infected with AIDS by using their shoe size would probably yield an *r* that is close to zero, i.e., there is virtually no relationship. Note that the *r* actually obtained from such an attempt to measure the correlation between shoe size and AIDS infection would not be exactly zero as, in any given sample, there will be some chance relationship between the two variables. So, we are likely to come up with an *r* of .02 or -.03, i.e., something very close to zero.

But that still doesn't tell us what, for example, an *r* of .1 or .46 actually means. These numbers are very important because the Hanson and Bussiere meta-analysis showed that clinical assessments had a correlation of $r = .1$ (10 studies, 1,453 sex offenders) while other research has shown that statistical risk prediction scales correlated more closely with recidivism, $r = .32$ to .35. Both clinical assessments and statistical risk prediction scales were "significantly" related to recidivism. In research, *significant* means that the relationship found was unlikely to be a chance finding, i.e., it is likely to

[14] While *r* can range from -1 to 1, it is important to keep in mind that it is the distance from zero (or how close the *r* is to 1 or -1 or the absolute value of *r*) that indicates the strength of the relationship, not whether the number is positive or negative. For example, the *r* of .97 as the hypothetical correlation between a positive HIV test and the presence of an actual HIV infection could also be presented as an *r* of -.97 between a positive HIV test and the patient being *HIV-free*. Or an *r* of -.97 would be obtained between a *negative* HIV test and the presence of an actual HIV infection. All of these are different ways of expressing the same degree of relationship. The strength of the relationship is indicated by the absolute value of *r*, not by whether it is greater or less than zero.

be a *real* relationship. This means that clinical assessments that yield a prediction of dangerousness will, in fact, on average, be found more often with offenders who recidivate than with those who do not.

But such "significance" doesn't tell us *how much more* often a clinical assessment of dangerousness will be associated with actual danger. We know that the statistical risk prediction scales will be more accurate (because a correlation of .33 is closer to a perfect prediction correlation of 1 than .1 is). But we still have no way of knowing *how much* more accurate, i.e., how much weight to give to either prediction. From what I have said so far, we have no way of determining how *valid* a prediction is if the correlation is .1 or .33. Validity means how much we can count on knowledge of one variable (clinical or statistical assessment of risk) to tell us what in fact will happen when we look at the other, in this case actual recidivism. There is a way to translate correlation coefficients into measures of validity. To illustrate how we can do this, let's turn to a hypothetical experiment.

An Illustration of the Meaning of Correlation and Prediction

Imagine that a high school of 1000 students has been called out to the football field by their high school principal who had an urge to teach the students about correlations. As each student enters the field, the principal tells each one that they are either a 1 or a 2 until exactly half (500) of the students have been assigned 1's and half have been assigned 2's. The principal then tells them that if everyone left and they reentered the field, he would give them the same number and that he didn't use a random method (e.g., whim or tossing a coin) for making the assignments; he used a reliable rule for making each assignment decision. He then challenges anyone to come up with an explanation for how he made his decisions: What factor did he use to make the assignment decisions? In the confusion that follows as people mill around on the football field trying to find a pattern, there does seem to be a lot of "noise" or randomness regarding who's a 1 and who's a 2. There are tall and short, male and female, black and white 1's. There are good students and poor ones, rich and poor, popular and unpopular who were assigned 2's. At first glance, there doesn't seem to be any rhyme or reason, no pattern to the decisions.

So, one student raises his hand and guesses that the principal made the assignments alphabetically by last name. The principal tells them to check it out to see if the first letter of their last names played any role in the decisions. A math wiz, who had just learned about correlation coefficients, takes out an alphabetical roster of names and assigns a 1 to the first 500 and a 2 to those whose names begin with letters from the second half of the alphabet. By calculating the correlation coefficient he could determine if, as an alphabetical scoring rose from one to two, there was any increase in the tendency for a student to be assigned a 1 or a 2.

If he obtained a correlation of 1, then every student whose last name began with A-K (the 500th student's last name was Kzirsky) would be a 1 and the L-Z's would all be 2's. (If the correlation was -1, then the assignments would be the reverse but every assignment could be known simply by knowing a student's last name.) If the correlation was close to zero, then the alphabetic principle could be discarded. If the score was clearly somewhere between 0 and plus or minus 1, then the student's names may have played *some* role in making the decision (along with other factors) or may be determined by a factor that is somewhat related to their names, e.g., the assignments could have been made by the order of their files in the school filing system that is maintained by the principal's very dyslexic secretary and would therefore be somewhere between .1 and .9 (or -.1 and -.9). The correlation actually calculated between last name and number assignment turns out to be *r* = .01, which is so close to zero that alphabetic assignments are ruled out.

One black student runs into a small group of three of his black friends and notices that three out of four of them was assigned a 2. And he also happens to know that exactly half of the students are of African-American descent. However, he also realizes that race could not be *the* determining factor (since one of this small groups of black students had been classified a 1). So he hesitates to guess that race is the factor. After a while, when no one comes up with a better answer, he does guess that race was a determining factor. The principal tells them to check it out. So, the math wiz assigns a 1 to every non-black and a 2 to every African-American. They calculate the correlation coefficient and it turns out to be .07 and *significant* (it is *unlikely* to be a chance relationship). But when they look around and count, they find that there are 233 black 1's and 267 black 2's; *a similar number of blacks are in both groups*. When they then include all minorities, "people-of-color," in the calculation, the correlation goes *down*. So they realize that minority race is *less* likely to be a factor than African ancestry (if that is, indeed, a factor at all and not just a weak, chance relationship that was accidentally enough to be deemed "significant" by a statistical analysis). If African ancestry is indeed a factor, it certainly isn't a major factor. The students think about what race, itself, could be correlated with. Someone recalls that African-Americans are, on average, taller than those Americans without any African ancestry.

They then measure everyone's height and give a 1 to the 500 shortest and a 2 to the 500 tallest and calculate that correlation: Does assignment to one of the two height categories correlate with the assignment of 1's or 2's to the students? The correlation coefficient turns out to be .34 and significant. Now they appear to be getting somewhere. Yet, though the 2's are clearly taller on average, there are a few short 2's and a few tall 1's. Finally, someone guesses that weight is the factor. They go and get the school nurse's scale and weigh everyone (the 500 lightest being assigned a 1 and the 500 heaviest assigned a 2) and the correlation turns out to be .98. That seems to be the key. If you are lighter than average, you are almost certain to be assigned a 1; if heavier, a 2. The principal admits that that is the factor he used. It was imperfect because the weights came off their medical records and some students near the middle have switched categories as they have gained or lost weight. So, these are the correlations they came up with:

Factor	Correlation	Significance[15]
First letter of last name	.01	not sign.
Minority race	.03	not sign.
African ancestry	.07	p<.05
Height	.34	p<.001
Weight	.98	p<.000001

When they tried to account for the seeming randomness—the way the principal's decisions seemed to *vary* willy-nilly—none of the *variation* ("varying" between being assigned a 1 or a 2) could be accounted for by the first letter of last name. Minority race accounted for not much more. African ancestry was significantly (or truly, validly) related, but the relationship was very weak, i.e., it was a very poor predictor. Height was both significant and more valid (it was a better predictor), but it left a lot of the variation unexplained. Weight, however, predicted or explained—could be used to *predict* what number would be assigned to an individual student or could *explain* (or *account for*) the pattern of number assignments for almost all students—most of the variation in whether a student was assigned a 1 or a 2. In fact, the percent of the *variance* (a measure of the variation) explained (or predicted or accounted for) is a more meaningful measure of the strength of the relationship between a factor (name, race, height, or weight) and the predicted variable (assignment of a 1 or a 2). So, we have to explain a little bit about what the *variance* in an attribute (or variable) is and how we measure it.

The Difference Between "Significance" and "Validity": Using the "Percent of the Variance" to Measure the Validity of Significant Predictor.

To understand the meaning of "percent of the variance," consider the following illustration. Imagine that we had a group of 9 schoolboys line up in order of weight (to the nearest kilogram) as measured on the school nurse's scale and we assigned them the categories 1 through 9 (first through ninth). The numbers would then provide information about the boys' relative weights, lightest (1) to heaviest (9):

	Lightest				mean				Heaviest
Category:	1	2	3	4	5	6	7	8	9
No. of Boys:	1	1	1	1	1	1	1	1	1

[15] Significance is determined by the probability that an observed pattern has *no* meaning, i.e., that it indicates nothing and is just a chance event. If the probability (p) of getting the pattern of we are seeing by chance is less than .05 (p < .05), we say we have a *significant* finding that represents something about the real world. If it is greater than .05, we say it is not significant and could very well just be a chance event. If p is much less than .05 (e.g., p < .001), then we have a higher degree of certainty that the pattern we are seeing is not a chance event.

If you were sitting in another room and were told that there were nine boys lined up in the other room in positions (or categories) 1 through 9, and you were asked to guess what category a randomly chosen boy is in, the best guess would be the mean, 5. If you always guess the mean, you will, on average, be closer to the correct number than any other guess.

The mean is the number of the category that is least distant from all the other categories in the group. That means that if you add up the distances between a single boy, such as the boy in category 1, and each of the others—i.e., how many categories you would have to move through or into to go from the boy in the first category to the boy in category 2 (one) *plus* from boy 1 to the boy in category 3 (two) *plus* from boy 1 to the boy in category 4 (three), etc.—the total is 36. The distances between the boy at the mean (5) and each of the other boys, on the other hand, add up to 20. The mean is the number closest, on average, to all the other numbers. If we have a single number, the mean, that is closest to all the others—a central point, so to speak—than we have a point around which we can measure the *variability* of the group of nine students.

The variability of this group could be measured by the *average distance from the mean*, which in this case is a little more than *two* (20 ÷ 9). If we took another sample of boys and they all happened to weigh the same (were within one kilogram of the others), there would be only one category, which would also be the mean:

	Lightest/mean/Heaviest
Category:	1
No. of Boys:	9

In this case, the *average distance from the mean* to each of the others would be *zero* (0 ÷ 9).

If some, but not all, of the boys belonged to the same weight categories, then there would not be nine positions (several would share the same ordinal position or category), e.g., three boys could belong to the middle category and two to the highest and lowest with one between the middle and each end:

	Lightest		mean		Heaviest
Category:	1	2	3	4	5
No. of Boys:	1	2	3	2	1

In this case, there would be 5 categories: 1 boy in the first, 2 in the second, 3 in the third, 2 in the fourth, and 1 in the fifth; the *average distance from the mean* to each of the boys would be about *one* (8 ÷ 9). Thus, the *variability* in this group (as measured by the average distance from the mean) would be about ½ of the variability in our first group.

Back to our original example with 9 categories. If you have the name of each boy on a card in the other room along with their height and weight, you could use this information (name, height, or weight) to venture a guess as to how the boys are lined up. For example, placing the cards in alphabetical order, you could guess that Zylberschmidt is last (farthest above the mean), Abel is farthest below the mean, and Kriegman is right at the mean. What you would be doing is trying to account for how the boys *vary* in their positions around the mean, does a boy appear to be above it and close? Below it and close? Above it but far? If you use alphabetical order, in a typical sample of boys lined up by weight, you won't account for *any* of the variance, i.e., you would be "predicting" none of the variance in the lineup order and your guesses would not improve on the one best guess, the mean. If you used height on the cards to suggest what the boys' positions were, you might actually account for a significant amount of the variance as height and weight are correlated, or height can be used to "predict" weight more accurately than random guessing or the first letter of the boy's last name. And if you used their weights on the cards to predict the order they are in—again, in what direction and how much they vary from the best single guess (the mean)—you would account for 100% of the variance around the mean and could pinpoint exactly which category each boy is in. Thus, when a factor (e.g., weight) accounts for 100% of the variance in another factor (position in the lineup), we have a perfect predictor (weight, in this case). Whereas a random, useless predictor (e.g., alphabetical order) accounts for 0% of the variance.

In the real world, we don't usually come up with a perfect predictor or a perfectly useless predictor. This is why we must make a distinction between *significance* and *validity*. Significance and validity are terms of art in psychological science. Going back to our example of the 1000 students assigned 1's and 2's, *significance* indicates that it is likely that there is a real relationship between African ancestry and the probability of being assigned a 2 as opposed to a 1. In fact, if there were no real relationship between African ancestry and being assigned to the two groups, we could get such a correlation (.07) about 3 times in 100 such experiments just by chance. Thus, it is unlikely—but not at all impossible—that our correlation is just a chance event. So, *significance* is a measure of the probability that there is a *real* relationship between two variables and not just a chance event.

Validity is a measure of the *strength* of that relationship, which in this example is very weak 267 African-Americans were assigned 2's versus 233 having been assigned 1's. A highly valid predictor would have a high correlation with the predicted variable and would provide a good guide to the predicted variable. A significant predictor with some, but very little validity, e.g., African ancestry in this example, would be a weak, poor guide to estimating the value of the predicted variable with a randomly chosen student. Validity can be either weak or strong and there can still be significance. It is much harder, however, to find the significant relationship between two weakly related variables. For example, the weak but real relationship between race and category in the principal's experiment could not be demonstrated if he only had 4, 10, or 20 students.

Without significance, the likelihood is that we are just seeing random fluctuations between the variables. In our example, there is probably a real (non-chance, non-random) and thus a *significant* relationship between African ancestry and having had a 2 assigned. However, the *validity* is very poor, i.e., African ancestry is a very poor predictor of category assignment.

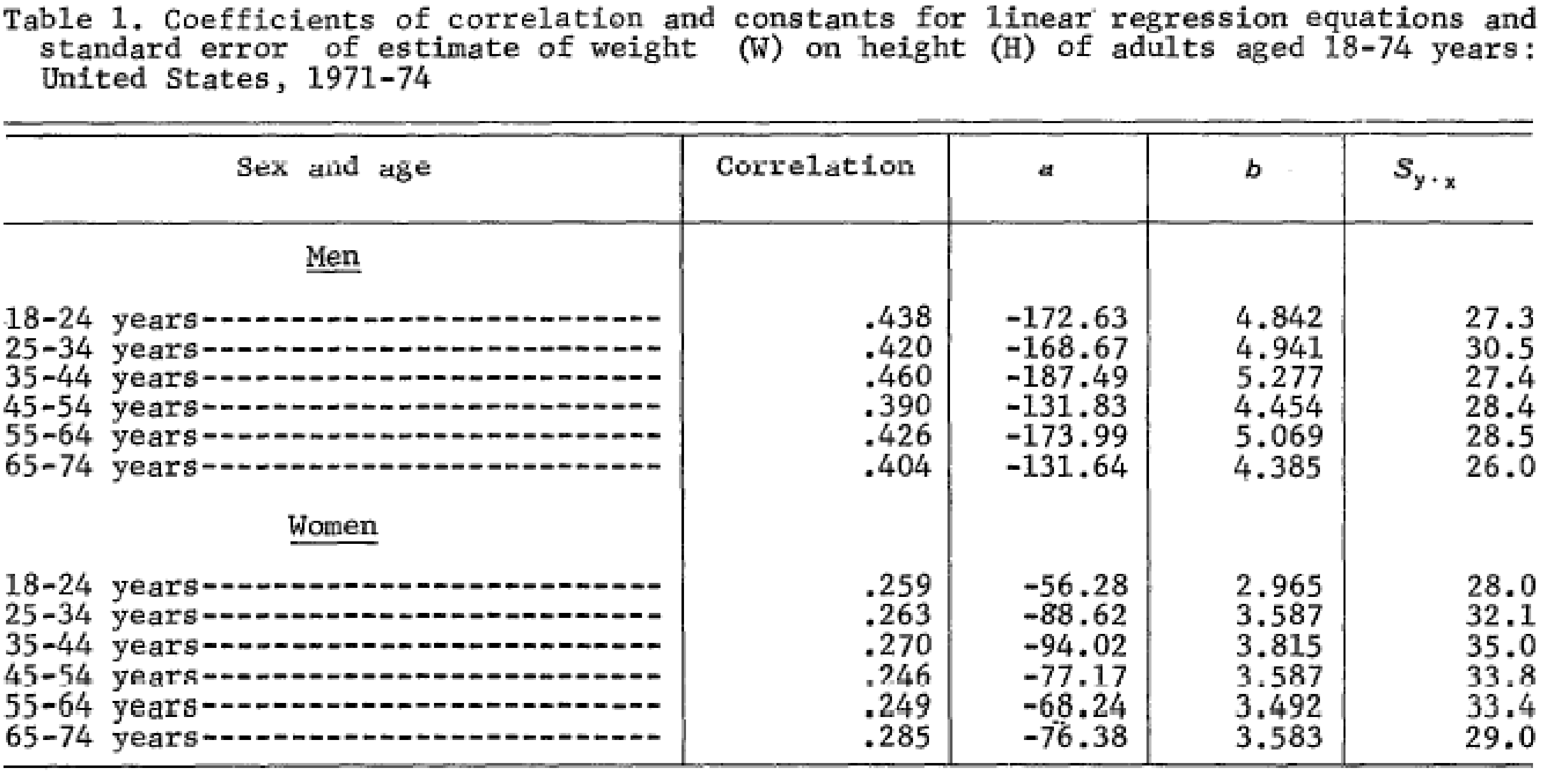


Table 1. Coefficients of correlation and constants for linear regression equations and standard error of estimate of weight (W) on height (H) of adults aged 18-74 years: United States, 1971-74

Sex and age	Correlation	a	b	$S_{y\cdot x}$
Men				
18-24 years--------------------------	.438	-172.63	4.842	27.3
25-34 years--------------------------	.420	-168.67	4.941	30.5
35-44 years--------------------------	.460	-187.49	5.277	27.4
45-54 years--------------------------	.390	-131.83	4.454	28.4
55-64 years--------------------------	.426	-173.99	5.069	28.5
65-74 years--------------------------	.404	-131.64	4.385	26.0
Women				
18-24 years--------------------------	.259	-56.28	2.965	28.0
25-34 years--------------------------	.263	-88.62	3.587	32.1
35-44 years--------------------------	.270	-94.02	3.815	35.0
45-54 years--------------------------	.246	-77.17	3.587	33.8
55-64 years--------------------------	.249	-68.24	3.492	33.4
65-74 years--------------------------	.285	-76.38	3.583	29.0

Average male correlation is .42

Average female correlation is .26

Average overall correlation is .34

Percent of variance in weight predicted by height is 12%

From: Us Department of Health, Education, and Welfare. 1977 Weight by Height and Age of Adults 18-74 Years: United States, 1971-74

In this case, African ancestry could explain a tiny fraction, about ½ of 1%, of the decision making, i.e., it can account for ½ of 1% of the variance. *The percent of the variance in one factor accounted for (predicted, explained by) another factor is one way of measuring the strength of the relationship, i.e., how good a "predictor" a factor is.* Height, in this example, "explained" (or could "account for" or could "predict") 12% of the *variance*. And weight that day on the nurse's scale—when correlated with the actual factor the principal actually used, which was weight as noted in the students' medical records—accounted for 96% of the *variance*. So, we can add a third column to our table that attempts to measure the amount of the seemingly random variation that can be made sense of by one of our factors. Another way to say this is that we can determine what part of the seemingly random variation can be *rendered non-random* or *accounted for* or *explained* or *predicted* by any particular factor.

Factor	Correlation (*r*)	Significance	**Part of the Overall Variance Accounted For**	**Predictive Utility**
First letter of last name	.01	not sign.	0.0 %	none
Minority race	.03	not sign.	0.1 %	none
African ancestry	.07	p<.05	0.5 %	none
Height	.34	p<.001	12.0 %	poor
Weight	.98	p<.000001	96.0 %	superb

Now we can make some sense out of our correlation coefficients. Hanson, as noted above, said that correlations of less than .1 would have little practical utility. We can see why: A correlation of .07 accounts for just ½ of 1% of the variance. Such correlations—even if non-random (not due to chance) and therefore *significant,* in research terminology—are close to useless when making predictions, i.e., they have almost no *validity*. Knowing a student's African ancestry added virtually nothing to our ability to predict which group a randomly chosen student would be assigned to, even though there is a *significant* (non-random or not due to chance) relationship between race and height (and through overall size, there is a relationship between race and weight).

Indeed, there is a formula for determining how much of the variation is explained, accounted for, or predicted by two factors with a correlation coefficient of *r*. The part of the overall variation that is accounted for is r^2. The square of the correlation coefficient (r^2) = the percent of the variation around the mean (as quantified using a mathematical measure of variation called the variance[16]) accounted for.

Clinical assessments can account for (predict) ½ of 1% to 1% of the variance (in this case, 0.5 to 1.0% of the variation in whether or not a person recidivates), while actuarial risk predictions can account for (predict) approximately 10% of the variance. Both are statistically *significant*. But clinical judgment has no utility in making accurate decisions in the real world. Note that the 10% of the variance that the actuarial method can predict is *not* very impressive (not very *valid*), but it is certainly worth knowing and using when appropriate. *And it is 5 to 20 times more accurate (accounts for up to 20 times more of the variance) than clinical prediction.*

Janus and Prentky (2003) in a thorough review of all the empirical data, point out that it makes no sense to use clinical prediction that has little or no utility and ignore the actuarials. Yet, the fact is that lawyers and judges still prefer clinical experts who offer clinical formulations and opinions and are not interested in research or actuarial evidence (Redding, et al., 2001). Indeed, if the courts do not struggle to comprehend what the real

[16] The calculation of variance is closely related to the way we examined the examples of the small groups of boys lined up in weight order. We calculated the *average distance from the mean* for the different groups of boys: approximately 2, 1, or 0 depending on how much the group varied around the mean in our different examples. If we squared the distance from the mean before averaging the distances, we could come up with the *average squared distance from the mean* or the "variance."

life data is telling us—which can only be done by turning to the research and understanding what the empirical evidence is about expert testimony and its relationship to sex offender recidivism—they will be unable to shake this preference and will continue to turn instead to unsound clinical judgment.

Whether or not the higher (yet low) level of accuracy in prediction attained through the use of the actuarial method is adequate for use in civil commitment is something for the court to decide. Note that there are some cases in which other factors may make this weakly valid predictive method a legitimate tool to use as *an aid* in making a commitment decision. For example, there are cases of chronic recidivism that are too unusual to have been included in an actuarial formula and in which repetitive sanctions have had no effect. If the actuarial tool says "high risk" and we have evidence of relatively recent compulsive, repetitive sexual misconduct beyond the level the actuarial tool can measure, a commitment decision may be sound despite the fact that the actuarial tool alone is a weak predictor. Alternatively, there are cases in which the weakly predictive actuarial indicates "high risk" but the offenses committed were unusually destructive (heinous) and thus the anticipated harm of a future sex offense and a weakly valid prediction of high probability of reoffense may make civil commitment justifiable.

Whether or not the weakly predictive actuarials should be allowed in the courtroom is clearly something for the court to determine and may vary from case to case depending on the additional information available. However, we can be certain that the court can make no rational determination of this issue without some solid understanding of the science of prediction and the very serious limitations of our best methods of predicting future sex offenses. When state experts say "the Static-99[17] indicates 'high risk,'" without carefully explaining to the fact finder that the underlying "science" is not very valid, the prejudicial impact is enormous. Here we have an "expert" who is using a "scientific, objective" tool that says this offender is "dangerous." Without knowing the serious limitations of the tool and the science, the fact finder is simply unable to make a rational decision that has any relationship to events in the real world.[18]

[17] The best researched and most widely used actuarial tool for predicting sex offense recidivism.

[18] And here we get into terminological and possibly ethical concerns. Hanson and Thornton, the developers of the Static-99, labeled high scorers "high risk" to differentiate them from those who scored in the middle (medium risk) or low scorers (low risk). But when an expert says the offender falls into Hanson and Thornton's "high risk" category, he *appears* to be saying the man is SD, as that is what the hearing is about.

Hanson and Thornton, however, were not calibrating the "high risk" Static-99 category by what would be considered an unacceptably high risk in a SD proceeding. There may be as little relationship between "high risk" and SD as in the movie "Spinal Tap" where a drugged-out "rocker" claimed that one amplifier was better than the other because one had a dial that went from 1 to 11 while the other one "only goes up to 10." Since there is no unit of measurement on the dials—their only function is to allow the user to set the volume at the same level in the future—there was no way to compare a 9 on one amp to an 11 on the other. This becomes an ethical concern when an expert opines in this manner (1) knowing that there is no clear relationship between Hanson and Thornton's terminological invention ("high risk") based on their personal attempt to suggest a meaning for the distribution of Static-99 scores and what the fact finder must decide ("sexually dangerous") and (2) without informing the fact finder of the serious limits of the science and the low validity of our best method of prediction.

Specific application of statistical concepts to findings of Sexual Dangerousness (Exhibit 6)

If we understand the difference between *significance* and *validity* and we have some method of evaluating the validity of a predictor (e.g., *percent of the variance predicted*), we still need to know what these things indicate about making *sexual dangerousness determinations* in the real world of the courtroom. That is, we need to have some way to understand in a humanly meaningful way (as opposed to just seeing an abstract set of numbers) what the numbers from the empirical data indicate about the real life decisions we must make.

Correlations in the real world of sex offense recidivism prediction

So, let's take a closer look at a sample of different correlations of 1.0, 0.0, and .07 and what they would mean, specifically when actually applied to sexual dangerousness decisions. Imagine a hypothetical sample of 100 sexual dangerousness hearings with a base-rate of recidivators (genuine SDP's) of 50%. This base-rate (50%) is consistent with Prentky, Knight, Lee, & Cerce (1997) who followed men (for a period of 25 years) after they were released from the Treatment Center for Sexually Dangerous Persons and estimated their recidivism rate if being *charged* with an offense meant that a new offense had been committed.[19]

Less than 10% of all sex offenders in Massachusetts were found to be "sexually dangerous" as were those in this study. All experts agree that the recidivism rate for the group of men previously adjudicated sexually dangerous, even after commitment and treatment, is higher than that of the general population of run-of-the-mill sanctioned and released sex offenders. In an SD commitment hearing, the question being considered is whether or not a man belongs in this high risk category.

In cases where experts participated in release decisions of men previously adjudicated SD, both the pattern of opining by experts and a reasonable recidivism

[19] In actuality, the ".07 Correlation Chart" presented later is as close to .07 as I could get with a sample of 101 and these characteristics. The exact correlation that this chart yields is .08. However, the feel one gets from looking at this chart of the degree of relationship between clinical judgment and actual recidivism—i.e., the sense of what a correlation of .07 means in the real world in these cases—is indistinguishable from a chart that had a correlation of exactly .07 and a recidivism base-rate of 50%.

A correlation of .07 between clinical prediction and actual recidivism is the estimate produced by Hanson and Bussiere's (1998) meta-analysis of studies of sex offender recidivism (Hanson, 1998). (Though Hanson suggested that the correlation was .10, when errors are corrected it comes out to .07. And, in any case, there is very little difference between the two in real life application; both are so close to zero that the utility of either is nil. The rationale for this correction is explained in more detail in Exhibit #8.

A recidivism rate of 50% was chosen because it is close to the *upper limit* (45%) of sex offender recidivism when *the most dangerous sex offender recidivists* were released and followed *for a period of 25 years* (see Attachment #4). The estimated base-rate Prentky, et al. produced for the probability of being charged with or arrested for a new sex offense over a 25 year period was 45%. See the discussion of the Prentky, et al. study (Exhibit #4) for the reasons to conclude that even their 45% figure is an overestimate. 50% is also conceptually easy to grasp as it corresponds to a common real life experience, the odds of tossing a coin and getting heads.

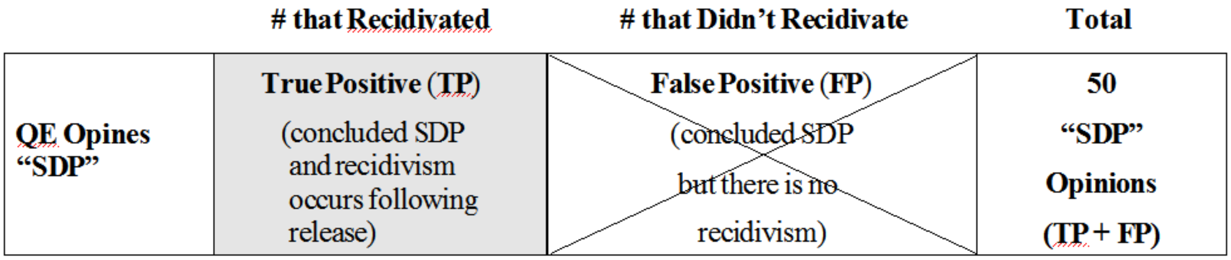
estimate is available, and thus the degree of accuracy (and the degree of bias, if any) in the experts' pattern of opining can be estimated.[20]

But before we get to that, let's be clear about what we know about the accuracy of prediction.

Accuracy Chart of Possible Outcomes

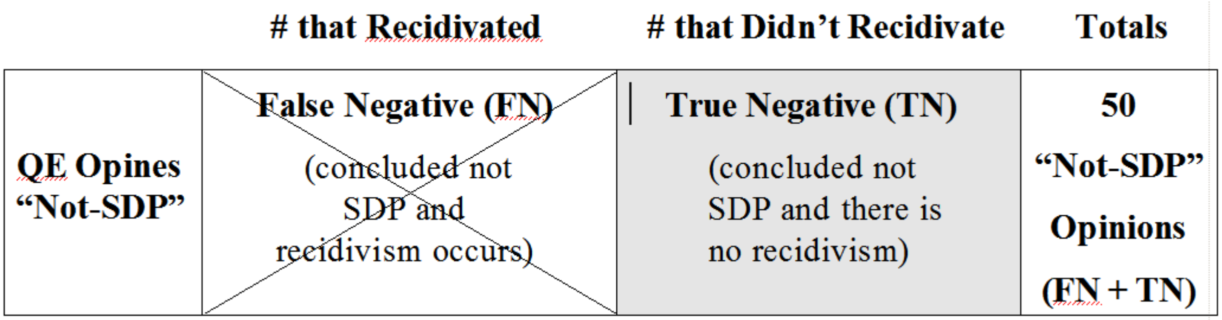
For 100 men released through Section 9 hearings,[21] the following are all possible outcomes. (Assuming a base-rate of recidivism of 50% and a pattern of opining that matches that base-rate, i.e., 50% of the time, the expert opines sexually dangerous.) Note: in the following charts, the term "Qualified Examiner" or QE is used to indicate the expert "qualified" by the state. QE thus signifies the state's expert.

When the Qualified Examiner opines, "sexually dangerous"



	# that Recidivated	# that Didn't Recidivate	Total
QE Opines "SDP"	True Positive (TP) (concluded SDP and recidivism occurs following release)	~~False Positive (FP)~~ ~~(concluded SDP but there is no recidivism)~~	50 "SDP" Opinions (TP + FP)

When the Qualified Examiner opines, "not sexually dangerous"



	# that Recidivated	# that Didn't Recidivate	Totals
QE Opines "Not-SDP"	~~False Negative (FN)~~ ~~(concluded not SDP and recidivism occurs)~~	True Negative (TN) (concluded not SDP and there is no recidivism)	50 "Not-SDP" Opinions (FN + TN)

[20] Note that while the data utilized comes from Massachusetts, there are reasons to believe that in Massachusetts (where on rare occasion SD men are released without opposition from the state) the pattern is *less conservative* than other states with SD statutes. For example, Washington's SD statute has been in operation for 19 years without any SD person being released. And the pattern in other states is similar to Washington's.

[21] A Section 9 hearing refers to a petition for release from civil commitment under Massachusetts General Law ch. 123A section 9.

All possible outcomes: When the Qualified Examiner opines either way

	# that Recidivated	**# that Didn't Recidivate**	**Totals**
QE Opines "SDP"	**True Positive (TP)** (concluded SDP and recidivism occurs following release)	**False Positive (FP)** (concluded SDP but there is no recidivism)	**50** **"SDP"** **Opinions**
QE Opines "Not-SDP"	**False Negative (FN)** (concluded not SDP and recidivism occurs)	**True Negative (TN)** (concluded not SDP and there is no recidivism)	**50** **"Not-SDP"** **Opinions**
Totals	**50 Recidivators,** **i.e., True SDP's**	**50 Non-Recidivators,** **i.e., Not SDP's**	**Total Men Examined 100**

The shaded boxes (on the downward diagonal line) are opinions/predictions that were accurate (TP + TN). The X'd boxes (the upward diagonal line) were errors in prediction (FN + FP).

The total number of predictions made or men evaluated is the sum of all the boxes or TP+TN+FP+FN.

Accuracy of prediction would be the Total along the downward diagonal (the shaded boxes) divided by the Total Number of Predictions made or Total Correct/Total Predictions or (TP+TN)/(TP+TN+FP+FN).

A Perfect Predictor: Correlation = 1.0

If the clinical opinions were perfect predictors (correlation = **1.0**), we would expect the following results:

	# that Recidivated	**# that Didn't Recidivate**	**Totals**
QE Opines "SDP"	**50** **True Positive** (concluded SDP and recidivism occurs following release)	**0** **False Positive** (concluded SDP but there is no recidivism)	**50** **"SDP" Opinions**
QE Opines "Not-SDP"	**0** **False Negative** (concluded not SDP and recidivism occurs)	**50** **True Negative** (concluded not SDP and there is no recidivism)	**50** **"Not-SDP" Opinions**
Totals	**50** **Recidivators, i.e., True SDP's**	**50** **Non-Recidivators, i.e., Not SDP's**	**Total Men Examined 100**

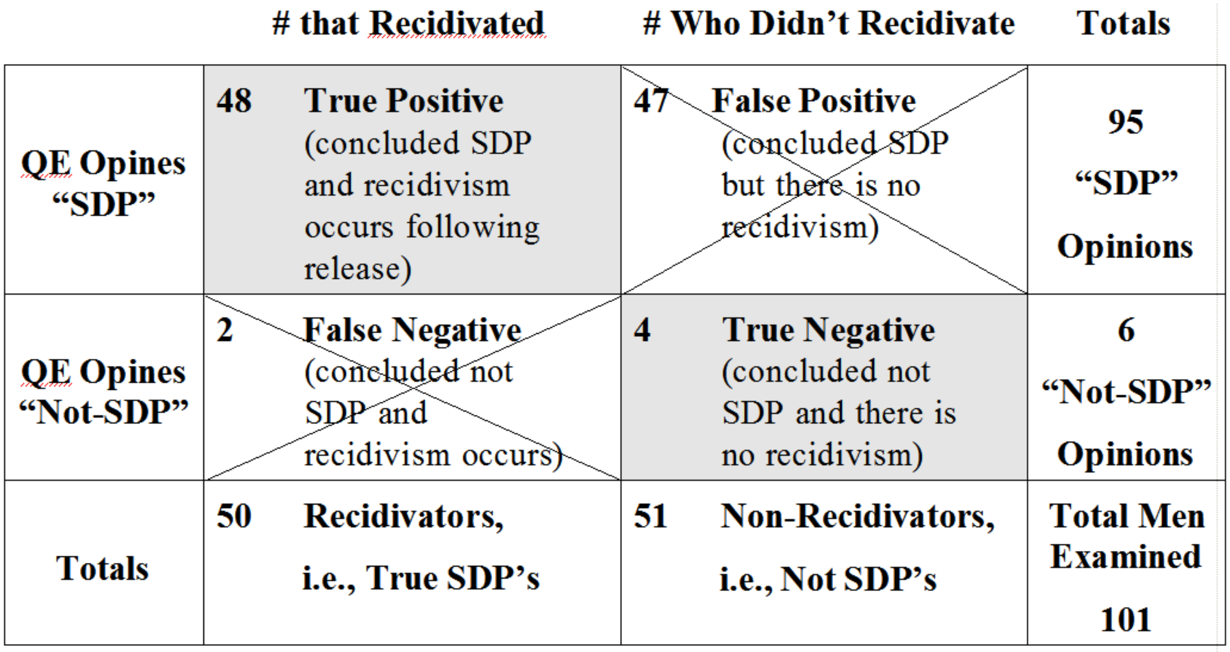
A Completely Random Predictor such as Tossing a Coin: Correlation = 0.0

If there were **no relationship** (no correlation, or *r* = 0.0) between expert predictions and recidivism we would expect the following results (on average):

	# that Recidivated	**# Who Didn't Recidivate**	**Totals**
Heads, QE Opines "SDP"	**25** **True Positive** (concluded SDP and recidivism occurs following release)	**25** **False Positive** (concluded SDP but there is no recidivism)	**50** **"SDP" Opinions**
Tails, QE Opines "Not-SDP"	**25** **False Negative** (concluded not SDP and recidivism occurs)	**25** **True Negative** (concluded not SDP and there is no recidivism)	**50** **"Not-SDP" Opinions**
Totals	**50** **Recidivators, i.e., True SDP's**	**50** **Non-Recidivators, i.e., Not SDP's**	**Total Men Examined 100**

What a Correlation of .07 "Looks Like" in The Real World

The actual rate at which the States' Examiners opine "SD" in cases where a commited SD person is being evaluated is greater than 95%.[22] The chart below represents a close approximation to what occurs in these hearings if the correlation between clinical judgment and outcome were approximately **.07**. This correlation (**.07**) is the correlation actually obtained between clinical judgment and recidivism in Hanson & Bussiere's (1998) meta-analysis of all known studies of predicting sex offender recidivism, involving 29,000 sex offenders.[23]



	# that Recidivated	# Who Didn't Recidivate	Totals
QE Opines "SDP"	48 **True Positive** (concluded SDP and recidivism occurs following release)	~~47 **False Positive** (concluded SDP but there is no recidivism)~~	95 "SDP" Opinions
QE Opines "Not-SDP"	~~2 **False Negative** (concluded not SDP and recidivism occurs)~~	4 **True Negative** (concluded not SDP and there is no recidivism)	6 "Not-SDP" Opinions
Totals	50 **Recidivators, i.e., True SDP's**	51 **Non-Recidivators, i.e., Not SDP's**	Total Men Examined 101

[22] Statistics available from testimony in Massachusetts from four experienced "qualified examiners" give approximate rates of concluding sexually dangerous when evaluating men whose commitments are being reviewed at the Treatment Center for Sexually Dangerous Offenders: 91% (82 out of 90), 98% (98 out of 100), 99% (237 out of 240), and 100% (11 out of 11), with the Community Access Board (which may have different experts for each review) at the Treatment Center averaging 99% (approximately 1185 out of 1200). The total is 1613 out of 1641 or 98% of the state's evaluators' opinions being SD. Note that the four specific QE's whose statistics are reported here were not selected from a larger group of Qualified Examiners; these are *all* of the examiners whose stats were available based on their sworn testimony. All of the other QE's have almost identical patterns of opining SD in such cases; there is rarely any discrepancy between the two state examiners on such cases. As noted, when these men are reviewed and released by the courts, the recidivism rate is under 45% (see Exhibit #4). It is important to note that Massachusetts is one of the *least* conservative states with SD statutes, i.e., in Massachusetts the experts do on rare occasion recommend (or do not oppose) release.

[23] As noted above, Hanson (1998) reported a correlation of .10 between clinical prediction and recidivism. However, this was due to a misclassification of a major study and an error in determining the correlation in another. When these errors were corrected, the correlation fell to .07. This correction is explained in more detail in Exhibit #8.

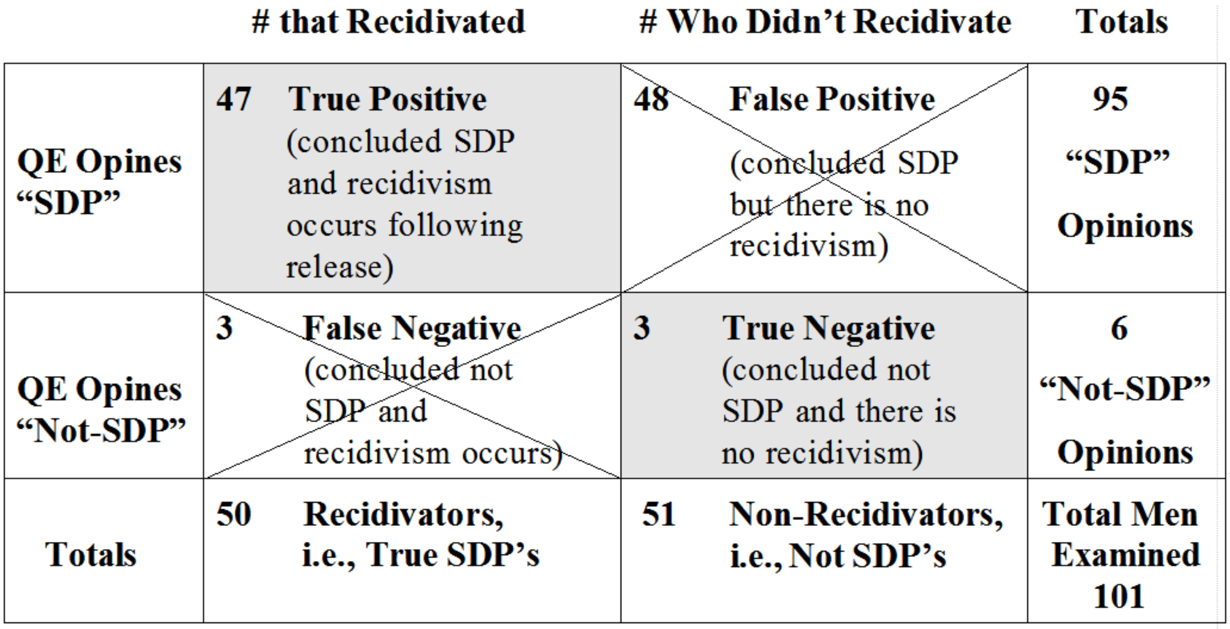
r = .08 (this was the closest I could get to .07 and required adding one more man to the sample, making the total 101. This gives the qualified examiners the benefit of the doubt as it slightly increases their accuracy and decreases their rate of over-predicting sexually dangerous.)

base-rate (of recidivists) = 50/101 = 50%

Note that the Correct Fraction = (True Positive + True Negative)/(Total Predictions Made) = 52/101 = 51%

Tossing a coin would, on average, yield a Correct Fraction of 50%.

If the QE's opined at the same rate (95 out of 101 opinions being "sexually dangerous) and there were **no relationship** (no correlation or *r* = 0) between QE opinions and recidivism we would expect the following results (on average):



	# that Recidivated	**# Who Didn't Recidivate**	**Totals**
QE Opines "SDP"	**47 True Positive** (concluded SDP and recidivism occurs following release)	~~**48 False Positive**~~ ~~(concluded SDP but there is no recidivism)~~	**95** **"SDP"** **Opinions**
QE Opines "Not-SDP"	~~**3 False Negative**~~ ~~(concluded not SDP and recidivism occurs)~~	**3 True Negative** (concluded not SDP and there is no recidivism)	**6** **"Not-SDP"** **Opinions**
Totals	**50 Recidivators, i.e., True SDP's**	**51 Non-Recidivators, i.e., Not SDP's**	**Total Men Examined 101**

Note the similarity between this chart (correlation = **0.0**) and the actual chart that we would expect given an **empirically obtained estimate of the recidivism rate** (**50%** or less), the **empirically obtained correlation between clinical prediction and actual recidivism** (**.07**), and the **actual opining pattern of the State's QE's** (greater than **95% "SDP"**) in commitment review cases.

The meaning of a correlation of .07

What this means is that if we were to *decide a priori* to find 95% of the men sexually dangerous and, in contrast, 50% of the men would actually recidivate, we could use a standard dartboard to determine how each man will be categorized. A standard dartboard has 20 "pie sliced" sections, each numbered 1 through 20. Each offender could chose a number between 1 and 20 and, with a blindfold on, the QE (or the offender, for that matter) could throw darts until the dartboard is hit, thus selecting one number. If it's the offender's number, he is declared "not sexually dangerous"; if not, he is declared "sexually dangerous." In this manner, 5% of the men would be found "not Sexually Dangerous" and 95% would be found to be "Sexually Dangerous" just as occurs in the real world.

Using this methodology, we would, on average, produce the results in the last chart. Using clinical methodology, however, we can expect to improve our accuracy to that in the previous chart. That is, we can expect an improvement over the Dartboard Methodology of 1 in both the True Positive and True Negative boxes based on the known *significant* correlation between clinical judgment and actual recidivism of **.07**!

More on the difference between *significance* and *validity*

Given a large enough sample, this increase in accuracy (from 50% to less than 52%) is "significant," i.e., there is a real, non-chance relationship between clinical prediction and recidivism; *clinicians* ***can*** *predict at a rate better than chance*. This means that the increase from 50% (chance) to 52% (the clinical method) is due to the fact that clinical predictions are *truly* (i.e., it is not a chance finding) better than coin tossing. But given the size of the increase in accuracy, we can understand why Hanson (1998) said that a correlation of less than .10 would have "little practical utility," i.e., it has almost no "**validity**" or meaning in the real world.

Compare the chart based on actual rates of opining and our best estimate of the upper limit of recidivism (with the empirically determined correlation between clinical opinion and actual recidivism of approximately .07) with the chart based on the same assumptions except the correlation is 0.0, i.e., there is no relationship between clinical prediction and recidivism. The difference between the two charts shows what a significant correlation of .07 indicates, i.e., the *difference* gives us a sense of the *validity* or utility of this *significant* relationship.

In our example in Exhibit #5, there was a real (i.e., genuine or significant) relationship between race and being assigned to category 2. But a correlation of .07 indicates that its strength (validity) is very low and that it accounts for approximately ½ of 1% of the variance, i.e., racial information alone is of very limited utility in making a decision about whether a specific individual student would be assigned to category 1 or category 2. Using this racial information, we would increase our accuracy to 53% from 50% (random guessing). If instead of being assigned a 1 or a 2 in the school gym for a school exercise, something of grave importance hung on this decision, we would be committing a grave injustice if we based our predictions on race. Just so, a large body of research now conclusively demonstrates that clinical predictions provide this same degree of information and are essentially useless in trying to predict sexual recidivism.

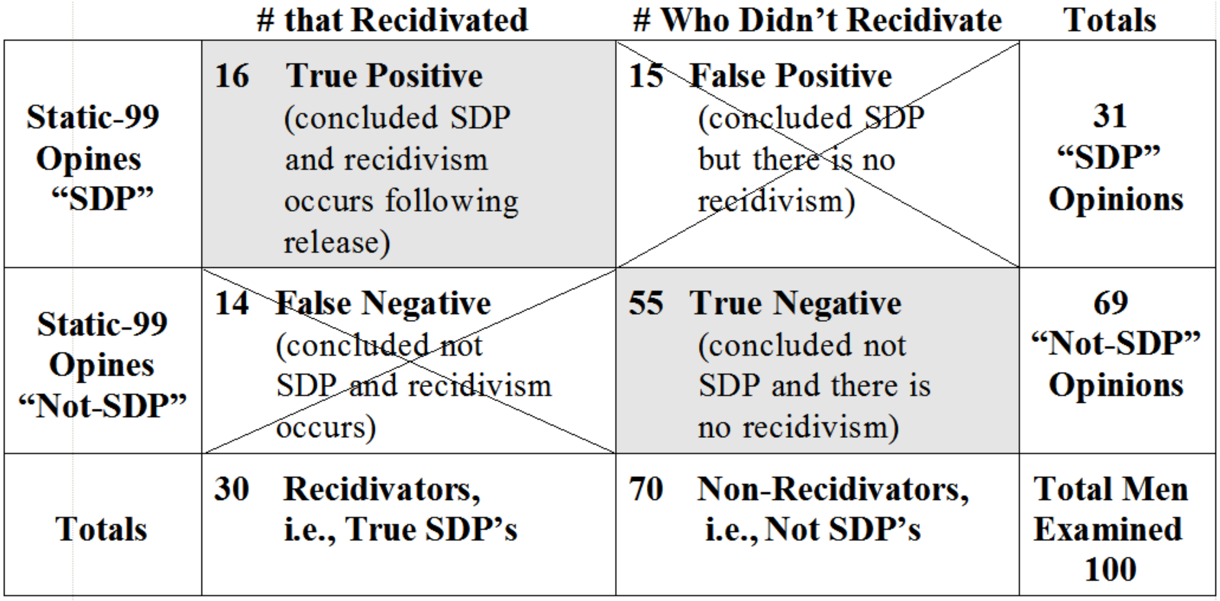
So, while "dropping out of treatment," for example—a frequently cited variable given a high degree of importance by the State "experts" making these predictions—is empirically *significant* when predicting recidivism (there is a *real* relationship between the two), the strength of the relationship is very weak, *validity* is low, and it is a poor predictor of recidivism. Thus, a factor such as "quitting treatment" that accounts for less than 3% of the variance (in whether or not an offender will reoffend) is *nearly* useless; 97% of the variance is unaccounted for. When properly combined with other factors (the way to do this will be discussed below), this significant factor may be able to provide some usable information, albeit, not a great deal. When improperly relied upon to buttress a foregone conclusion of sexual dangerousness, the factor can be used to provide impressive weight to an expert's opinion unless the fact finder is told about its very poor validity and has some understanding of what that means.

This also explains why psychologists who use clinical assessments to make predictions make so many errors and are worse at making accurate predictions than lay people: When you use factors with little or no relationship to the outcome being predicted and overemphasize those factors, you are clearly going to degrade the predictive utility of genuine factors. When one uses information unrelated to the occurrence of a behavior in the future to predict that behavior, then the information that may be used to accurately predict future behavior is devalued and the error rate always goes up. If there is a clear bias toward reaching a particular conclusion, then the unrelated information will not merely degrade the value of the related information and add random noise. Rather, it will tend to completely eclipse the impact of any truly relevant data and render the opinion almost totally non-predictive. If such opinions are nearly always "sexually dangerous," as the State's expert's opinions are in recommitment hearings, then we have valueless predictions that almost always lead to the conclusion of sexual dangerousness.

The Application of Actuarial Correlations (Exhibit 7)

Since the actuarial method is far superior in accuracy to the clinical method, how much can we rely on it? We know that the correlation between actuarials and recidivism is between .30 and .34. Therefore, the percent of the variance in recidivism predicted by the actuarials is between 10 and 12%.[24] What does that tell us about how much we can rely on actuarial prediction in making SD decisions?

So what does a correlation between .30 and .34 look like in actual practice?



	# that Recidivated	# Who Didn't Recidivate	Totals
Static-99 Opines "SDP"	**16 True Positive** (concluded SDP and recidivism occurs following release)	~~**15 False Positive** (concluded SDP but there is no recidivism)~~	**31 "SDP" Opinions**
Static-99 Opines "Not-SDP"	~~**14 False Negative** (concluded not SDP and recidivism occurs)~~	**55 True Negative** (concluded not SDP and there is no recidivism)	**69 "Not-SDP" Opinions**
Totals	**30 Recidivators, i.e., True SDP's**	**70 Non-Recidivators, i.e., Not SDP's**	**Total Men Examined 100**

This chart presents an average example of what an *r* = .32 (between actuarial prediction and recidivism) accounting for 10% of the variance (in recidivism) would look like in actual practice, if 30% of the men are SD. Note that 30% is in the upper end of the range of reasonable estimates of the base rate of recidivism of sex offenders who have *not* previously been adjudicated SD. That is, it is a conservative estimate (probably an over-estimate) of the base rate of recidivism for men initially evaluated for commitment. If we employ the Static-99 to determine who is and who isn't SD and we set a cutoff Static-99 score so that it renders 30% SD predictions (in order to match our base rate estimate so that we do not introduce any bias one way or another), something similar to this chart is what we should expect to see.

In this example, which comes closest to a real life best method possibility, when saying SD, the actuarial method would be right a little more than 50% of the time (16 out of 31) and wrong a little less than 50% of the time (15 out of 31). Could that possibly meet the "clear and convincing" standard of evidence necessary to deprive a man of his liberty, potentially for the rest of his life?

[24] See Exhibits 5 and 6.

State experts typically say "He scored in the 'high risk' category on the Static-99." From the chart above, if we set the high risk category so that it matches the base rate and doesn't produce bias, when an offender "scores in the 'high risk' category" and based on that the expert opines SD, the offender will actually be SD slightly more than 50% of the time. It seems clear that presenting this actuarial data to a fact finder with the implication that it is objective (which it is), scientifically sound (which it is), and valid (which it is) is grossly misleading if the limits of the validity are not carefully explained to a fact finder who understands the distinction between validity and scientific significance.

Clinical Variations (Exhibit 8)

More recent studies have again shown some potential for the guided clinical approach (Hanson & Morton-Bourgon; 2004). However, these studies suggest that the guided clinical approach is midway in accuracy between the invalidated pure clinical approach and the minimally accurate actuarial approach. When we turn to understand the implications of the degree of validity of the most accurate method—the actuarial approach—we see that its validity may not be sufficient to meet the legal criteria for most SD cases (Exhibit #7 and #10). Since the guided clinical approach appears to be significantly less valid, if the low validity of the actuarial is insufficient, we need not consider the guided clinical approach. And this is compounded by the fact that a true guided clinical approach is rarely employed by state's experts.[25]

On the other hand, the adjusted actuarial approach may have some utility. But as yet, there have been no studies of the adjusted actuarial method that demonstrate its validity. Indeed, there are reasons to believe that allowing *any* clinical adjustments lowers validity. For example, in the discussion of the Goldberg Rule, a simple mechanical prediction tool for making diagnostic predictions (discussed in Exhibit #2), I noted the following:

> Goldberg gave some of the judges (including all of the experts) the information from the Goldberg Rule on each case and allowed them to use the extra information (that the rule in its simplicity could not make use of) and their clinical judgment to use or modify the prediction made by the rule. Though there were some gains in accuracy, no judge did as well as the rule: Every judge would have been more accurate if they avoided using clinical judgment and always used the rule alone.

Golberg demonstrated gains in accuracy when the information from the simple mechanical measure was available to the clinicians. But allowing them to use clinical judgment along with the actuarial-like mechanical measure decreased the validity of their conclusions from what would have been the result if they went with the unadjusted actuarial.

While adjusting actuarial assessments may at times be necessary and can be reasonably defended in certain circumstances, there is no empirical evidence that this improves accuracy and considerable evidence that it decreases accuracy (Goldberg, 1959; Quinsey & Maguire, 1986; Dawes, Faust, and Meehl, 1989; Webster, Harris, Rice, Cormier, & Quinsey, 1994; Quinsey, Harris, Rice, & Cormier, 1998). For another example, consider the best predictor of non-sexual violence, the Violence Risk Appraisal Guide (VRAG). The VRAG was also intended to be an adjusted actuarial tool, when Webster, et al. (1994) created it. After less than a

[25] In Massachusetts in recent years, when asked about the methodology they employed, the state's experts typically claim to be using the guided clinical approach. I have now heard such claims made on more than 50 occasions. Yet, I have never heard of a single case in which a guided clinical method of the type that Hanson reported as showing some validity was actually employed. The claim is now often made that an adjusted actuarial method is being employed. Again, the method utilized may include the use of an actuarial, but the clinical speculations used to adjust it make the state experts' "adjusted actuarial method" unlike any that has been researched and shown to be valid.

decade of experience with the measure, they recanted and eliminated all clinical adjustment (Quinsey, Harris, Rice, & Cormier, 1998).

With no studies demonstrating any improvement in accuracy when using the adjusted actuarial approach, consider the following discussion of the problem with *minor* adjustments to actuarial measures, i.e., the "cautiously adjusted actuarial" approach.

> Clinicians might be able to gain an advantage by recognizing rare events that are not included in the actuarial formula (due to their infrequency) and that countervail the actuarial conclusion . . . In psychology, this circumstance has come to be known as the "broken leg" problem on the basis of an illustration in which an actuarial formula is highly successful in predicting an individual's weekly attendance at a movie but should be discarded upon discovering that the subject is in a cast with a fractured femur . . . The clinician may beat the actuarial method if able to detect the rare fact and decide accordingly . . .
>
> The broken leg possibility is easily studied by providing clinicians with both the available data and the actuarial conclusion and allowing them to use or countervail the latter at their discretion. The limited research [now less limited] examining this possibility, however, all shows that greater overall accuracy is achieved when clinicians rely uniformly on actuarial conclusions and avoid discretionary judgments . . . When operating freely, clinicians apparently identify too many "exceptions," that is, the actuarial conclusions correctly modified are outnumbered by those incorrectly modified . . . The available research suggests that formal inclusion of the clinician's input does not enhance the accuracy, nor necessarily the utility, of the actuarial formula and that informal or subjective attempts at adjustment can easily do more harm than good . . . (Dawes, et al., 1989, pp. 1570-1671)

Rather than making harmful adjustments to an accurate actuarial prediction, the States' evaluators almost always override the actuarial prediction altogether and present a hodgepodge of "adjustments."

In summary, at this point our best methodology is the actuarial approach and there are serious limitations to that method. At some point, the courts will have to address the issue of whether—outside of extreme cases which do occur (and, I would suggest, for which we have little or no need for an expert)—we have a predictive methodology that can, with sufficient accuracy to deprive an individual of liberty potentially for life, differentiate the dangerous sexual recidivist from the ordinary sex offender.

The Psychological Limitations of Opining Sexual Dangeroussness (Exhibit 9)

Sexual Dangerousness: A legal term that seems to cause behavior

It is highly misleading to talk about the diagnoses—and even more so, to talk about legal terms—as sources or causes of behavior. They *are* behavior. What causes the particular constellation of behaviors and reports of feeling states that we call "depression" can't be "depression." No phenomenon (other than God) has ever been proposed as its own cause. As the American Psychiatric Association (author and publisher of the *Diagnostic and Statistical Manual of Mental Disorders*) put it:

> [P]sychiatric predictions of violent conduct unduly facilitate a jury's finding of future dangerousness by providing a clinical explanation of what is, at best, only an assessment of statistical probabilities. A medical diagnosis of antisocial personality disorder is in essence a descriptive label for past aberrant behavior of a particular type. When stated to a jury, however, the term . . . "antisocial personality disorder" . . . conveys the erroneous impression of an endogenous disorder which bears a cause-and-effect relationship to similar, future behavior. Medical opinions in this area thus offer a jury a seductively facile—but wholly unfounded—explanation as to why a particular individual having the statistical "symptoms" of a recidivist will in fact be a recidivist.[26]

Legal Terms vs. Diagnoses: The meaning of "sexual dangerousness"

Unlike a finding of sexual dangerousness—which is, in large part, a *prediction* that an offender *will* commit a new offense—a finding of not-sexually dangerous is *not a prediction* that an offender *will not* commit a new offense. Confusion about this results from conflating *clinical diagnoses* with a *legally defined term* that sounds like a clinical illness. "Sexual Dangerousness," is not a mental disorder or condition; it is a legal term. One can not have "sexual dangerousness" like one can have "chicken pox." Therefore, it is not meaningful to reify the term and consider it a condition within the patient.

Qualified Examiners can use diagnoses of conditions that exist within the patient—aspects of the patient's personality, characteristic ways of coping, ways of organizing one's experience of the world, etc.—to predict behavior, that is, to ascertain whether such behavior is *likely*. Leaving aside the fact that such clinical predictions have poor reliability and poor validity, the determination of *the likelihood of a future offense* can be a prediction of behavior based, in part, on diagnoses and an understanding of what goes on within a patient's psyche, but certainly should not be treated as something that itself exists within the patient. Thus, unlike medical tests for infections or conditions and diagnoses, when an expert says that a man is not sexually dangerous, he is not saying that a "condition" doesn't exist; he is not claiming that the man is free of an illness called sexual dangerousness and thus poses little or no risk.

[26] *Amicus* Brief for the American Psychiatric Association at 13, Barefoot v. Estelle, 463 U.S. (1983) (No. 82-6080).

Thus, when an expert says there is no basis upon which to conclude with reasonable scientific or professional certainty that a man is sexually dangerous, he is not saying the individual is free of a "disorder" called "sexual dangerousness," and is therefore *unlikely* to commit a new offense. As the American Psychiatric Association put it:

> Because most psychiatrists do not believe that they possess the expertise to make long-term predictions of dangerousness, they cannot dispute the conclusions of the few who do. Instead of offering a countervailing prediction of nondangerousness in a particular case, they (or defense counsel on cross-examination) can only advance a challenge to the asserted "expertise" of the prosecution psychiatrist. The likely result, therefore, is that the jury will hear not the traditional battle of expert conclusions, but only a dispute over one expert's ability to reach the conclusion that he did.[27]

This cripples the defense who has to face experts who say their client is dangerous using experts who refuse to say their client is not dangerous and typically will only say there is no foundation for an opinion that he is dangerous. Indeed, we would need the same foundation—the same ability to accurately predict long term behavior using clinical judgment—in order to use clinical judgment to say that the offender is unlikely to reoffend.

However, despite this problem, in many cases the defense expert can firmly declare that the offender is *not sexually dangerous*. How can this be? To answer this, we need to keep in mind the fact that "sexually dangerous" is not a psychological term; the court does *not* turn to the expert for help in understanding what "sexually dangerous" means; the court wants to know whether there is a professional or scientific basis with a high degree of certainty/validity to apply the term *defined by the legislature* and *interpreted by the court* to a particular offender. In fact, I have frequently not been allowed to testify as to its meaning because it is a legally defined term, created by the legislature and interpreted by the courts. I note this because legal terms have implicit aspects to their meaning that are not part of psychological terms. For example, someone can be said to suffer from "depression" (a clinical term) if depression is the best (i.e., most probable, or "to a preponderance of the evidence") diagnosis. However, a person is not "guilty" (a legal term) of an offense if they "to a preponderance of the evidence, probably" did it.

With a legally defined term, *a standard of proof* is an implicit—usually made explicit at some point by the courts—part of the definition. In this particular case, if

> the court finds by clear and convincing evidence that the person is a sexually dangerous person, the court shall commit the person. [The] term "sexually dangerous person" means a person suffering from a serious mental illness, abnormality, or disorder, as a result of which the individual would have serious difficulty in refraining from sexually violent conduct or child molestation. (from the *Adam Walsh Act*)

[27] *Amicus* Brief for the American Psychiatric Association at 13, Barefoot v. Estelle, 463 U.S. (1983) (No. 82-6080).

Note that this term "sexually dangerous person" does not exist in psychology or psychiatry. Sexual dangerousness is a legal *finding* that can only be made when there is "clear and convincing evidence" that "a person [is] suffering from a serious mental illness, abnormality, or disorder, as a result of which the individual would have serious difficulty in refraining from sexually violent conduct or child molestation." Thus, by the very definition of this *legal* term, the meaning of "sexual dangerousness" is that a court has found that there is clear and convincing evidence that the individual suffers from a serious mental disorder that results in creating serious difficulty in refraining from sexually violent conduct or child molestation.[28]

As a legal term, if an expert opines "sexually dangerous" or "not sexually dangerous," the expert is giving an opinion on a legally defined matter (i.e., not clinically defined by either psychology or psychiatry); this is called for by the statute. Thus, when an expert opines that an offender is "not sexually dangerous," they are saying that the criteria are not met for such a conclusion, i.e., when there must be "clear and convincing evidence" that the offender meets the criteria in order to be considered SD, if it cannot be proved at that level, it is not so. E.g., "innocent until proven guilty" does not mean that someone who is "not guilty" has had their innocence proven. Innocence is the assumed state (does not need to be proven) and is logically equivalent to "the inability to establish guilt." In this case, if "sexual dangerousness" cannot be established to the fairly high degree of certainty based on "clear and convincing evidence," the individual is *not* SD.

"Sexual dangerousness" is a legal term that indicates that *all* of the following five conditions are present:

1. the respondent has engaged in sexual misconduct, *and*

2. has a serious mental illness, abnormality, or disorder, *and*

3. there is reason to believe that the abnormality/disorder has an influence on future behavior, *and*

4. the influence must be of the type that creates a serious difficulty in controlling sexual behavior creating a significant risk of sexually violent conduct or child molestation, *and*

5. we must have evidence and reasoning to establish all of the above to a high enough level of certainty that a reasonable person could conclude that they had been established by "clear and convincing evidence."

If any of these five factors is missing or if any of the first four factors cannot be established by clear and convincing evidence, the person is *not* sexually dangerous. Thus, "sexually dangerous" means we have established *all of these*. "Not sexually dangerous" does

[28] We also know from the *Crane* decision that the "serious difficulty in refraining from . . ." can also be construed to mean that, due to the mental disorder, the offender has a serious difficulty in controlling behavior. *Kansas v. Crane*, 534 U.S. 407 (2002).

not mean the opposite (i.e., that we are certain that the person is unlikely to commit a new sexual offense); it simply means that any one of the five elements cannot be established.

Why an Actuarial Prediction Can Never Be a Sufficient Basis for a Finding of Sexual Dangerousness (Exhibit 10)

Randomization

If we take a *random* sample from a larger population, then our sample, on average, should reflect the characteristics of the larger population. The entire purpose of using *randomization* to select our sample is precisely *to avoid introducing* ***any*** *bias to our sample*, i.e., to ensure that our sample is essentially the same as (is fairly representative of) the larger population from which it was drawn. The famous picture of Truman holding up the headline "Dewey Wins" was the result of the use of a non-random telephone sampling of voters. At that time, Republicans were more likely to own telephones. So bias was introduced by using a telephone poll; it was not a truly *random* sample of American voters and thus the *sample* polled did not accurately reflect the true voting pattern of the larger *population* of all voters.

If we take a truly *random sample* of sex offenders, the recidivism rate in our sample should, on average, be the same as the recidivism rate in the larger *population* of all sex offenders. So if we toss a coin for each sex offender and place all those who come up heads into Sexually Dangerous (SD) category and all those who come up tails into the not-Sexually Dangerous (not-SD) category---regardless of our labels, because of the random *method* we used to select the members of the labeled groups---the SD group and the not-SD group will be *random samples*. As such, the two samples will, on average, have the same rate of recidivism as the larger *population* of sex offenders (as well as the same rate of recidivism as the other group).

Thus, using a method of predicting that has no validity—which is exactly the same as using a completely random method of assignment such as tossing a coin—would produce groups of sex offenders that have the same recidivism rates as the larger population that they were selected from. The rate of recidivism in our SD sample would, therefore, be the same as the *Base Rate* of recidivism in the larger *population* of all sex offenders.[29]

The goal of prediction by an expert is to help us select out the *most* dangerous sex offenders. If the expert used a random method—i.e., one which like coin tossing were accurate 50% of the time—then there would be no difference between those labeled SD and those labeled not-SD. And both groups would have the same *Base Rate* of recidivism as the larger population of sex offenders. On average, a member of such a randomly selected SD group would be no more or less dangerous than the average sex offender.

The goal of using an expert is to identify sex offenders who are more likely to reoffend than the average sex offender, i.e., those truly SD offenders whose rate of recidivism is higher than the Base Rate for the average sex offender. If we have no accurate method of selecting the more dangerous from the less dangerous, then experts are of no use in helping us differentiate the SD from the average sex offender. But we do have some demonstrable accuracy in

[29] While the probability of a new sex offense is just one element of determining SD, it is an essential element. Even if an offender committed a horrendous sexual offense and suffered from a serious mental condition that caused him to do so, if we somehow knew he would never commit another sexual offense then he could not possibly be SD. The analysis being presented in this attachment only looks at what we would need to have to meet this *necessary* condition for an SD determination: Whether a new sexual offense can be clearly and convincingly shown to be sufficiently likely to make an offender SD.

prediction. So we need to know if the accuracy we have can reasonably be expected to produce evidence that is clear and convincing.

Measuring Accuracy of an SD Opinion: The Positive Predictive Value

We have a way of measuring how much help an expert can offer us: the Positive Predictive Value of the expert's opinions. This is a simple notion: it is the percent of the time the expert is correct *when he opines SD*. Thus, for example, if an expert were to opine SD 100 times and those 100 offenders were released and followed over a 60-year period during which 30 recidivate and 70 did not, the expert's SD predictions would have been correct 30% of the time: the Positive Predictive Value = 30%.

How can we know if this is an improvement in accuracy over chance, over coin tossing or random guessing? To know that, we would have to know the Base Rate in the larger population. If 15% of sex offenders reoffend, then that expert's method would be a significant improvement over random guessing/coin tossing which would have produced a SD sample that had the same rate of recidivism as (the Base Rate of) the larger population or 15%. 100 SD predictions using a completely random method would produce a SD group of 100 sex offenders, of whom 15 would recidivate. Since 30 out of 100 of the expert's SD opinions recidivated, the expert was much better than (in fact, twice as good as) chance/coin tossing. The expert's method was clearly *not* random.

If, however, the Base Rate of the larger population is 30%, then our expert did no better than chance. He would have been just as accurate tossing a coin to select his SD group of 100 and his predictive method and "expert" opinion is useless in determining SD. In this case, he would have given us a group or sample of sex offenders which he labeled SD that has no greater propensity to reoffend than the average sex offender.

The question then is, "Does the Positive Predictive Value of a method indicate an improvement over chance or random guessing?" And since random guessing will produce a group with the same recidivism Base Rate as the general population of sex offenders, this question is the same as "Does the Positive Predictive Value of a method produce a selected group that is more likely to reoffend than the Base Rate of recidivism of the average sex offender?" If we use a random method (like coin tossing) and our predictions of SD are accurate 50% of the time, then the Positive Predictive Value and the Base Rate will be the same.

The Positive Predictive Value vs. the Hit Rate

When our predictions of SD are accurate 50% of the time (i.e., tossing a coin), then the

Base Rate = Positive Predictive Value (where the Positive Predictive Value = percent of the time we are correct when we conclude SD)

That is, we can never improve over the Base Rate when our method is random like coin tossing. On average, a random method will give us the exact Base Rate of accuracy within the group we identify as SD using that random method. That's basically another way of defining a "random method," i.e., one with no accuracy and no bias, completely random.

However, we can improve our **Hit Rate**, even when using a random method.

Hit Rate = the percentage of truly SD men who are identified as SD, i.e., what percentage of SD men is our method "capturing." This is also known in the medical testing literature as the "sensitivity" of the test.

When using a random method, we can capture more SD's simply by opining SD more often. For example, we can toss two coins for each sex offender and only opine "not-SD" when *both* coins come up tails. Since two tails will occur 25% of the time, this will produce 75% SD conclusions and will capture 75% of the SD men even though the method is completely random. If we *always* opine SD, we will capture 100% of the SD men, even when coin tossing is our method of opining.[30] While our Hit Rate (how many SD's we capture) will be 100%, our Positive Predictive Value will be the Base Rate of Recidivism in the general population of sex offenders (since we will be selecting ALL sex offenders for inclusion in our SD group) and our predictions will be random and will provide us with no additional valid meaning whatsoever.

Thus, Hit Rate (percent of SD men "captured" by our method) is highly dependent on *how often we opine SD* and the Hit Rate does NOT equal the Positive Predictive Value (*the percent of the time we are correct when we do opine "SD"*). We can bring our Hit Rate to 100% by always opining SD, even though, if we are using a random, useless predictive method, our Positive Predictive Value may be 15, 20, 30, or 50% (i.e., whatever the Base Rate is).

The effect of Base Rates on the Positive Predictive Value

If 50% of all sex offenders are truly SD (if 50% of all sex offenders actually reoffend when released) and we toss a coin for 100 men (Heads = SD), then, on average, we will get 50 men who we diagnose as SD of which 25, or 50% will, in fact, be SD. In the following chart, the red row highlights the information needed to determine the Positive Predictive Value of an expert's opinion that an offender is SD.

[30] How can we always opine SD if we are using coin tossing as our predictive method? Simple. Toss 20 coins and only opine not-SD if all 20 come up tails at the same time. Though that will happen once in a million or so tosses, if necessary, we can easily increase the number of coins so that it never happens.

Base Rate = 50%

	# Who Recidivated	# Who Didn't Recidivate	Totals
Heads, Expert Opines "SD"	**25 True Positive** (concluded SD and recidivism occurs following release)	~~**25 False Positive** (concluded SD but there is no recidivism)~~	**50 "SD" Opinions**
Tails, Expert Opines "Not-SD"	~~**25 False Negative** (concluded not SD and recidivism occurs)~~	**25 True Negative** (concluded not SD and there is no recidivism)	**50 "Not-SD" Opinions**
Totals	**50 Recidivators, i.e., Truly SD**	**50 Non-Recidivators, i.e., Not SD**	**Total Men Examined 100**

We will have correctly identified 50% of the 50 truly SD men (Hit Rate = 50%) and, as we saw above, the Positive Predictive Value when using a random method for concluding SD will also be the Base Rate or, in this case, 50%. Thus, half the time when our expert opines "SD" he will be right, using a method that is totally random.

However, if 25% of the men are SD and we toss a coin for 100 men (heads = SD), then we will again get 50 men who we diagnose as SD, but only 12.5 (on average), or 25% of the 50 SD opinions will actually be correct.

Base Rate = 25%

	# Who Recidivated	# Who Didn't Recidivate	Totals
Heads, Expert Opines "SD"	**12.5 True Positive** (concluded SD and recidivism occurs following release)	~~**37.5 False Positive** (concluded SD but there is no recidivism)~~	**50 "SD" Opinions**
Tails, Expert Opines "Not-SD"	~~**12.5 False Negative** (concluded not SD and recidivism occurs)~~	**37.5 True Negative** (concluded not SD and there is no recidivism)	**50 "Not-SD" Opinions**
Totals	**25 Recidivators, i.e., Truly SD**	**75 Non-Recidivators, i.e., Not SD**	**Total Men Examined 100**

In this case, we will have identified 50% of the 25 truly SD men and our Hit Rate will be 50% (which it always will be if we toss a coin or use a truly random method that produces a SD conclusion half the time), but our Positive Predictive Value will be 25%. 25% of the time when we say SD we will be right, the man will be SD; 75% of the time when we say SD we will be wrong, the man will not be SD.

This means that the probability that an SD opinion offered by an expert is correct is highly dependent on the Base Rate in the larger population. If almost all sex offenders are SD, then an expert—even one using a minimally accurate method—will be right almost all those times when he opines SD. If almost no sex offenders are SD, it will be far less likely that an expert using such a method will be right when opining SD

To repeat where we started, the probability of being actually SD given an expert opinion of SD (i.e., the Positive Predictive Value) based on a test that is 50% accurate is always the same as the Base Rate. That is, a random method like tossing a coin will neither improve nor degrade the Base Rate estimate and we will be identifying as SD a random group of sex offenders. Truly *random methods* will yield a group of identified SD men who are no different from the general population of sex offenders, i.e., a group with the same Base Rate of true SD.

Thus, *the probability of being SD when diagnosing SD (the Positive Predictive Value) only rises above the Base Rate when conclusions of SD are made using a nonrandom method that is accurate more than 50% of the time* (and the Positive Predictive Value is lower than the Base Rate if the test is less than 50% accurate).

Because we have methods that are more accurate than random guessing---and we are told that those methods have achieved acceptance in the scientific community as being reliable and *valid* (see Attachment #5)—the question is whether they are sufficiently valid to produce evidence that could be clear and convincing. Let’s take a look.

“Clear and Convincing Evidence” and the Positive Predictive Value

As a psychologist, I cannot translate into an exact probability figure the level of certainty needed to reach the *legal* standard of “clear and convincing.” If such a figure exists, it would be for the courts to determine. However, expert opinions that we know are wrong more often than they are right—opinions that could not even be the basis for “a preponderance of the evidence”—could not possibly provide the essential foundation for the “clear and convincing evidence” necessary for a finding of SD. Since the Positive Predictive Value is the percent of times we are correct when we clinicians conclude and opine SD, if our conclusions are wrong more often than they are right—if the Positive Predictive Value is less than 50%—then such conclusions could never be “clear and convincing.”

Thus, if we only commit men as SD when we are fairly sure (i.e., when the “evidence is clear and convincing”) that they are SD (and thus we are not committing a larger number of men who are not-SD), then we only want to use expert opinions as a primary component of such evidence when they are at least correct more often than they are wrong. Thus, such expert opinion could only be a basis for clear and convincing evidence when the Positive Predictive Value of expert opinions is well above 50%,[31] i.e., significantly better than tossing a coin. As

[31] Note that this is NOT the same as saying there is a greater than 50% chance that *the offender* will reoffend. What we are saying is that *the expert’s opinion* has a greater than 50% chance of being accurate, i.e., the expert’s opinion is more accurate than tossing a coin. Even when the expert opines not SD, we

we will see shortly, if the Base Rate is well below 50%, we would need a *very* accurate method to get our Positive Predictive Value above 50%.

We have no such method. Given our current limitations, using our best methods, as we shall see, we may never be able to have clear and convincing evidence that a particular man whom we are considering and who is in the highest risk category by our best predictive measure is in fact truly high risk; more often than not we would be wrong, and that cannot be the basis for clear and convincing evidence.

The need for substantial accuracy when Base Rates are low

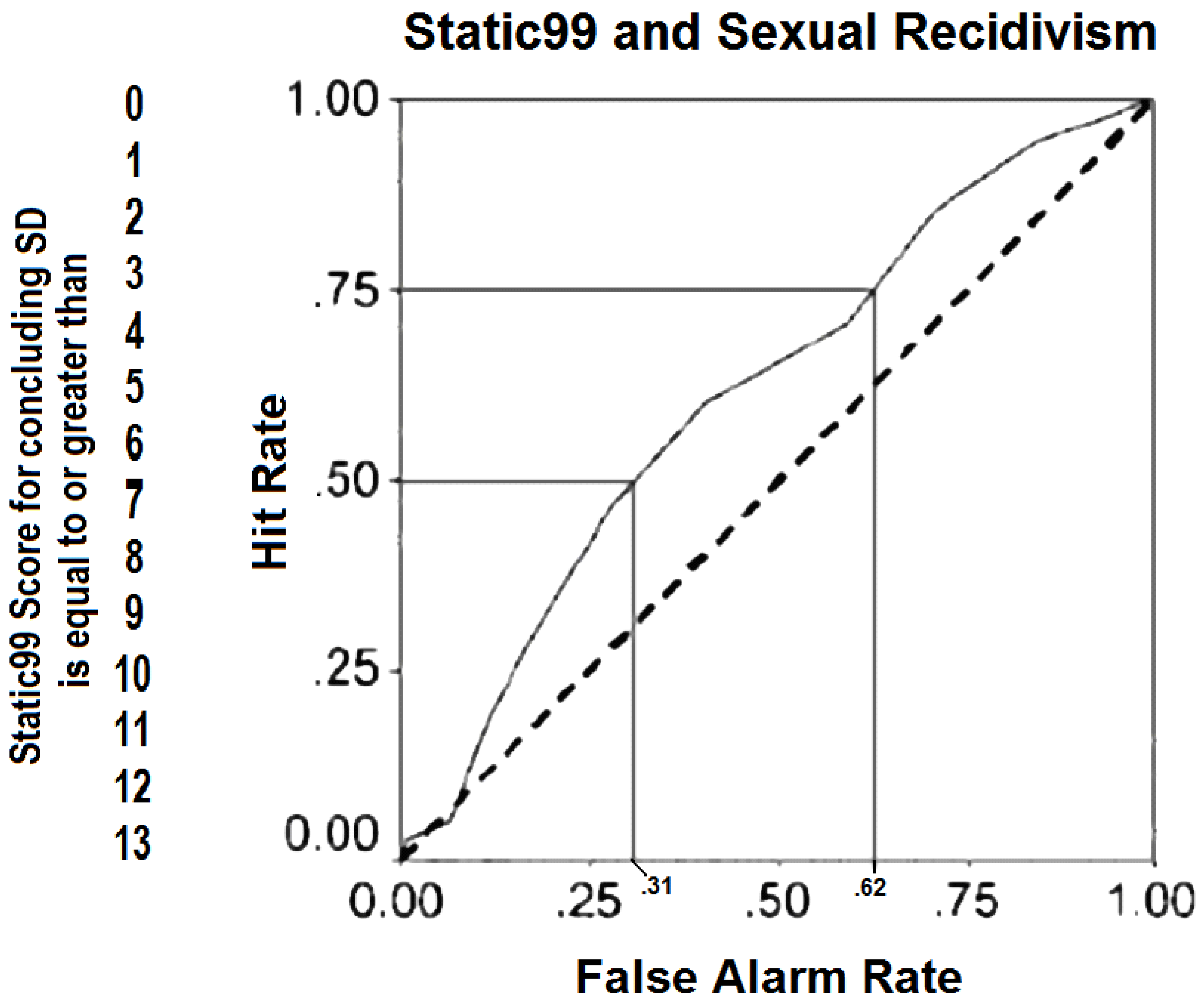
To understand why a low Base Rate necessitates a very accurate method (more accurate than any method we currently possess) for an opinion of SD to be correct more often than it is wrong, first consider that all prediction methods use "cutoffs," i.e., above a certain level the man is SD and below it not-SD. This is true of the subjective clinical method when the "expert's" feelings/intuition reach a certain subjective level and he opines SD, as well as of objective methods where we use an actual numerical cutoff score. Given that all methods have cutoffs, we can now look at our predictive method and see how the Base Rate interacts with our accuracy in terms of whether we have evidence that could be "clear and convincing."

In the following chart,[32] the diagonal dotted line is what we would get by tossing a coin. The wavy line is comprised of the plotted results from the real life use of the Static-99—and the Static-99 is an example of one of our best (most accurate) predictive methods—on actual samples of sex offenders. Note that the plotted ROC[33] curve shows how the Hit Rate rises (we "capture" more of the truly SD using the Static-99) along with the False Positive Rate (more not-SD Dolphins get caught in our SD Tuna net) when we lower the threshold (the Static-99 cutoff score) for reaching a conclusion of SD.

would need a Positive Predictive Value greater than 50% to know that that opinion has any validity. We are examining when we should listen to an expert's opinion about the probability of a reoffense and when we should ignore it and use an estimate of the base rate as our best predictor.

[32] The chart is taken from "A Multisite Comparison of Actuarial Risk Instruments for Sex Offenders," Harris, Rice, Quinsey, Lalumie, & Boer, 2003, *Psychological Assessment*, 15, 3, 413–425, with information added about the different cutoff scores and the specific Hit Rates at 50 and 75%.

[33] Receiver Operator Characteristics curve. This type of curve was derived from signal detection theory and is something you need know nothing about to understand this analysis. I include this footnote merely for reference as other people will talk about ROC curves and AUC's (the area under the wavy line or the "Area Under the Curve"). The dotted line produced by random guessing has exactly half of all possibilities in the universe of Hit Rates and False Positive Rates underneath it; the AUC is 0.5. The greater the AUC, the greater the accuracy of our test. When studied in real life samples, the best actuarials for the prediction of sex offense recidivism tend to produce AUC's of between .60 and .75.



Note that as discussed above we can increase or decrease our Hit Rate, the percentage of SD men whom we "capture," by changing our cutoff, e.g., by lowering the threshold for opining SD and opining SD 75% or 100% of the time. We can thus easily increase our Hit Rate, but only at the expense of also increasing our False Positive rate (the rate at which we conclude SD for offenders who are not actually SD).

Based on the chart above, we see that in real life, using one of our best predictive methods, at a Hit Rate of 50% (we use a Cutoff score, C-50, that "captures" 50% of the truly SD), 31% of those who are NOT SD will be identified as SD (False Positive rate of 31%). If we try to increase our Hit Rate by 50%—from a Hit Rate that captures half of the truly SD to a Hit Rate that captures three-quarters of the truly SD (by using a lower threshold or Cutoff score and thus opining SD more often)—then 62% of those who are NOT SD will be identified as SD (False Positive rate of 62%). In this real life example, we can only increase our Hit Rate 50% by doubling our False Positive rate.Using the information from the figure above, in the next chart, we begin to analyze the effect that using our Most Accurate Method has on the Positive

Predictive Value. Our Most Accurate Method at Cutoff C-50 identifies 50% of true SD's while yielding a false positive rate of 31%. This is well above chance and is, in fact, the accuracy of the Static-99 (ibid.). If 100 men were evaluated and if the true SD *Base Rate* were 50%, at Cutoff C-50—which we set to make the Hit Rate 50%—25 out of 50 true SD's would be accurately identified and 15.5 (on average) not-SD men would be "captured" in the SD conclusion (31% of 50 not SD men or 15.5 false positives).

Base Rate = 50% Cutoff set for a Hit Rate of 50%

	# Who Recidivated	**# Who Didn't Recidivate**	**Totals**
Static-99 at C-50 Indicates "SD"	**25 True Positive** (concluded SD and recidivism occurs following release)	~~**15.5 False Positive**~~ ~~(concluded SD but there is no recidivism)~~	**40.5 "SD" Indications**
Static-99 at C-50 Indicates "Not-SD"	~~**25 False Negative**~~ ~~(concluded not SD and recidivism occurs)~~	**34.5 True Negative** (concluded not SD and there is no recidivism)	**59.5 "Not-SD" Indications**
Totals	**50 Recidivators, i.e., Truly SD**	**50 Non-Recidivators, i.e., Not SD**	**Total Men Examined 100**

Thus when concluding SD, this method would be right 25/40.5 of the time or the Positive Predictive Value will be 62%. Can that meet the clear and convincing standard when evaluating a man for commitment? That would be up to the fact finders and the courts to decide.

But this example is not typical of what happens in SD proceedings. Our state experts and prosecutors appear to set the cutoff at a hit rate of around 75% or higher. With a 75% hit rate (identifying 75% of true SD's) the situation actually seen in court is more like this:

Our Most Accurate Method at Cutoff C-75 identifies 75% of true SD's while yielding a false positive rate of 62%. Again, this is well above chance and is, in fact, the accuracy of the Static-99 (ibid.).

Base Rate = 50% Cutoff set for a Hit Rate of 75%

	# Who Recidivated	# Who Didn't Recidivate	Totals
Static-99 at C-75 Indicates "SD"	**37.5 True Positive** (concluded SD and recidivism occurs following release)	~~**31 False Positive** (concluded SD but there is no recidivism)~~	**68.5 "SD" Indications**
Static-99 at C-75 Indicates "Not-SD"	~~12.5 **False Negative** (concluded not SD and recidivism occurs)~~	**19 True Negative** (concluded not SD and there is no recidivism)	**31.5 "Not-SD" Indications**
Totals	**50 Recidivators, i.e., Truly SD**	**50 Non-Recidivators, i.e., Not SD**	**Total Men Examined 100**

If 100 men are evaluated and the SD Base Rate is 50% then, at C-75, 37.5 out of 50 true SD's will be accurately identified (Hit Rate = 75%) and 31 not SD men will be captured in the same net (a False Positive rate of 62% of 50 not-SD men or 31 false positives). Thus when concluding SD, this method will be right 37.5/68.5 of the time or the Positive Predictive Value is 54%.

While this, too, would be up to the court to decide, it starts to be doubtful as to whether that could meet the clear and convincing standard when evaluating a man for commitment.

But this example---though somewhat more similar to what we actually see---is *still* not typical of what actually happens in SD proceedings. To reflect the reality of SD proceedings, we would have to consider what happens when we use a more realistic Base Rate, i.e., the Base Rate of all sex offenders who are in custody and who will at least be reviewed to determine if there should be a petition for commit. Since as we saw above, Base Rates influence our Positive Predictive Value, let's take a look at what the real life situation is when, instead of assuming a Base Rate of 50%, we use a more realistic estimate of the true Base Rate, an estimate near the higher end of the range (15 to 35%) of reasonable recidivism Base Rate estimates, say 30%.

Again, our Most Accurate Method at Cutoff C-50 would identify 50% of true SD's while yielding a false positive rate of 31%. And again, this is well above chance and is, in fact, the accuracy of the Static-99 (ibid.). If 100 men are evaluated and the SD Base Rate is 30% then 15 out of 30 SD's will be identified (Hit Rate = 50%) and 22 not SD men will be captured in the same net (31% of 70 not SD men or 22 False Positives).

Base Rate = 30%   Cutoff set for a Hit Rate of 50%

	# Who Recidivated	# Who Didn't Recidivate	Totals
Static-99 at C-50 Indicates "SD"	**15 True Positive** (concluded SD and recidivism occurs following release)	~~**22 False Positive** (concluded SD but there is no recidivism)~~	**37 "SD" Indications**
Static-99 at C-50 Indicates "Not-SD"	~~**15 False Negative** (concluded not SD and recidivism occurs)~~	**48 True Negative** (concluded not SD and there is no recidivism)	**63 "Not-SD" Indications**
Totals	**30 Recidivators, i.e., Truly SD**	**70 Non-Recidivators, i.e., Not SD**	**Total Men Examined 100**

Thus when concluding SD, this method will be right 15/37 of the time or the Positive Predictive Value will be 41%. An opinion of SD will be wrong more often than not or 59% of the time. An SD conclusion which is thus wrong 3 out of 5 times could not be a primary component of clear and convincing evidence for concluding that a specific man is SD.

However, no state expert in Massachusetts, for example, opines SD as low as 37 out of 100 cases. If we go with a 75% hit rate (identifying 75% of true SD's) the result is more in line with actual practice in which 50 to 80% of the time (depending on the expert) the state experts opine SD. This more realistic example of the true situation we see in court looks like this:

Our Most Accurate Method at Cutoff C-75 identifies 75% of true SD's while yielding a false positive rate of 62%. Again, this is well above chance and is, in fact, the accuracy of the Static-99 (ibid.). If 100 men are evaluated and the SD Base Rate is 30% then 22.5 out of 30 SD's will be identified (Hit Rate = 75%) and 43 not SD men will be swept up in the same net (62% of 70 not SD men or 43 False Positives).

Base Rate = 30% Cutoff set for a Hit Rate of 75%

	# Who Recidivated	# Who Didn't Recidivate	Totals
Static-99 at C-75 Indicates "SD"	**22.5 True Positive** (concluded SD and recidivism occurs following release)	~~**43 False Positive** (concluded SD but there is no recidivism)~~	**65.5 "SD" Indications**
Static-99 at C-75 Indicates "Not-SD"	~~**7.5 False Negative** (concluded not SD and recidivism occurs)~~	**27 True Negative** (concluded not SD and there is no recidivism)	**34.5 "Not-SD" Indications**
Totals	**30 Recidivators, i.e., Truly SD**	**70 Non-Recidivators, i.e., Not SD**	**Total Men Examined 100**

Thus when concluding SD, this method will be right 22.5/65.5 of the time or the Positive Predictive Value will be 34%. In this illustration of the use of actuarials in real SD hearings, an actuarial-based SD opinion will be wrong twice as often as it is correct or 66% of the time. An SD conclusion which is thus wrong 2/3 of the time could never meet the clear and convincing standard when evaluating a specific man for commitment.

Thus, given our best estimate[34] of the true Base Rate of sex offense recidivism of significantly below 50%, and using our best predictive methods in which opinions of SD are accurate less than 70% of the time, expert opinions of SD using such a method could never provide a basis for the clear and convincing evidence needed for commitment. The offender would need to have features not measured by our best method that clearly (and convincingly) ups the risk estimate.[35]

Expert witness testimony can NOT be an indication of such features because expert witness testimony by state experts almost ALWAYS ups the risk to SD whenever an offender

[34] By "best estimate," I mean an estimate that lies within the reasonable range of Base Rate estimates and is at the end of that range (the high end) that gives the State the benefit of the doubt. If using such a Base Rate makes actuarial-based SD conclusions incapable of producing clear and convincing evidence to support an SD finding, we can be fairly certain that current actuarial-based methods can never provide a basis for a finding of SD.

[35] The latter is what is typically claimed by state experts. Other attachments provide the evidence that the adjustments to the best method predictions (actuarial or mechanical prediction) are, in practice, almost never improvements in accuracy. While truly unbiased, skilled clinicians might, in a small number of cases, be able to improve accuracy by cautious use of additional information, such an ability has never been demonstrated.

falls in the highest risk category using our Most Accurate Method. The only thing the expert *actually* adds is his *ipse dixit*.[36]

Therefore, falling in the highest risk category using actuarial assessment and having a state expert say the offender is exceptional and truly high risk are almost always one and the same thing. In such a state of affairs, the state expert's opinion about features that can up the probability estimate could never provide the truly exceptional information beyond actuarial-based prediction that would be necessary to reach a clear and convincing standard.

[36] We know this to be the case because if state experts almost always opine SD (which they do) when, for example, the Static-99 score falls in Hanson's "highest risk" category (Static-99 score of 6+), then the accuracy of the test in that range would have to be much higher (the wavy curve would have to move further away from the dotted line in the ROC figure) than it is (see the ROC figure which shows no such tendency).

Release from Commitment in Massachusetts (Exhibit 11)

Under the Jimmy Ryce Civil Commitment Program, the only way for a person adjudicated "sexually dangerous" (SD) to be discharged occurs

> When the Director of the facility in which a person is placed pursuant to subsection (d) determines that the person's condition is such that he is no longer sexually dangerous to others, or will not be sexually dangerous to others if released under a prescribed regimen of medical, psychiatric, or psychological care or treatment, he shall promptly file a certificate to that effect with the clerk of the court that ordered the commitment.

Since the Program aims to place SD men in state facilities when these are available, presumably SD men in Massachusetts (and possibly in other New England states) will be committed to the Massachusetts Treatment Center for Sexually Dangerous Persons. But whether they are placed there or some other facility, we can analyze, hypothetically, the burdens of such placement. In this scenario, the “Director of the facility in which a person is placed” would then be the Director of the Massachusetts Treatment Center for Sexually Dangerous Persons. So, the Program would rely on said Director to initiate release proceedings for men who are no longer sexually dangerous. We have data in Massachusetts (and the pattern is almost identical in other states) that suggests how this would actually operate.

How Directors of sex offender treatment facilities have operated

In Massachusetts, more than fifty (less than 100, exact number unknown) men who were deemed SD have been found to be no longer SD during the time when the "treatment" facility has been run by the Department of Corrections (DOC). In over 98% of those cases, the State has opposed their release, and in no case has the DOC petitioned for their release.

That means that there have been *at least* fifty occasions when men who were no longer sexually dangerous remained committed to the facility. Though the Massachusetts SDP law, Section 9, states that "The department of correction may file a petition at any time if it believes a person is no longer a sexually dangerous person," this has never happened and it is almost certain that it never will happen. Indeed, the DOC has a track record of almost always opposing the release of men who were, in fact (as determined by the courts in the subsequent hearing), no longer SD.

Furthermore, we have accurate figures about the average length of treatment in the Massachusetts SOTP *according to the clinical experts who oversee that treatment program*. In Massachusetts, the Community Access Board (CAB) reviews the treatment and SD status of each committed man annually and determines if they remain SD. Year after year, they find that the men remain SD well over 99% of the time. That’s an astounding figure for any treatment program of any treatable disorder of any type. It represents a “cure” or a “sufficient reduction in risk due to treatment" rate of well under 1% per year. At that rate, we can estimate the *average* length of "treatment" to be well over 50 years.[37]

In Massachusetts, SD men are entitled to petition the court annually for a hearing to review their SD status and for release if they are found to no longer be SD. In actual practice, they get a hearing about once every three years. Because of this, the actual length of commitments on average is between 15 and 25 years[38] with almost all (close to 100%) of the releases being actively opposed by "the Director of the facility" (or his designee). While the actual length of treatment is therefore much lower than 50 plus years, this shorter length of treatment is due to death in prison and releases by the judicial system that are almost always opposed by "the Director of the facility." In regard to other state programs, the situation is almost identical or worse; in many states the official evaluation teams have never found a man to be no longer SD.

Indeed, it would be hard to imagine a Director of a facility or department that was appointed by an elected official who would ever recommend a "sexually dangerous" offender for release. If one such recommended for release offender went out and committed a new sexual offense, it would mean the abrupt end of the career of the politician under whose watch the Director recommended the offender for release.

Furthermore, under the Program, the facility Director would have to be willing to recommend the release of an "uncured" sexually dangerous offender. The following is a description by the Department of Correction of their own SOTP

> **Sex Offender Treatment Program**
> The Sex Offender Treatment Program is a comprehensive treatment program for all inmates identified as sex offenders including: those committed for a sex offense, those with any history of a sex offense conviction, and/or those individuals with sexual overtones in the reading of their official version. The major treatment components of the program include: modifying interpersonal relations; exploration of the roots of the problem; enhancing coping skills; building empathy; identifying cognitive distortions; modifying deviant arousal; developing a relapse prevention plan; transition to the community; and maintaining recovery. *This treatment program is based on the concept that sex*

[37] A "cure" rate of 1% per year means that in any given year there is a 1% chance that a randomly selected (average) individual will be cured. If one looks at a population of 100 individuals suffering from drug resistant tuberculosis, for example, such a cure rate means that one will be deemed cured in any given year. Of course, this is an average figure and in one year there may be no cures and in another there may be several. So some may be tuberculosis free in the first few years, while many others will die of old age with tuberculosis even if the tuberculosis doesn't kill them. On average at that rate of cure, it would take 100 years to cure the entire population of 100. The "time to cure" for the *average* offender would then be fifty years. And this is based on a cure rate of 1%. In actual practice, the official "cure rate" for SD across the nation appears to be *much lower* than 1%, which means the average time-to-cure could well be over 75 years based on the evaluations by those who are running these programs.

[38] The actual average is unknown. This estimate is based on my personal experience of over thirty years working at or evaluating men at the Massachusetts Treatment Center. Rarely has a man been released after being committed for less than 10 years and most men are released before 30 years have passed. A larger number of men are never released and die at the Treatment Center than those released in less than 15 years. So, an average of between 15 and 25 years is fairly likely to be accurate.

> *offenders cannot be cured, but through a lifetime commitment to treatment and supervision, the chance of a relapse will be greatly reduced.* (Emphasis added. From http://www.state.ma.us/doc/PROGRAMS/deptsrvs.htm, the Department of Correction's web site. 2/22/04)

In today's political climate, is it really possible to imagine an elected official allowing one of his subordinates (the Director of the facility) to recommend release into the community of an "uncured" sex offender, regardless of the supervision available?[39]

[39] Since the Adam Walsh Act aims to place those deemed SD under the care of state-run programs, in almost all cases, the Director of the facility would be appointed by an elected official or his/her designee. It is just not reasonable to expect that, in today's political climate, the subordinates of elected officials would petition for the discharge of a man who had been deemed "Sexually Dangerous," even when there is no longer evidence that a person is SD. This explains why the official boards evaluating SD men rarely, if ever, (and in some states, never) recommend release.

BIBLIOGRAPHY (Exhibit 12)

Andrews, D. A., & Bonta, J. (1998). *The Psychology of Criminal Conduct (2nd ed.)*. Cincinnati, OH: Anderson.

Barbaree, H. E., Blanchard, R., & Langton, C. M. (2003). The development of sexual aggression through the life span: The effect of age on sexual arousal and recidivism among sex offenders. In R.A. Prentky, E. S. Janus, & M. C. Seto (Eds.), *Sexually coercive behavior: Understanding and Management* (pp. 59-71). New York: Annals of the New York Academy of Sciences (Vol. 989).

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

Berlin, F. S., Balbreath, N. W., Geary, B. McGlone, G. (2003). The use of actuarials at civil commitment hearings to predict the likelihood of future sexual violence. *Sex Abuse: A Journal of Research and Treatment*, 15, 4, 377-382.

Blanchard, R., & Barbaree, H. E. (2005). The Strength of Sexual Arousal as a Function of the Age of the Sex Offender: Comparisons Among Pedophiles, Hebephiles, and Teleiophiles. *Sexual Abuse: A Journal of Research and Treatment*, 441-456.

Bonta, J., Law, M., & Hanson, K (1998). The Prediction of Criminal and Violent Recidivism Among Mentally Disordered Offenders: A Meta-Analysis. *Psychological Bulletin*, 123, 123-142.

Borum, R. (1996). Improving the clinical practice of violence risk assessment. *American Psychologist, 51,* 945—956.

Chumlea, W. C., Schubert, C. M., Roche, A. F., Kulin, H. E., Lee, P. A., Himes, J. H., & Sun, S. S. (2003). Age at menarche and racial comparisons in US girls. *Pediatrics*. Jan;111(1):110-3.

Cohen, M. L., Groth, A. N., & Siegel, R. (1978). The clinical prediction of dangerousness. *Crime and Delinquency*, January, 28 - 39.

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

Dawes, R. M., Faust, D., & Meehl, P. E. (1993). Statistical prediction versus clinical prediction: Improving what works. In G. Keren, & C. Lewis (Eds.), *A Handbook for Data Analysis in the*

Behavioral Sciences: Methodological Issues (pp. 351-367). Hillsdale, NJ: Lawrence Erlbaum.

DeBruin, D. W., Stevens, T. N., & Gilfoyle, N. F. P. (2005). Brief of *Amicus Curiae* American Psychological Association in support of defendant-appellant (in Case No. 04-50393, U.S. Court of Appeals for the Fifth Circuit, U.S.A. v. Sherman Lamont Fields).

Epperson, D.L., Kaul, J.D., & Huot, S.J. (1995). Predicting risk of recidivism for incarcerated sex offenders: Updated development on the Sex Offender Screening Tool (SOST). Paper presented at the 14th Annual Conference of the Association for the Treatment of Sexual Abusers, New Orleans, LA.

Epperson, D.L., Kaul, J.D., & Hesselton, D. (1999). Minnesota Sex Offender Screening Tool–Revised (MnSOST-R): Development, Performance, and Recommended Risk Level Cut Scores. Minnesota Department of Corrections.

Faust, D., & Ziskin, J. (1988). The expert witness in psychology and psychiatry. *Science*, 241, 31-35.

Goldberg, L. (1959). The effectiveness of clinician's judgment: Diagnosis of organic brain damage from the Bender Gestalt test. *Journal of Consulting Psychology, 23*, 25-33.

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

Gottfredson, S., & Gottfredson, D. (1994). Behavioral prediction and the problem of incapacitation. *Criminology*, 32(3), 441-474.

Greenfeld, L. A. (1997). Sex Offenses and Offenders. *Bureau of Justice Statistics*, U.S. Department of Justice.

Grisso, T. (2000). Ethical Issues in Evaluations for Sex Offender Re-offending. Paper presented at Symposium on Sex Offender Re-offense Risk Prediction, Madison, WI (March 6).

Grove, W.M., & Meehl, P.E. (1996). Comparative efficiency of formal (mechanical, algorithmic) and informal (subjective, impressionistic) prediction procedures: The clinical/statistical controversy. *Psychology, Public Policy & Law*, 2, 293-323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12, 1, 19-30.
Grubin, D. (1999). Actuarial and clinical assessment of risk in sex offenders. *Journal of Interpersonal Violence*, 14 (3), 331-343.

Hall, G. C. N. (1995). Sexual offender recdividism revisited: A meta-analysis of recent treatment studies. *Journal of Consulting and Clinical Psychology*, 63, 5, 802-809.

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

Hanson, R. K. (2000). Using Static-99 to estimate risk for offenders who have remained offense free in the community. Office of the Solicitor General, Canada.

Hanson, R. K. (2002). Recidivism and age: Follow-up data on 4,673 sexual offenders. *Journal of Interpersonal Violence,* 17, 1046-1062.

Hanson, R. K. (2006). Does Static-99 Predict Recidivism Among Older Sexual Offenders? *Sex Abuse: A Journal of Research and Treatment*, 18, 343-355.

Hanson, R. K., Bussiere, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism. *Journal of Consulting and Clinical Psychology*, 66, 348-362.

Hanson, R. K., Gordon, A., Harris, A. J. R., Marques, J. K., Murphy, W., Quinsey, V. L., and Seto, M. C. (2002). First report of the collaborative outcome data project on the effectiveness of psychological treatment for sex offenders. *Sexual Abuse: A Journal of Research and Treatment*, 14, 2, 169-194.

Hanson, R. K., & Harris, A. (1998). Dynamic Predictors of Sexual Recidivism. Office of the Solicitor General, Canada.

Hanson, R. K., & Harris, A. (2000). The Sex Offender Need Assessment Rating (SONAR): A method for measuring change in risk levels. Office of the Solicitor General, Canada.

Hanson, R. K., & Morton-Bourgon, K. (2005). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology*, 66, 348-362.

Hanson, R. K., & Morton-Bourgon, K. (2007). The Accuracy of Recidivism Risk Assessments for Sexual Offenders: A Meta-Analysis. Public Safety and Emergency Preparedness Canada.

Hanson, R. K., & Thornton, D. (1999). *Static 99: Improving the predictive accuracy of actuarial risk assessments for sex offenders*. Ottawa: Public Works and Government Services Canada.

Harris, A., Phenix, A., Hanson, R. K., & Thornton, D. (2003). STATIC-99 Coding Rules Revised - 2003. Office of the Solicitor General, Canada.

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

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

Hart, S. D. (2003). Actuarial risk assessment: Commentary on Berlin et al. *Sex Abuse: A Journal of Research and Treatment*, 15, 4, 377-382.

Howe, E. (1994). Judged person dangerousness as weighted averaging. *Journal of Applied Social Psychology*, 24 (14), 1270-1290.

Howells, K. (1981). Adult sexual interest in children: Considerations relevant to theories of aetiology. In Cook, M. & Howells, K. (eds.), *Adult Sexual Interest in Children*, London: Academic Press, 55-94.

Jackson, R. L., Rogers, R., & Shuman, D.W. (2004). The Adequacy and Accuracy of Sexually Violent Predator Evaluations: Contextualized Risk Assessment in Clinical Practice.

International Journal of Forensic Mental Health, Vol. 3, No. 2, pages 115-129
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. A. (2003). Forensic Use of Actuarial Risk Assessment with Sex Offenders: Accuracy, Admissibility and Accountability. *American Criminal Law Review, 40,* 1143-1489.

Kozol, H. L., Boucher, R. J., & Garafolo, R F.. (1972). The diagnosis and treatment of dangerousness. *Crime and Delinquency*, October, 371-392.

Kriegman, D. (2006). The reduction of sexual offense recidivism following commitment and psychodynamic treatment: A challenge to the dominant cognitive-behavioral model. *The Journal of Sexual Offender Civil Commitment: Science and the Law, 1*, 90-98.

Langan, P. A., Levin, D. J. (2002). Recidivism of prisoners released in 1994. *Bureau of Justice Statistics*, U.S. Department of Justice.

Langan, P. A., Schmitt, E. L., & Durose, M. R. (2003). Recidivism of sex offenders released from prison in 1994. *Bureau of Justice Statistics*, U.S. Department of Justice.

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

Meehl, P. E. (1954). *Clinical versus statistical prediction.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment*, *50*, 370—375.

Meehl, P. E. (l996, August). *Credentialed persons, credentialed knowledge.* Paper presented at the 104th Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

Milner, J., & Campbell, J. (1995). Prediction issues for practitioners. In J. C. Campbell (Ed.), *Assessing dangerousness: Violence by sexual offenders, batterers, and child abusers* (pp. 20-40). Thousand Oaks, California: Sage Publications, Inc.

Menzies, R. J., Webster, C. D., & Sepejak, D. S. (1985). Hitting the forensic sound barrier: Predictions of dangerousness in a pre-trial clinic. In C. D. Webster, M. H. Ben-Aron, & S. J. Hucker (Eds.), *Dangerousness: Probability and prediction, psychiatric and public policy* (pp. 115—143). New York: Cambridge University Press.

Menzies, R., Webster, C. D., McMain, S., Staley, S., & Scaglione, R. (1994). The dimensions of dangerousness revisited: Assessing forensic predictions about violence. *Law and Human Behavior, 18* (1), 1-28.

Monahan, J. (1981). *Predicting violent behavior: An assessment of clinical techniques.* Beverly Hills, CA: Sage.

Monahan, J. (1984). The prediction of violent behaviour: Toward a second generation of theory and policy. *American Journal of Psychiatry*, 141(1), 10-15.

Monahan, J. (1992). Mental disorder and violent behavior. *American Psychologist, 47, 511—521.*

Monahan, J. (1995). Review of The Violence Prediction Scheme: Assessing Dangerousness in High-risk Men, by C. D. Webster, G. T. Harris, M. E. Rice, C. Cormier, & V. L. Quinsey. *Criminal Justice and Behavior*, 22, 446-447.

Monahan, J. (1996). Violence prediction: the past twenty and the next twenty years. *Criminal Justice and Behavior*, *23*, 107-120.

Monahan, J. (Chair), Feshbach, S., Holder, W., Howe, R. A., Kittrie, N., Loevinger, J., McDonough, L., Messinger, S., Repucci, N. D., Schoen, K., Tapp, J. L., & Wasserstrom, R. (1978). Report of the Task Force on the role of psychology in the criminal justice system. *American Psychologist*, 33, 12, 1099-1113.

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

Okami, P. & Goldberg, A. 1992. Personality Correlates of Pedophilia: Are They Reliable Indicators? *Journal of Sex Research, Vol. 29,* No. 3, 297-328.

Prentky, R., Harris, B., Frizzell, K., & Righthand, S. (2000). An actuarial procedure for assessing risk with juvenile sex offenders. *Sexual Abuse: A Journal of Research and Treatment, 12*, 71-93.
Prentky, R., Janus, E. S., Barbaree, H. E., Schwartz, B. K., & Kafka, M. P. (2006). Sexually violent predators in the courtroom: Science on trial. *Psychology, Public Policy, and Law*, 12, 4, 357–393.

Prentky, R. A., Knight, R. A., & Lee, A. F. (1997). Risk factors associated with recidivism among extrafamilial child molesters. *Journal of Consulting and Clinical Psychology*, 65(1), 141–139.

Prentky, R. A., & Lee, A. F. (2007). Effect of Age-at-Release on Long Term Sexual Re-offense Rates in Civilly Committed Sexual Offenders. *Sexual Abuse: A Journal of Research and Treatment*, 19, 43-59.

Prentky, R. A., Lee, A. F., Knight, R. A., & Cerce, D. (1997). Recidivism rates among child molesters and rapists: A methodological analysis. *Law and Human Behavior*, 21, 6, 635-659.

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

Quinsey, V. L., Maguire, A. (1986). Maximum Security Psychiatric Patients: Actuarial and Clinical Prediction of Dangerousness. *Journal of Interpersonal Violence*, 1, 2, 143-171.

Quinsey, V. L., Rice, M. E., & Harris, G. T. (1995). The actuarial prediction of sexual recidivism. *Journal of lnterpersonal Violence,* 10, 85—105.

Rice, M. E. (1997). Violent offender research and implications for the criminal justice system. *American Psychologist*, 52, 414-423.

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

Thorne, F. C. (1972). Clinical Judgment. In R. H. Woody, *Clinical Assessment in Counseling and Psychotherapy*. Englewood Cliffs, NJ: Prentice Hall.

Thornton, D. (2006). Age and sexual recidivism: A variable connection. *Sexual Abuse: A Journal of Research and Treatment*, 18(2), 123–135.

Webster, C. D., Harris, 0. T., Rice, M. E., Cormier, C., & Quinsey, V. L. (1994). *The Violence Prediction Scheme*. Toronto, Ontario, Canada:University of Toronto, Centre of Criminology.

Wollert, R. (2006). Low base rates limit expert certainty when actuarials are used to identify sexually violent: An application of Bayes's Theorem. *Psychology, Public Policy, and Law*, 12, 56-85.

Zonana, H., Abel, G., Bradford, J., Hoge, S. K., & Metzner, J. (1998). *APA Task Force Report On Sexually Dangerous Offenders*. American Psychiatric Association.

CURRICULUM VITAE

DANIEL KRIEGMAN, PH.D.
20 Dorcar Road
Newton, MA 02467-3021
(617) 527-1631
kriegman@aol.com

EDUCATION

B.A. State University of New York at Buffalo, 1974
Major: Psychology
Honors: Departmental and University Honors,
Magna Cum Laude

M.A. Boston University, 1977

Ph.D. Boston University, 1980
Area: Clinical Psychology
NIMH Fellowship: 1974-75, 1975-76
Teaching Fellowship: 1974-75

CLINICAL and ADMINISTRATIVE EXPERIENCE

1981 - Private Practice, Cambridge and Newton, MA. *Individual, couple, family, and group psychotherapy, clinical supervision, evaluations of Sexual Dangerousness, expert witness evaluations, and testimony for the courts (risk assessment, aid-to-sentencing, treatment planning, custody evaluations).*

1989 - 1995 Clinical Director, Counseling and Psychotherapy Center of Greater Boston (a private, group practice).

1987 - 1993 Founder & President, Human Services Cooperative, Inc. (a worker-owned, human service provider operating through contracts with state agencies, third party payers, and client fees). *Oversaw the operation and rapid growth (beginning in fiscal year 1990, revenues exceeded 1.2 million dollars) of a human service provider agency. Contracts included the Crisis Services Program at the Solomon Mental Health Center in Lowell (staff of 18 highly experienced masters and doctoral level clinicians and psychiatrists providing the first line of contact [including containment and stabilization, if necessary] and referral for mental health clients in the Lowell catchment area) and all of the psychological and psychiatric services (e.g., court evaluations and psychopharmacology) at the Massachusetts Treatment Center for the Sexually Dangerous Offender.*

1986 - 1987 Director of Clinical Services, Counseling Services of Massachusetts Mentor, Inc. (a licensed mental health clinic with three sites). *Responsible for selection of all staff, establishment of job descriptions and assignments, overall supervision of staff, policies and procedures, program evaluation, in-service training for professional staff, and budget management.*

1977 - 1986 Chief Psychologist, Massachusetts Treatment Center for the Sexually Dangerous Offender (a two hundred fifty bed, long-term, secure inpatient facility). *Began in 1977 as a Staff Psychologist in a part-time position consisting of intensive individual and group psychotherapy, intake evaluations, diagnostic testing, and evaluations for community access. Later held the position of Unit Director for approximately 130 committed patients, responsible for all clinical services on the unit. Most recent position included two separate areas of responsibility: Director of Supervision and Training for the direct service staff (over 50 non-custodial, clinical, DMH employees); Director of Intake and Treatment Planning for all patients sent to the Treatment Center for evaluation for commitment (Intake), and those who were newly committed (Treatment Planning).*

1981 - 1983 Psychologist, North Shore Professional Associates, Inc., Beverly, MA; (a private group practice). *Part-time position consisting of individual, couples, and family treatment, psychodiagnostic testing, psychotherapy supervision, and custody evaluations for the courts.*

1981 - 1982 Family Therapist, Family Continuity Program, Beverly, MA (an outpatient alternative to hospitalization that provides a variety of intensive services to the family of an “identified” adolescent patient). *Part-time position consisting of family therapy (often provided in the home of a "dysfunctional" family), individual and group psychotherapy, and clinical supervision.*

1979 - 1981 Clinical Director, Adolescent Rehabilitation Program at the Dr. Solomon Carter Fuller Mental Health Center (an intensive, highly-funded, long-term, secure, inpatient unit for severely disturbed, acting out adolescents). *Full-time position supervising 40 clinical, educational, and milieu staff responsible for the overall treatment, education, and management of 12 adolescents, and expert witness evaluations and testimony for the courts (commitment hearings).*

RESEARCH and TEACHING

2005 - Editorial Board, *The Journal of Sexual Offender Civil Commitment: Science and the Law*. A peer-reviewed journal publishing articles of original research, review, and commentary on issues related to the civil commitment of sexual offenders, including assessment and treatment of sexual offenders.

1998 - Faculty, Psychoanalytic Couple and Family Institute of New England

1993 - Faculty, Massachusetts Institute for Psychoanalysis

1981 - 1988 Seminars at the Northeast Society for Group Psychotherapy and seminars and courses at the Boston Institute for Psychotherapy.

1974 - 1980 Doctoral Thesis, Boston University: A psychosocial study of religious cults from the perspective of self psychology.

1977 - 1979 Instructor: Boston University, Lesley College, and Massasoit College. Seven courses in these areas: General Psychology, Introduction to Abnormal Psychology, and Personality Theory.

1977 - 1978 Psychological Testing Supervisor, Boston University Graduate School: Psychodiagnostic Testing.

1974 - 1975 Teaching Fellow, Boston University: Personality Theory, Abnormal Psychology.

1973 - 1974 Teaching Assistant, SUNY at Buffalo: Research Methods.

PUBLICATIONS

Kriegman, D. & Biederman, I. (1980). How many letters in Bidwell's ghost? An investigation of the upper limits of full report from a brief visual stimulus. *Journal of Perception and Psychophysics*, 28, 82-84. Featured in *Scientific American*, 252, 2, 126-127, 1985.

Kriegman, D. & Solomon, L. (1985a). Cult groups and the narcissistic personality: The offer to heal defects in the self. *International Journal of Group Psychotherapy,* 35, 2, 239-261.

Kriegman, D. & Solomon, L. (1985b). Psychotherapy and the "new religions": Are they the same? *Cultic Studies Journal*, 2, 1, 2-16.

Kriegman, D. (1986). The treatment of the sexually dangerous offender: a multi-modal approach utilizing the perspective of self psychology. Paper presented at The Massachusetts Treatment Center for Sexually Dangerous Persons.

Kriegman, D. (1988). Self psychology from the perspective of evolutionary biology: Toward a biological foundation for self psychology. In A. Goldberg (Ed.), *Progress in Self Psychology* (Vol. 3, pp. 253-274). Hillsdale, New Jersey: The Analytic Press.

Slavin, M. O. & Kriegman, D. (1988). Freud, biology, and sociobiology. *American Psychologist*, 43, 658-661.

Kriegman, D. & Knight, C. (1988). Social evolution, psychoanalysis, and human nature. *Social Policy*, 19, 2, 49-55.

Kriegman, D. & Slavin, M. O. (1989). The myth of the repetition compulsion and the negative therapeutic reaction: An evolutionary biological analysis. In A. Goldberg (Ed.), *Progress in Self Psychology,* (Vol. 5, pp. 209-253). Hillsdale, NJ: Analytic Press.

Slavin, M. O. & Kriegman, D. (1990). Toward a new paradigm for psychoanalysis: An evolutionary biological perspective on the classical-relational dialectic. *Psychoanalytic Psychology,* 7, 5-31.

Kriegman, D. & Slavin, M. O. (1990). On the resistance to self psychology: Clues from evolutionary biology. In A. Goldberg (Ed.), *Progress in Self Psychology,* (Vol. 6, pp. 217-250). Hillsdale, NJ: Analytic Press.

Kriegman, D. (1990). Compassion and altruism in psychoanalytic theory: An evolutionary analysis of self psychology. *Journal of the American Academy of Psychoanalysis*, 18, 2, 342-367.

Slavin, M. O. & Kriegman, D. (1992). Psychoanalysis as a Darwinian depth psychology: Evolutionary biology and the classical-relational dialectic in psychoanalytic theory. In J. Barron, M. Eagle, and D. Wolitzky (Eds.), *The Interface of Psychoanalysis and Psychology* (pp. 37-76). Washington, DC: American Psychological Association.

Slavin, M. O. & Kriegman, D. (1992). *The Adaptive Design of the Human Psyche: Psychoanalysis, Evolutionary Biology, and the Therapeutic Process*. NY: Guilford Press.

Kriegman, D. (1996). On the existential/subjectivism-scientific/objectivism dialectic in self psychology: A view from evolutionary biology. In A. Goldberg (Ed.), *Progress in Self Psychology* (Vol. 12, pp. 85-119). Hillsdale, NJ: Analytic Press.

Kriegman, D. (1996). The effectiveness of medication: The *Consumer Reports* study. *American Psychologist*, 51, 10, 881.

Kriegman, D. and Kriegman, O. (1997, in preparation). War and the evolution of the human propensity to form nations, cults, and religions. Paper presented at the Human Behavior and Evolution Society Annual Conference (June 7, Tucson, AZ) and at the Association for Politics and the Life Sciences Annual Conference (September 3, 1998, Boston, MA)

Kriegman, D. (1998). Interpretation, the unconscious, and psychoanalytic authority: Toward an evolutionary, biological integration of the empirical/scientific method with the field-defining, empathic stance. In R.F. Bornstein & J.M. Masling (Eds.), *Empirical Perspectives on the Psychoanalytic Unconscious* (pp. 187-272). Washington, DC: American Psychological Association.

Slavin, M. O. & Kriegman, D. (1998). An evolutionary biological perspective on psychoanalysis. In Robert Langs, (Ed.), *Theories in Psychoanalysis* (pp. 255-296). NY: International Universities Press.

Slavin, M. O. & Kriegman, D. (1998). Why the analyst needs to change: Toward a theory of conflict, negotiation, and mutual influence in the therapeutic process. *Psychoanalytic Dialogues*, 8, 2, 247-284.

Slavin, M. O. & Kriegman, D. (1998). Bigger than both of us: Double binds, conflicting interests, and the inherent paradoxes of human relatedness. *Psychoanalytic Dialogues*, 8, 2, 317-327.

Slavin, M. O. & Kriegman, D. (1998). Paradox and conflict, meta-communication and negotiation in psychoanalysis: Response to Dr. Ringstrom's discussion. *Psychoanalytic Dialogues*, 8, 2, 293-296.

Slavin, M. O. & Kriegman, D. (1998). Conflicting interests and the creation of a third space. *Psychoanalytic Dialogues*, 8, 3.

Kriegman, D. (1998). Evolutionary psychoanalysis: An advance in understanding the human psyche or a phylogenetic fantasy. *Contemporary Psychology*, 43, 2, 138-139.

Kriegman, D. (1998). Of quantum leaps and oxymorons: A reply to Langs. *Contemporary Psychology.*

Teicholz, J. G. & Kriegman, D. (Eds.). (1998). *Trauma, Repetition, & Affect Regulation: The Work of Paul Russell.* New York: The Other Press.

Kriegman, D. (1999). Trauma, conflict, and countertransference: A discussion of Peter Thomson's paper. *Canadian Journal of Psychoanalysis*, 7, 1, 59-62.

Kriegman, D. (1999). Parental investment, sexual selection, and evolved mating strategies: Implications for psychoanalysis. *Psychoanalytic Psychology*, 16, 4, 1-26.

Kriegman, D. (2000). Evolutionary psychoanalysis: Toward an adaptive, biological perspective on the clinical process in psychoanalytic psychotherapy. In P. Gilbert and K. Bailey (Eds.), *Genes on the Couch: Explorations in Evolutionary Psychology* (pp. 71-92). East Sussex, England: Psychology Press.

Kriegman, D. (2002). Interpreting & Negotiating Conflicts of Interests in the Analytic Relationship. In A. Goldberg (Ed.), *Progress in Self Psychology* (Vol. 18, pp. 87-112). Hillsdale, NJ: Analytic Press.

Kriegman, D. (2006). The reduction of sexual offense recidivism following commitment and psychodynamic treatment: A challenge to the dominant cognitive-behavioral model. *The Journal of Sexual Offender Civil Commitment: Science and the Law*, 1, 90-98.

Kriegman, D. (2006, in press). Conflict in the Analytic Relationship: The Psychoanalytic Treatment of a "Schizophrenic" (without drugs). *Journal of the American Psychoanalytic Association*. (Accepted for publication pending revisions.)

MEMBERSHIPS, LICENSURE, and CERTIFICATIONS

American Psychological Association
Massachusetts Institute for Psychoanalysis
Division of Psychoanalysis of the American Psychological Association (Division 39)
Founding Board Member and Faculty, The Psychoanalytic Couple & Family Institute of New England
Human Behavior and Evolution Society
Licensed Psychologist, Massachusetts (since 1981)
Health Service Provider under M.G.L. c. 112, § 120
Qualified Psychologist for Section 12 commitments pursuant to 104 CMR 3.20 (2)
under M.G.L. c. 123 § 12A
Qualified Examiner for evaluations of "Sexual Dangerousness"
under M.G.L. c. 123A § 1 (designated in 1986 by the Department of Mental Health)