Prediction Issues and the Role of Incremental Validity in Sex Offender Risk Assessment

Martin D. Lloyd

University of Minnesota
Abstract

We review issues of prediction and discuss their relevance to the prediction of sexual offender recidivism. First, we briefly discuss the different types of sex offenses, rates of these offenses, and rates of recidivism among sex offenders. We then discuss key issues in the field of prediction, with special attention given to the relative advantages and disadvantages of clinical and statistical prediction, as the general superiority of the latter approach shown in the literature. In our discussion of sex offender recidivism, we review known predictors of recidivism. We further discuss the various approaches to predicting sexual recidivism, including the actuarial instruments that have been developed for this purpose. Finally, we discuss incremental validity methodology and its potential utility for understanding and improving upon risk assessments as they are actually used in applied settings.
Sexual offending is a problem that has been receiving increasing attention, both on a societal level and in academic circles. This attention is not surprising, given the high costs of sexual offending. To the victim, the experience can involve severe emotional trauma, which may lead to difficulties in daily functioning and necessitate therapeutic care; there may also be physical injury. The victims, families, third-party payors and often governmental bodies must shoulder the financial costs of caring for victims. Society bears the costs of incarcerating or otherwise controlling the perpetrators of these crimes to lessen or eliminate future harm to the victim or to others. It is this goal of preventing future serious harms that requires identifying the most dangerous offenders, so that resources can be most efficiently allocated to prevent future heinous offenses. Hence, considerable effort has been devoted to developing ways of determining which offenders pose the greatest risk (Harris, Rice, Quinsey, Lalumiere, Boer, & Lang, 2003). This paper will review methods developed to predict risk of sexual reoffending, variously defined to include what are traditionally considered violent acts (e.g. forcible rape) as well as some acts (such as sexual offenses against children) that may not involve physical force, as well as means of evaluating the performance of these tools and ways of developing new, improved assessments.

We will first examine the background of these issues by reviewing evidence on the prevalence of sexual offenses and the legal context that mandates assessments to predict the risk of offender recidivism. Second, we will examine issues generally involved in behavioral prediction, to determine what this broad literature can tell us is likely to be true the specific problem of predicting sex offender recidivism. Third, we will review various schemes developed for predicting sex offense recidivism. Finally, we will examine the critical issue of incremental
validity in assessment, as it places important limits on what can be accomplished with an
assessment instrument possessing a given level of validity, whether it is used to predict a general
type of behavior or a specific behavior pattern such as sexual reoffending.

Background of the Problem

The seriousness of the sexual offending problem is undeniable, given the devastating
effects on its victims and its cost to society. However, the public may overestimate the actual
prevalence of sexual offending in society today. According to the Uniform Crime Report (UCR),
a compilation of arrest statistics gathered by the Federal Bureau of Investigation (FBI) and
published by the U.S. Department of Justice (DOJ), 95,136 acts of forcible rape (33 per 100,000
people in the U.S.), defined as any “carnal knowledge” of an unwilling female in which force is
used, were reported to the police in 2002 (DOJ, 2003). Also in 2002, another 95,066 arrests were
made for other sex offenses, which include statutory rape and those crimes termed public
indecency but do not include prostitution (DOJ, 2003). These figures are useful in demonstrating
the wide distribution of sexual offending. However, it is well if anecdotally documented that
many victims of such crimes do not report them to the police (Russell, 1982). Hence, anonymous
victimization surveys may present a clearer picture of the phenomena. According to the National
Crime Victimization Survey (NCVS), 247,730 people in the U.S. reported being rape or sexual
assault victims in 2002, where rape was defined as forcible intercourse and sexual assault
included forcible sex acts not involving intercourse as well as non-forcible sex acts with minors
or vulnerable adults (DOJ, 2004). These victimizations included 90,390 rapes, 77,470 attempted
rapes, and 79,870 other sexual assaults. Looking specifically at sexual offenses against children,
among
against a minor, with about 20% victimizing a child age 12 or younger (Greenfeld, 1996). Of the state prison inmates convicted of sexual assault other than forcible rape, over two thirds had committed their offenses against minors and nearly half committing their offenses against children 12 and younger (Greenfeld, 1996). To look at the problem another way, of the approximately 1.4 million individuals incarcerated in the U.S., over 100,000 of them, or about 14% of the prison population, are incarcerated for sex offenses (DOJ, 2004). From these figures, we can see that sexual offending is by no means rare. However, it does not directly affect as much of the population as many another problem, such violent crimes, property crimes, or for that matter poverty. For example, in 2002, extrapolating from survey data, 247,730 individuals in the U.S. experienced having been the victims of rape or sexual assault, whereas over 4.5 million people experienced having been non-sexually assaulted; more than 17.5 million experienced property crimes (DOJ, 2004). The sources above are generally regarded as the best available for determining the prevalence of most crimes. Sexual violence, however, is arguably different from most other crimes in that it is thought to be substantially less likely to be reported to police. Hence, official figures on sex crimes are likely to be more extreme underestimates of the true rates than are reported rates for most crimes (Russell, 1982). While we do not know the true rates of the various types of sexual crimes, it must be considered a serious social problem. This is so even if its prevalence does not match the epidemic levels that many in society may believe that it has, owing to the extreme moral repugnance of many of these crimes, and the expectation of great emotional (if not always physical) harm to the victim in many cases of these crimes.

From the standpoint of this review, even more troubling than the general population rate of sexual offenses is the base rate of sexual recidivism, i.e., the rate at which convicted sexual offenders, once released from incarceration, commit one or more new sexual offenses. This latter
rate is critical to mental health professionals, because clinicians are often charged with predicting it. It seems mental health professionals are seldom if ever asked to predict initial offenses in the community. For pragmatic, legal, and ethical reasons, society is unlikely to undertake a comprehensive effort to identify future sex offenders before their first sex crime; therefore, we anticipate that psychologists and other mental health professionals will continue to be uninvolved in large-scale efforts to predict first offenses.

On the other hand, attempting to predict whether convicted but released sex offenders are likely to commit further sex offenses is inherently more manageable (if still daunting). It is certainly considered by many to be a socially more acceptable task.

The best available single estimate of the base rate of sex offender recidivism is probably 13.4%, reported by Hanson and Bussière (1998), calculated as the rate at which individuals move from the category of non-reoffender to reoffender with regard to any sexual reoffense. This figure was averaged across 61 meta-analyzed studies, comprising offenders who had committed widely varying types of sexual offenses, followed up over varying (albeit not lengthy) intervals after release, and yielding a total \( N \) of nearly 29,000. Obviously, the aggregate recidivist figure is possessed of substantial statistical precision. The figure is widely cited and has come to receive general recognition if not universal acceptance; it is not without its criticisms, however. As Russell (1982) noted, sexual offenses suffer more underreporting than other crimes. Most studies of recidivism, including those cited by Hanson and Bussière (1998), use reconviction or, more rarely, rearrest as proxy measures for recidivism. They also looked at the number of offenders who committed one or more new sexual offenses, not the number of new offenses committed. Given the underreporting of sex offenses and the fallibility of police work and of our legal system, using rearrest and reconviction as reoffending measures will both produce false negative
“diagnoses” of recidivism (Doren, 1998). On the other hand, the same fallible police work and the same fallible judicial system can produce false positives as well, as has been shown in a number of cases by DNA evidence (Innocence Project, September 24, 2005). Overall, the general consensus of informed opinion is that the balance of these opposing influences considerably favors false negatives, and so the true recidivist rate, over follow-up intervals similar to those in studies reviewed by Hanson and Bussière, is likely to be significantly greater than 13.4%. Marshall and Barbaree (1990) used unofficial sources (e.g., files from Canada’s Child Protective Agency and local police agencies, containing evidence linking offenders in their study to probable new offenses), estimated the rate of probable sexual reoffending at 2.4 times the reported rate over a follow-up of somewhat greater than four years. Their ratio pertains to the number of new offenses, not the number of reoffenders. When Marshall and Barbaree (1990) looked at reoffenders instead, they found that approximately 20% of released offenders appeared to have committed one or more further offenses; this is not dramatically higher than the 13.4% rate reported by Hanson and Bussière.

That it matters whether one looks at reoffenders or reoffenses is not surprising. A small fraction of the offender population, if they are especially criminally active, may reoffend many times (Janus & Meehl, 1997); this would lead to most released offenders not being identified to authorities as probable reoffenders, while the high-activity criminals would be at high risk of being caught and labeled (because each fresh offense increases their chance of detection). If even a fraction of their crimes are detected, they are likely not only to be labeled as reoffenders, but also labeled as multiple offenders. On the other hand, multiple reoffenses by each member of a great majority of the offender population would likewise lead to high rates of capture in official and/or unofficial statistics (again, because each act exposes them to increased risk of detection),
high recorded ratios of definite or probable reoffenses per reoffender; and relatively few
individuals unknown to authorities as definite or probable reoffenders. Contrariwise, multiple
reoffenses by a more modest fraction of the offender population, but with the possibility that a
larger fraction commits reoffenses at a significantly lower rate (say, about one offense per
recidivist), would lead to much more numerous individuals being unknown to authorities as
probable reoffenders because they would be likely often to escape detection. This pattern is
perhaps most consistent with Marshall and Barbaree’s (1990) findings. On balance, the recidivist
rate reported by Hanson and Bussière (1998) may not be such an underestimate of the true
recidivist rate as has been claimed by critics.

Another potential criticism that could be leveled at Hanson and Bussière (1998) is that
the studies included in their meta-analysis only followed offenders for a period of four to five
years after release. Offenders who first re-offended after that period would have been counted in
the underlying studies, and in the meta-analysis, as non-recidivists. At first blush this is a telling
criticism; it points out an undeniable truth: recidivism rates are time-dependent statistics.
However, the criticism is really only a legitimate methodological demurer against Hanson and
Bussière if readers are in the habit of misunderstanding Hanson and Bussière’s figure as being a
lifetime risk of becoming a recidivist. However, these authors expressly pointed out their risk
figure was no such thing. Indeed, studies by Hanson himself (and by others) have followed
offenders over a period of 20 years and have reported sexual recidivism rates considerably
higher, in the range of 30–40% (Hanson, Steffy, & Gauthier, 1993; Prentky, Lee, Knight, &
Cerce, 1997), when reconviction for a new sexual offense is used as the recidivism measure. The
five-year sexual recidivism rate at five years in Prentky, et al. (1997) was close to the Hanson-
Bussière rate, suggesting that the increased rate by 20 years is related to lengthened follow-up,
and not to some difference between the Hanson-Bussière and Prentky et al. (1997) samples. The general trend in the literature would seem to indicate that while the Hanson-Bussière rate substantially underestimates the long-term recidivist rate, it is likely a fair estimate of the detected recidivist rate over five years post-release. It is also important to note, as is usual in follow-up studies, that most of the “relapse” takes place early on—the rate seen at five years does not double by twice five years, it approximately doubles by $4 \times 5 = 20$ years, indicating that the hazard (instantaneous risk of a detected first new offense) is probably declining over time since release.

While far fewer studies have looked at long-term recidivist rates than short-term rates, the available evidence suggests that the cumulative rate continues to grow between five and 20 years after release while the instantaneous rate continues to decline, with an interval estimate of the twenty-year rate being 30–40%. These are certainly sizeable rates. However, they are considerably lower than the rates for most other crimes, including most non-sexual violent crimes (Beck & Shipley, 1989).

The heinousness of sexual crimes and high victim and social costs of sexual offending, despite the less than extraordinarily high recidivist rates, have 17 states (Eric Janus, personal communication, September 22, 2005) to pass laws to protect the public from at least some sex offenders. These laws generally use a process of civil commitment for ostensible treatment, in which those offenders judged most dangerous, or most likely to recidivate, are committed to a secure inpatient facility after completion of their prison terms, for specialized sex offender treatment. As of 2004, 17 states were undertaking the civil commitment of some persons judged to be “sexually violent persons” or similar locutions (Caidwell, 2002; Seto, 2005).
These special civil commitment laws require trained mental health professionals to make predictions about the future likelihood of sex offenses, violent offenses, general criminal behavior, or some combination of these, for sex offenders about to be released from prison after completing their terms of incarceration—or, in certain cases, offenders who have already been released from prison and who have not been arrested for any new offense (Doren, 2001; Harris, et al., 2003). Essentially, the finding that an individual is dangerous enough to be civilly committed is based on the conclusion that the offender would probably recidivate if released into the community (Covington, 1998); this conclusion is based on predictions made by mental health professionals as to the offender’s likelihood of future reoffense. More specifically, what the state, and therefore the state’s experts, are required to prove in order to commit an offender are three elements: prior sexual misconduct, current mental disorder (broadly defined to include any personality disorder, at least in most jurisdictions), and likelihood of engaging in future sexual misconduct (Janus & Meehl, 1997). The states employing civil commitment laws, therefore, assume that future sexual violence can be predicted with sufficient accuracy to allow the state’s interest in protecting the public to counterbalance the offender’s liberty interest, a belief that has been found to have no constitutional impediments and which courts have refused to question (Janus & Meehl, 1997). Despite the courts’ unwillingness to question whether sexual recidivism can be sufficiently accurately predicted (e.g., on Frye or Daubert grounds; Frye v. United States, 293 F. 1013, 1014, D.C. Cir. 1923; Daubert v. Merrell Dow Pharmaceuticals, 509 N.W.2d 579 1993), the question is ultimately an empirical one.

As will be discussed in this review, studies have identified a number of impediments to our ability to make the predictions required of us by Hendricks-type civil commitment laws. The idea that the most dangerous offenders can be identified with sufficient accuracy for the law’s
Risk Assessment 11

proper working presupposes not only (a) that the probability of future offending can be measured, but also (b) that experts can discriminate adequately between higher and lower probabilities of recidivism, (c) that there are standards in place to allow commitments of those in the higher probability group while exempting the lower probability group, and (d) that these standards are actually consistently employed (Janus & Meehl, 1997). If any of these assumptions does not hold, then excessive numbers of incorrect decisions will be made. Dangerous offenders will be released into the general population, while offenders who would not actually re-offend if released will be unnecessarily deprived of liberty through civil commitment, at high cost to the state and the individual. Clearly, neither of these outcomes is desirable.

As previously discussed, sexual offending (especially violent offending) causes considerable suffering to victims. When perpetrated by recidivists, this suffering could in theory have been avoided (in jurisdictions with Hendricks-type statutes), and so seems especially grievous. On the other hand, unnecessary commitment entails deprivation of liberty for the offender, other deprivations for their family and friends, losses to society of positive contributions the offender would have made if at large, and of course a sizable and utterly pointless expense to the state. Most baneful to those wishing to optimize the use of scarce resources is the occupation of a security hospital bed by someone needing neither preventive detention nor recidivism-preventing treatment, in place of somebody else who is arguably in need of both. Civil commitments generally cost $100,000 or more per offender per year in 1997 dollars (over $120,800 in 2005 dollars, according to the Consumer Price Index, compounded over the interval 1997–2005; Department of Labor Bureau of Labor Statistics, 2005), which is more than four times the cost of incarcerating the same offenders in prisons (deFiebre, 1995; Janus & Meehl, 1997). Given such high costs of false negative and false positive predictions, it is
crucial to determine how accurate we can expect recidivism predictions to be. It is this issue to which we now turn.

Common Problems in Behavioral Prediction

Prediction is a common enterprise in numerous fields, both in and out of psychology. Within psychology, required predictions range from determining how much recovery of functioning a patient with traumatic brain injury is likely to show over time, determining which patients are most likely to attempt or commit suicide, to the prediction at the heart of this review, determining which sexual offenders are most likely to commit future sexual offenses. Before turning to issues unique to sex offender risk assessments, we will first examine issues that affect the accuracy of all behavioral predictions, as these problems apply equally to risk assessments.

Base Rates

A first major problem is the relationship between base rates and predictive accuracy. A base rate is the a priori probability of a condition being present, or a behavior being emitted, in a specified population. By a priori probability (prior probability for short) we mean the probability one would rationally impute to an individual, knowing their population membership but ignorant of any personal information about that individual. The base rate, together with the operating characteristics of a test (sensitivity, specificity), determine the test’s accuracy when used with a particular population and also dictate the optimum use of the test (Bayes, 1764; Meehl & Rosen, 1955). Base rates influence accuracy in that predictive instruments make the most correct predictions when the base rate of the behavior being predicted is 50%. For this reason, rare events are inherently difficult to predict. In particular, an instrument with insufficient specificity (probability of correctly labeling a non-recidivist), used with a population having a low recidivism rate, may yield extremely numerous false positives. Indeed, a test which is valid (i.e.,
correlates with the criterion above zero) can be less accurate than “betting the base rate” (i.e.,
predicting the modal outcome for everyone), due to misplacement of the cutting score for the low
base rate population.

Quantifying Predictive Accuracy

There are several ways to quantify the accuracy of predictions, which it is convenient to
distinguish here before reviewing the literature on predictive validity. One accuracy measure is
the Pearson correlation coefficient $r$ and its special forms for dichotomous criteria, the point
biserial $r$ (used with a quasi-continuous predictor) and the $\phi$ coefficient (used with a
dichotomous predictor). Negative values indicate that, as coded, higher values of the predictor
are associated with lower values of the criterion; zero values indicate no linear association with
the criterion. The absolute magnitude ranges from zero to one, with one indicating a perfect
linear relationship. A related measure is $R^2$, which is the percentage of variance in the outcome
variable that is explained by a linear combination of one or more predictor variables. While
commonly used in studies of prediction, $r$ and $R^2$ are not ideal measures of predictive accuracy
for dichotomous outcomes, because they change in an unintuitive (nonlinear) fashion with
changing base rates.

A second way to quantify accuracy, developed in epidemiology is through the use of two
statistics: sensitivity and specificity. Sensitivity (which we will denote by $\alpha$, following one
common notational convention) is, for our purposes, the probability that a future recidivist will
be correctly predicted to be such. Specificity (denoted $\beta$) is the probability that a non-recidivist
will be correctly identified as such. These quantities are designed to be base rate-free. Hence,
base rate information must be added to determine how well a test, with a given sensitivity and
specificity, will perform in a population with that base rate. By the same token, as base rate-free
accuracy measures, they do not tell one whether a test will outperform “betting the base rates,”
let alone by how much. In assessing predictive accuracy, sensitivity and specificity are
nevertheless generally much more appropriate than $r$ and $R^2$ because they relate directly to the
dichotomous nature of the criterion (Mossman, 1994a).

In recent research on prediction, the use of receiver operating characteristic (ROC)
analyses has become increasingly common. Derived from signal detection theory, ROC analyses
provide accuracy measures that are independent of base rates (Mossman, 1994b). An ROC curve
traces out an infinite set of paired sensitivity-specificity tradeoffs, as a function of a diagnostic
test cutting score that sweeps from its most conservative possible value (no positive diagnoses)
to its most liberal value (no negative diagnoses). The conventional ROC curve shows $\alpha$ plotted
against $(1 - \beta)$ (Mossman, 1994a). Various measures of ROC performance can be derived, but
most commonly the ROC curve is summarized by the area under the curve (AUC), which ranges
from .50 (unless the test manages to perform worse than chance) to 1.0 (perfect accuracy). The
value of an AUC can be interpreted for our problem as the probability that a randomly selected
recidivist scores higher on the predictive instrument than does a randomly selected non-recidivist
(Mossman, 1994a; Nunes, Firestone, Bradford, Greenberg, & Broom, 2002). The AUC has the
defect of its virtues. Since it is independent of base rates (and of cutting scores), it does not tell
how well a test will perform in a given population with a fixed base rate, or with a specific
cutting score (which may not be optimal for that population. Like sensitivity and specificity,
ROC measures are particularly appropriate to predictions with dichotomous outcomes.

Clinical Versus Statistical Prediction

A sizable literature has been devoted to debate on the relative merits of clinical versus
statistical prediction (see, for example, Grove & Meehl, 1996; Grove, Zald, Lebow, Snitz, &
Briefly, clinical predictions are those in which the data on which the prediction is based are aggregated in an informal manner without a 100% reproducible scheme for combining predictive data, such as when a clinician reviews the information collected on a patient and makes a prediction based on his or her global interpretation of that information. Statistical prediction is a subclass of a general class of prediction schemes that also includes actuarial prediction and mechanical prediction, terms that are generally, albeit unfortunately, used interchangeably. More precisely, mechanical prediction means only that the data combining scheme is 100% reproducible and requires no professional judgment in its application; the data combination scheme need not be empirically based or involve a statistical formula. Actuarial prediction, in the narrow sense, means that the predictive data are combined in a fashion similar to an old-fashioned life insurance actuarial table. Statistical prediction, again in its narrow sense, means that predictive variables are combined using some type of statistical model (e.g., multiple regression) so that variable weights and generally decision-making cutting scores as well, are determined based on what is optimal in a training set of data. All statistical and actuarial predictions are mechanical, but the converse is not true.

Mechanical prediction schemes involve collecting specified pieces of information and attaching specified weights to each piece of information in order to make the prediction (The weights can be as simple as zero and one.). The issue of clinical versus mechanical prediction is central to the current discussion, as either type of prediction can be used in predicting sexual recidivism; indeed, when clinical and mechanical prediction are defined as they are above, exactly one of them can be used. There is a certain sense in which mechanical and clinical predictions can be used together, as when a clinician is given the output of a mechanical prediction scheme, and is allowed to follow that prediction or alter it as he or she sees fit.
However, this is technically a form of clinical prediction, since the combination of the mechanical prediction and the clinician’s own insights is informal, require professional judgment, and the resulting predictions are not 100% reproducible.

Before turning to specific issues of the validity of these techniques in predicting sexual recidivism, we will first explore the relative merits of the two major modes of prediction generally, beginning with clinical prediction. Clinicians can be called upon to make an extraordinary variety of predictions: whether a patient is likely to attempt suicide, whether a patient will act out violently against others, whether someone is likely to succeed in a certain occupation, whether someone is likely to experience cognitive decline, and so on. Typically, these predictions are made informally, without the application of statistical formulas or the use of specialized data-combining instruments (Otto, 2000). Unfortunately, the vast preponderance of the evidence indicates that such clinical judgments are often error prone and at best perform as well as mechanical predictions; most often, they are inferior to mechanical predictions (Grove, et al., 2000). This is apparently because clinical judgment is subject to a wide variety of biases and heuristic short cuts that tend to introduce both random and systematic error into predictions. In a comprehensive review of the clinical judgment literature, Garb (1998) found that clinical predictions of suicide were of low validity; clinical predictions of violence were better, generally showing moderate validity, but were out-preformed by statistical methods. Though noting that the issue had not been directly studied, Garb (1998) also suggested that clinical prognoses were likely to be of limited validity. In another review, Ennis and Litwack (1974), in addition to noting clinicians’ difficulties predicting violence, found that clinicians made poor predictions of the need for hospitalization, as a substantial number of patients recommended for commitment but released by the courts made satisfactory adjustments to community life. Clinicians were also
found to make low validity predictions of response to treatment. Ennis and Litwack’s (1974) review looked specifically at the judgments of psychiatrists, though their findings would certainly seem applicable to all clinicians, given the frequent findings of poor clinical predictions. In yet another review, Faust and Ziskin (1988) noted that clinical predictions generally showed higher rates of errors than of correct judgments.

Despite the numerous areas in which it has been found wanting, clinical prediction is not without its supporters. Holt (1958) argued that clinical judgment may appear less valid in research studies than it actually is, as research studies tend not to include many of the tasks in which clinicians tend to engage when making judgments, such as determining what data are needed and effectively gathering those data. A number of authors have agreed about the importance of the former task. Westen and Weinberger (2004) pointed out that clinicians must ultimately decide what variables are important enough to be considered in making predictions. However, this point does not address the data combination issue, as clinicians could, in the development of a statistical or actuarial prediction scheme, decide which variables are worth examining, but use statistics to determine which variables are really valuable, and how best to combine (weight) the variables.

It has also been argued that clinicians may be well suited to identifying idiosyncratic variables that improve on mechanical judgments in individual cases. However, it has been repeatedly found that using clinical judgment to overrule mechanical predictions does not improve upon statistical prediction (Dawes, Faust, & Meehl, 1989; Grove, et al., 2000).

Regarding Holt’s (1958) point about the importance of clinicians effectively gathering information, many agree. Sawyer (1966) noted that clinical observation is often a useful tool in prediction. Dawes, Faust, & Meehl (1989) noted that many clinical observations, such as
observation and categorization of visual cues like body language and facial expressions, cannot (yet) be replicated by artificial means. However, Meehl pointed out at the outset of his 1954 book on the subject that a sharp distinction must be drawn between the source of data to be combined, and the method of combining data. Clinical observations may be very valuable, but they need not be combined clinically; they can be encoded and then included in mechanical prediction schemes (Dawes, et al., 1989; Westen & Weinberger, 2004).

In and of itself, the rather consistent finding that clinical prediction is subject to a considerable amount of error would seem to outweigh the few arguments in favor of its use. Nonetheless, a number of authors have attempted to explain the specific problems with clinical judgment. One notable problem stems from the complexity of the decisions which clinicians are required to make; these predictions require consideration of a substantial number of variables. Human judgment, and therefore clinical judgment, seems to be best suited to dealing with simple situations in which only a manageable number of variables without many complicated interactions are considered. Given that the predictions that clinicians are frequently called upon to make tend to be quite complex, they are ideally suited to the simple, consistent rules offered by mechanical means of prediction, which human judges are generally not equipped to use (Hammond & Summers, 1972; Shulman & Elstein, 1975). In addition to the difficulties that clinicians have in processing complex situations, they have also been found to have difficulties being consistent across cases. Essentially, even when clinicians develop valid decision rules, they seldom apply those rules consistently over a large number of cases (Dawes, 1971; Goldberg, 1970), negating the advantage of having a workable rule as it does not improve predictive accuracy across multiple cases. The inaccuracy so often seen in clinical judgment, then, seems to
be due at least in part to the cognitive limitations of human judges and these judges’
inconsistencies in applying simple rules.

While some overall cognitive limitations may limit clinicians’ ability to make
predictions, numerous authors have also noted specific cognitive errors that clinicians tend to
make, which limit the accuracy of their predictions. If these errors were avoided, predictive
accuracy could potentially improve. Many of the errors that affect clinical judgment involve the
use of heuristics, or general principles designed to simplify complex judgments. As discussed
above, simplifying complex situations is often useful in improving upon human judgment, which
may not be optimal in dealing with such situations, and heuristics can therefore be quite useful.
However, in simplifying situations in which the data is of limited validity, heuristics introduce
systematic errors into judgments (Tversky & Kahneman, 1974). Chief among the heuristics that
could contribute to inaccuracy in clinical judgment is the representativeness heuristic, which
involves making judgments about whether an individual belongs to a particular class by
comparing that individual to other members of the class whom one has previously encountered
(Garb, 1996; Tversky & Kahneman, 1974). A number of authors have stated that this heuristic is
commonly used in clinical judgment (Dawes, 1986; Garb, 1996). A problem with using the
representativeness heuristic in clinical settings, such as when making a diagnosis, for example, is
that a clinician attending to whether a case resembles an exemplar is likely not attending to more
formal diagnostic criteria (Garb, 1996). This practice is especially problematic when the
exemplars are less than ideal members of the particular class, thus leading the clinician to assign
a case to a class based on comparison with something that does not really describe that class.
This heuristic could certainly lead to error if used in prediction. However, Garb (1996)
conducted three studies in which clinicians were asked to make judgments about case studies and
also to rate the cases on how similar they were to cases with which the clinician had previously dealt. He reported that clinicians relied on representativeness in making diagnoses but not in making predictions. Instead, past behavior tended to most strongly influence predictions, which Garb (1996) called the past behavior heuristic. This heuristic is problematic because individuals can adapt and change their behavior, leading to misprediction. A final heuristic is availability, whereby predictions are affected by the ease with which one can bring to mind exemplars of the predicted event (Tversky & Kahneman, 1974), either because of commonness or vividness.

An additional class of judgment problems is biases: giving undue weight to a factor, or even weighting it the wrong way, due to pre- and ill-conceived notions about the relationship between the factor and the criterion. Garb (1998) identified several such biases. The risk of violence was overestimated for African-Americans, but not Caucasians, in psychiatric settings; similarly, it was overestimated for men but not women. Psychiatric prognosis was biased downward by low socioeconomic status in some studies, and also by advanced age. African-Americans were more likely than Caucasians to be misdiagnosed as schizophrenic (usually paranoid subtype), whereas Caucasians were more likely to be misdiagnosed as having psychotic affective disorders. Correspondingly, African-American patients were misperceived as requiring hospitalization more often than Caucasians with similar symptomatology.

Even the manner of presenting data can bias clinical judgment. When data are acquired sequentially, judges tend to over-predict the likelihood of an event, which some authors have attributed to judges treating each new piece of information as lacking in uncertainty (Elstein, 1976; Gettys, Kelly, & Peterson, 1973; Kahneman & Tversky, 1973). Contrariwise, when data are presented together, judges tend to underestimate the probability of an event (Elstein, 1976; Shulman & Elstein, 1975; Slovic & Lichtenstein, 1971).
Westen and Weinberger (2004) identified an additional set of factors that may have deleterious effects on clinical predictions. First, the very items of data available may vary from one case to the next; to make strictly comparable predictions, a clinician must have access to the same data in each case or else risk varying validity as a function of varying information sources. Experience with the predictive task may affect proficiency, and it is to be expected that opportunity to engage in reinforced learning of the task skill would influence accuracy. By contrast with statistical prediction, in which variable selection and weighting explicitly detects and minimizes misprediction, clinical judges often receive little or no valid, let alone timely, feedback about their predictions’ validity. It is little wonder that they have great difficulty identifying their errors (Dawes, et al., 1989) or making needed changes.

In sex offender risk assessment, there may be occasions when a clinician receives predictive accuracy feedback—specifically the highly negative feedback that arrives when the media announce that a sex offender (whom the clinician assessed) has reoffended in a particularly heinous fashion. However, this kind of feedback is arguably more likely to feed an accuracy-damaging availability heuristic, than to assist in trial-and-error learning; by analogy with the Skinner box, it is as if the rat gets reinforced once every 1,000 bar presses—and that time it gets a rock dropped on its head. One doesn’t expect much learning to take place under such a schedule. For a predictive task like recidivism prediction, regularly and reasonably promptly delivered valid feedback, especially positive feedback that shapes predictions that are partially if not entirely accurate, would be expected to result in improved clinical predictions through a process analogous to that by which actuarial instruments are validated (Westen & Weinberger, 2004). These authors also note that clinical judgments are most likely to be
inaccurate when the clinician makes predictions they were not trained to make; the structured clinical judgment format discussed below could help in training and thus improve accuracy.

It is an empirical question whether, even with all due effort expended to optimize clinical judgment, the predictions of human judgments can be expected to outperform mechanical, actuarial, or statistical data combination schemes. Over a large and ever-growing literature, it is the rather consistent finding that clinical prediction is less accurate than, or at best equivalent to, mechanical, actuarial, or statistical prediction (e.g., Grove, et al., 2000). This suggests one of two conclusions. Either the optimizations suggested by advocates of refined clinical judgments like Westen and Weinberger are typically not effected or, if they are, they do not increase validity to the desired degree.

The alternative to clinical prediction, to which we now turn our attention, is mechanical prediction. As defined by Grove and Meehl (1996), mechanical (formal, algorithmic) predictive methods are those methods that are 100% reproducible and require no professional judgment in application. This includes statistical prediction, which uses formulae such as regression equations to make predictions, algorithmic prediction, in which judgments are produced in a consistent manner (e.g., by a computer program), and actuarial prediction, which involves the use of experience tables. These methods are usually treated as if they were interchangeable, which conceptually they certainly are not. However, we will follow this practice here because the empirical literature turns out to sanction this loose usage; we will follow custom and refer to them all as “statistical” predictions even though they need only be algorithmic.

The validity of statistical predictions has been very widely studied in relationship to clinical predictions. Simply put, statistical methods of prediction are essentially always found to be as accurate as, or more accurate than, clinical methods. In the largest published review of this
area, Grove, et al. (2000) meta-analyzed 136 within-study comparisons of statistical and clinical predictions. Statistical predictions averaged about 10% more accurate than clinical predictions. Considering ± 1 in transformed effect size (equivalent to about ±10% in hit rate) to be a zone of equivalence between the two modes of prediction, statistical methods outperformed clinical predictions in 47% of studies, 46% of studies showed approximate equivalence, and 6% showed superiority for clinical prediction. The studies showing clinical prediction superior did not cluster in any one predictive domain, type of judge, or any other common methodological feature that the authors could discern. These findings, based on far more studies (the largest previous review had just 45 studies, less than 30 of which constituted within-study comparisons of statistical versus clinical prediction) and sounder methodology (no previous review employed meta-analysis), mirror the findings of all previous reviews (Dawes, et al., 1989; Garb, 1994; Marchese, 1992; Meehl, 1954; Sawyer, 1966; Sines, 1971; Wiggins, 1981), except for Holt (1970).

Holt (1970), the outstanding critic of Meehl (1954) and, since the mid-1950s, of what he saw as a pro-actuarial attack on the whole enterprise of clinical psychology, did not present a systematic review of the empirical literature. Rather, he criticized what was then the largest such review, Sawyer (1966). The criticism was not entirely relevant to our discussion, since Sawyer (1966) concerned structured versus unstructured data collection, as well as clinical versus statistical data combination. Holt (1970) argued that one should examine clinical vs. statistical methods at all stages of the judgment process, not just at the data combination stage. This was an odd criticism of Sawyer (1966), since Sawyer’s main contribution over Meehl (1954) was to examine “statistical” (actually, objective/psychometric) assessment versus “clinical” (actually, unstructured/nonpsychometric) assessment, as well as clinical versus statistical data combination. Moreover, Holt muddled Meehl’s (1954) absolutely crucial distinction between
data sources and data combination. Furthermore, Holt (1970) argued that clinicians should not be judged on their success at making predictions, because it is the identification of traits (especially source, rather than surface, traits), not making behavioral predictions, that is the hallmark of the clinical enterprise. However, we (and others) have argued strongly that behavioral prediction is an important and frequent aspect of clinical work. Finally, Holt (1970) cited not one single empirical study to counter Meehl’s, and Sawyer’s, conclusion that statistical data combination is as accurate as, or more accurate than, clinical data combination.

The success of statistical prediction has been attributed to a number of factors. First, statistical prediction automatically overcomes limitations that plague the clinician. Statistical prediction schemes are always consistent, while humans are subject to practice effects, fatigue, context effects, and whimsy (Dawes, 1971; Goldberg, 1970; Hammond & Summers, 1972; Shulman & Elstein, 1975). Statistical predictions can process effectively unlimited amounts of data, whereas humans saturate at less than ten simultaneous pieces of information. Statistical predictions manage redundancy far better than do human judges. Clinicians often over-estimate redundant information, such as multiple tests to measure the same construct, believing that convergent information strengthens validity. This belief can lend excess weight to “me too” information, leading to not gathering, ignoring, or down-weighting non-overlapping data that might reduce unexplained variance in the criterion. Over reliance on redundant measures can even lead to loss of validity, if an additional measure correlates with the predictive construct but not the criterion. Statistical algorithms identify invalid redundancies and root them out; they down-weight variables that, while redundant, still reduce unexplained variance in the criterion (Kahneman & Tversky, 1973). These advantages—consistency, unlimited data
capacity, elimination of redundancy, and optimal weight assignment—likely account, at least in part, for the consistently reported greater validity of statistical predictions.

Despite the consistent findings of its superior accuracy, statistical prediction is not without its detractors. A common complaint is that statistical predictions lack the flexibility of their clinical counterparts. Clinical predictions can in principle consider different variables for each case, allowing consideration of factors that may not be common enough for inclusion in a formal prediction scheme, but which are nevertheless relevant in individual cases. This phenomenon is referred to in the literature as Meehl’s “broken leg” problem (Meehl, 1954): suppose we construct an actuarial table to predict whether X will go to the movies on a given night. It is well established that $\Pr\{X \text{ goes to movies } | \text{ Tuesday}\} = .8$. However, we happen to know that on this particular Tuesday, X has just broken his leg and has been fitted with a hip cast. Now, when we constructed our experience table, we never observed X breaking his leg; it was too rare an event, and so it’s not included among the variables making up the actuarial table. However, common sense (or, if you like, the geometry of hip casts versus theater seats) tells us that X’s probability of going to the movies is now essentially zero. In this case, a clinician would be correct in using X’s broken leg to overrule the actuarial prediction. Cogent as this argument is in principle, it does not establish that clinicians routinely detect “broken leg” variables, overrule actuarial predictions only when it obviously pays to do so, and thus make net improvements on actuarial predictions. Whether clinicians handle “broken leg” cases in this fashion is an empirical matter, and the studies analyzed in Grove, et al. (2000) directly addressed to this question uniformly favor the hypothesis that clinicians do not improve on statistical predictions. Instead, they overrule the statistical predictions too often and at the wrong times, and end up being bested by the straight mechanical predictions.
The inflexible nature of statistical prediction has led some (e.g., Douglas & Ogloff, 2003) to suggest that statistical prediction may only outperform clinical prediction in the circumstances in which the statistical scheme was developed. In other words, mechanical schemes may be best suited to make predictions about people who are similar to the scheme’s validation sample, while clinical predictions can adapt to populations that differ in some important way from the validation sample. Grove and Meehl (1996) counter that there is no reason to believe that an actuarial scheme that has demonstrated validity in several populations will not also be valid in another population. (Of course, this presumes a fact that may not be in evidence, namely that a statistical scheme has had its validity generalization tested.) If Douglas and Ogloff were correct, then validity shrinkage should adversely affect the accuracy of statistical prediction schemes in validity generalization samples, such that clinical prediction (with its great flexibility) would prove superior in such samples. However, in the Grove et al. (2000) meta-analysis, this very issue was addressed in 14 studies, with the following results. The weighted mean transformed accuracy for clinical prediction was .94 (weighted SD=1.40), versus .98 (1.20) for statistical prediction in validity generalization samples, t(13)=1.86 (p < .085), showing a trend in the opposite direction from that predicted by Douglas and Ogloff.¹

Otto (2000) points out that one of the biggest problems of the actuarial approach is the limited availability of actuarial schemes. This is a valid criticism to the extent that it is impracticable to create an actuarial scheme for a particular problem (Grove & Meehl, 1996). However, this criticism obviously does not apply to situations in which an actuarial scheme is available.

¹ The accuracy statistics are transformed to put hit rates, correlation coefficients, areas under receiver operating characteristic curves, etc., into a common metric. The resulting metric is arbitrary, but is comparable between clinical and statistical prediction schemes. The weights have to do with the fact that hit rates, correlation coefficients, etc., have different asymptotic sampling variances, and different studies have different Ns (Grove, personal communication, September, 24, 2005.)
Otto (2000) also neglects another, seemingly little known but very important psychometric fact, with profound ramifications for practical prediction work. Simple equal-weights linear models lose at most a few percent of predictable variance, on average, when compared to optimal least squares weights in multiple regression (Grove, submitted). If key predictive variables can be identified based on research or even expert consensus, and if that consensus is sufficiently accurate to dictate the correct sign of the regression coefficient for each variable, then a statistical formula can be constructed with no more work required than to find out the means and standard deviations of the predictor variables; the criterion variable need not be measured at all. Hence, those who are concerned that developing statistical prediction schemes is very expensive, time-consuming work are only partly right. Developing a carefully optimized statistical prediction scheme for long-term, difficult to measure outcomes is expensive and time-consuming. However, developing a useful statistical prediction scheme that is likely to do as well as, and quite possibly better than, clinical judgment, may be accomplished rather cheaply and easily.

While the vast majority of the research literature favors mechanical prediction either on accuracy or economy grounds or on both, some authors have discussed a number of potential advantages of clinical prediction. As neither method has achieved anything close to perfect predictive validity, many have attempted to combine the two in order to utilize the advantages of each. Commonly, clinicians are given a statistical prediction and allowed to make the final judgment (Goldman, et al., 1982; Leli & Filskov, 1981, 1984; Sacks, 1977; Schofield & Garrard, 1975; Watley, 1963; Wexler, Swender, Tunnessen, & Oski, 1975). Note that this is not, as is often claimed, a true hybrid of statistical and clinical prediction; there is no true hybrid in Meehl’s (1954) sense of these terms. Since the final prediction employs professional judgment
and is not 100% repeatable, “hybrid” prediction is actually clinical in nature. In Grove, et al. (2000), seven studies of “hybrid” predictions showed a weighted mean effect size (transformed clinical accuracy minus statistical accuracy) of 0.078, almost exactly the same as the mean for all studies analyzed (overall weighted mean=0.086). No study showed superiority of hybrid judgment over statistical prediction. In fact, these numbers can be interpreted as indicating that the clinician subtracted 1% in accuracy by adding their judgment. This, of course, calls into grave question the advisability of allowing clinicians to “adjust” statistical predictions, which is a common (indeed, modal) practice in sex offender risk assessments (W. M. Grove, personal communication, June 15, 2005).

Some have attempted to combine the perceived advantages of the two predictive modalities through a more formal sort of combination (though still not a true hybrid in Meehl’s terms): the structured professional judgment method (SPJ; Douglas & Kropp, 2002; Hart, 1998), which attempts to combine the reliability, structured procedure, and strong relationship to predicted outcome of actuarial approaches with the flexibility of the clinical approach. SPJ assessments use a specified item set with scoring criteria, much as do the actuarial schemes. Rather than referring a total score to an actuarial table to determine level of risk, in SPJ the clinician makes the final judgment of low, medium, or high risk by following, albeit not slavishly, a set of guidelines (Douglas & Ogloff, 2003). This approach appears somewhat similar to other procedures in which a clinician is given the results of an actuarial prediction scheme and allowed to deviate from its output in making his or her own prediction (Goldman, et al., 1982; Leli & Filskov, 1984; Leli & Filskov, 1981; Sacks, 1977; Schofield & Garrard, 1975; Watley, 1963; Wexler, Swender, Tunnessen, & Oski, 1975); these other attempts have never proven to be more effective than straight actuarial prediction. Douglas and Ogloff (2003) cite several studies
that supposedly show such SPJ outperforming the numerical scores on the instruments on which
the judgments are based (Kropp & Hart, 2000; Dempster, 1998; de Vogel, 2002; Douglas, 2001;
Douglas, Ogloff, & Hart, 2003; Gretton & Abramowitz, 2002). Of these studies, however,
Dempster (1998) was an unpublished Master’s thesis and de Vogel (2002), Gretton and
Abramowitz (2002), and Douglas (2001) were conference presentations, none of which were
subject to peer review, and the claims cannot be readily verified. A later article describing the
results of the de Vogel presentation (de Vogel, de Ruiter, van Beek, & Mead, 2004), which will
be discussed below, does in fact show SPJ outperforming mechanical usage of an instrument, but
it is the only obtainable study cited by Douglas and Ogloff (2003) to contain a comparison
showing the reported pattern. Douglas and Ogloff’s claims thus appear to be largely
unsubstantiated.

Prediction of Sexual and “Violent” Recidivism

The literature on prediction has become increasingly relevant to the study of sexual
offenses, given requirements by numerous legislatures that clinicians predict which sexual
offenders are likely to commit future offenses (Doren, 2001; Harris, et al., 2003). Before
addressing approaches used for recidivism prediction and research on their validity, we will first
discuss what criteria researchers are attempting to predict, then what variables have been found
to predict these criteria.

Defining the Criterion: What is Recidivism?

The sex offender recidivism prediction literature commonly uses four criterion variables:

1. Sexual reoffending per se. Any sexual crime including, e.g., soliciting a prostitute,
promoting a lewd or lascivious act (such as an old New York State law, Sec. 175,
that forbade man #1 to allow man #2 to perform a lewd act on man #1—but said
nothing about women), selling legal pornography too close to a school, offering to sell sexual acts for money, exhibitionism, window peeping in certain jurisdictions (e.g., Wisconsin), uninvited touching by a psychotherapist of a client’s buttocks, etc. would all be counted under this heading. These are equally “sexual offenses” along with acts most people think of as sex crimes: the sexual mistreatment of children, sexual acts involving force or the threat of force, acts involving penetration, and so on. Interestingly, status offenses (acts that would not be illegal for many but may be illegal for some, owing to their status) such as possessing legal pornography for a sex offender parolee, would count as sex offender recidivism here.

2. “Violent” recidivism. The quotations around the word “violent” are placed there to alert the reader to the fact that the crimes counted under this heading mostly do not qualify as violent crimes according to the FBI UCR classification (which includes just four crimes—murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault), and many of them might not be considered “violent” by typical speakers of the English language. For example, merely touching another person without their consent legally constitutes the crime of battery. If no appreciable force is used, and there is zero risk of injury, most people would probably not consider such an act to be violent. Nevertheless, battery would be counted in the recidivism literature as a violent crime (but not an FBI violent crime), and a sex offender who battered in such a fashion would be counted a recidivist;
3. “Violent crime” including any sexual offense. This includes the union of (1) and (2); and

4. Crime in the broad sense. Used less often than the preceding three criteria, this variable is often narrowed so that only felonies are counted. Property crimes would be counted here. Committing a drug offense (by far the single most common reason for state imprisonment in the U.S. today) or shoplifting a Clark bar would equally constitute recidivism for a sex offender under this definition.

Obviously, a researcher can study the prediction of any criterion they choose, or several criteria at once. However, to the extent that research chiefly aims to assisting in implementing Hendricks-type commitment laws, some recidivism definitions are more defensible than others. We have not reviewed each of the seventeen state laws enabling civil commitment of sex offenders. However, based on the three we have reviewed from the standpoint of statutory language and legislative intent (drafting history and legislative floor debate arguments offered by bills’ proponents), namely Minnesota, Iowa, and Kansas, we would argue that the first criterion, namely sex offenses proper, is the most defensible criterion for recidivism prediction. The major concerns driving the enactment of these statutes seem to have been (regardless of the actual empirical facts), that sex offenders are sui generis among criminals, with phenomenally strong propensities to reoffend, obsessive (in the broad sense) drives toward aberrant sexual objects and activities, and near-total inability to control their deviant sexual impulses over the long term. Any idea that, while they were out endangering the community with their marked propensity to rape women and molest children, they might also hold up the nearest 7-11, snatch purses, use illegal drugs, commit stock fraud, or pass bad checks, does not seem to have loomed large in proponents’ reasons for championing civil commitment statutes.
It is quite true that one can find in the literature, and in experts’ courtroom testimony since these laws were passed, prominent arguments that committing sex offenders also protects the community from violent (and non-violent) non-sexual crimes. However, this seems really to have been an afterthought, from the standpoint of legislative intent. Because of this reading of the historical record, we will focus major interest on the ability of clinicians and actuarial risk assessment instruments to predict narrowly defined sexual reoffending. Nevertheless, the other criteria will be examined as well, to the extent that the literature addresses them.

**Assessment of Risk**

Throughout this review, we have used and will continue to use the term “risk assessment” to describe the process of predicting recidivism. While it is possible that one could define risk assessment as the assessment of a hypothetical construct, the disposition to reoffend or risk of reoffending, and therefore a separate venture from the prediction of actual reoffense vs. non-reoffense, we believe that the term should be considered to have the latter meaning, as these assessments do not seem to serve the State’s interest without being able to predict actual reoffending. In equating risk assessment with recidivism prediction, we share ground with the Association for the Treatment of Sexual Abusers (ATSA), the largest American professional organization of individuals who would be expected to routinely perform these evaluations; in their informational package on risk assessment (Hanson, 2000), risk assessment is clearly discussed in terms of recidivism prediction.

**Factors Reported to Predict Recidivism**

The risk assessment literature has identified numerous variables that have been found to predict sex offender recidivism. Some of the seemingly more salient predictors are highlighted in Table 1. Most of the literature on sexual recidivism prediction has dealt primarily with what are
called “static,” or relatively unchanging, factors. Hanson and Bussière (1998), in their meta-
analysis of 61 studies, found that the factor that best predicted sexual recidivism was a phallometric index of deviant sexual arousal. From among the static variables in a sample of 208 sexual recidivists and 201 nonrecidivists, Hanson and Harris (2000) found three static variables (two are listed in table 1; the third was a combination variable including psychiatric history, criminal history, and PCL-R scores) to be significantly associated with recidivism (multiple $R = .40$) and to make a unique contribution to the prediction of recidivism above the contribution of the dynamic variables named below (change in $R^2 = .035$). However, three variables splitting three and one-half percent of added predictable variance is hardly an impressive demonstration of incremental validity; it corresponds to a gain in hit rate of just over 2.3% under the most favorable conditions (i.e., $P = 1/2$), and just over 0.8% additional correct predictions at $P = .134$.

Others have noted that the best predictor of any violence (construed by many sex offender researchers to include all sex offenses; Rice & Harris, 1995; Sjostedt & Langstrom, 2001), is past history of violence (Monahan, 1981), suggesting that those who were more violent before will be more violent henceforth. In a study of 342 sexual offenders, Hall (1988) found that commission of new sexual felonies against adults was associated with various forms of past violence, both sexual and non-sexual (see Table 1). Hall (1988) did not identify a single variable that significantly predicted sexual reoffending against children. The reasons for this failure are not obvious. It is not attributable to a paucity of child molesters, because “previous offenses against children” was a predictor (implying that some subjects were child molesters), and 34 subjects committed new sexual felonies against children. Some other studies have also reported
problems in predicting child molestation, at least of the intrafamilial kind (e.g., Epperson, Kaul, & Hesselton, 1997).

While much of the risk literature has focused on stable variables, more recent studies have examined dynamic predictors, ones more capable of change over time. These are commonly divided into stable and acute dynamic factors (Beech, Fisher, & Thornton, 2003; Hanson & Harris, 2001). Stable dynamic factors are those tending to remain constant unless special effort is made to change them, e.g., in therapy. Acute factors are mercurial and may strongly affect imminent risk.

Stable dynamic factors implicated in sexual recidivism include having significant relationships with individuals rated as negative influences by probation officers, attitudes toward sexual assault (especially feeling entitled to express one’s sex drive and not seeing oneself as at risk; Hanson & Harris, 2000), non-cooperation with supervision, poor sexual self-regulation, and poor general self-regulation (Hanson & Harris, 2001). In a study of 208 sexual recidivists and 201 nonrecidivist sex offenders, Hanson and Harris (2000) found the three most predictive stable dynamic factors (see Table 1) to be more strongly associated with recidivism than were the three most predictive static factors (multiple $R = .53$ versus .40 for the static variables alone); this corresponds to an increase in hit rate of under 2% at $P = .134$. In another study (Thornton, 2002) attitudes supportive of sexual assault were predictive of recidivism for all sexual offenders but were most common in child molesters. Likewise, intimacy deficits, which encompass low self-esteem, emotional overidentification with children, lack of intimacy with adults, and aggressive thinking, have been found at higher levels in offenders who recidivate (Thornton, 2002).
Among the acute factors, or factors prone to change over short spans of time, found to be most predictive of sexual recidivism were access to victims, noncooperation with supervision, and anger (Hanson & Harris, 2000). These variables had a significant association with recidivism (multiple $R = .32$), and made a unique but small contribution to the prediction of recidivism of equivalent magnitude to the most predictive static predictors. Based on their previous study as well as the results from the Hanson and Bussière (1998) meta-analysis, Hanson and Harris (2001) identified several additional acute factors for inclusion on an instrument designed to measure dynamic factors. These additional variables were emotional collapse (defined as a severe emotional crisis or a loss of social support), substance abuse, and sexual preoccupations. Additionally, they included “unique factors,” which are essentially events that idiosyncratically affect an individual offender, seemingly driving him to reoffend (e.g., a painful anniversary, an intercurrent illness, or painful recollections of personal trauma).

<table>
<thead>
<tr>
<th>Study</th>
<th>Predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hanson and Bussière (1998)</td>
<td>Phallometric index of deviant sexual arousal</td>
</tr>
<tr>
<td></td>
<td>Arousal to children</td>
</tr>
<tr>
<td></td>
<td>Arousal to boys</td>
</tr>
<tr>
<td>Rice &amp; Harris (1997)</td>
<td>Deviant arousal (assessed by phallometry)</td>
</tr>
<tr>
<td></td>
<td>Psychopathy</td>
</tr>
<tr>
<td></td>
<td>Interaction of above</td>
</tr>
<tr>
<td>Hanson &amp; Harris (2000)</td>
<td>Static</td>
</tr>
<tr>
<td></td>
<td>Sexual deviance</td>
</tr>
<tr>
<td></td>
<td>IQ score (low)</td>
</tr>
<tr>
<td></td>
<td>Stable Dynamic</td>
</tr>
<tr>
<td></td>
<td>Not seeing self as risk</td>
</tr>
<tr>
<td></td>
<td>negative peer influences</td>
</tr>
<tr>
<td></td>
<td>Sexual entitlement (feeling entitled to express one’s sex drive)</td>
</tr>
<tr>
<td></td>
<td>Acute Dynamic</td>
</tr>
<tr>
<td></td>
<td>Access to victims</td>
</tr>
<tr>
<td></td>
<td>Noncooperation with supervision</td>
</tr>
<tr>
<td></td>
<td>Anger</td>
</tr>
<tr>
<td>Hall (1988)</td>
<td>Prior sex offenses against children</td>
</tr>
</tbody>
</table>
Table 1. Factors Predictive of Specifically Sexual Recidivism

<table>
<thead>
<tr>
<th>Study</th>
<th>Predictor</th>
</tr>
</thead>
</table>
| Thornton (2002)        | Prior nonsexual violent offenses  
                        | Prior sex offenses against adults  
                        | MMPI Mf score (high)  
                        | Sexual deviance  
                        | Attitudes supportive of sexual assault (e.g. believing that women want and/or deserve to be raped)  
                        | Intimacy deficits  
                        | Self-management problems (conceptually similar to impulsivity and measured by Factor 2 of the PCL-R) |

Many studies looking at predictors of recidivism have not identified any relevant acute dynamic factors, which is not entirely surprising given that most of these studies rely on assessments of offenders conducted while they are incarcerated, potentially months or years before they reoffend. Acute dynamic factors are generally hypothesized to predict immediate, rather than long-term, risk for reoffense. Thus, such factors would likely only show a strong association with recidivism if assessment occurred close to the time of reoffense, and so associations will seldom be detected unless the research methodology relies on prospective methodology (Hanson & Harris, 2000), which is potentially very expensive and would be prohibitively expensive to implement as a method of actually tracking offenders in the community. Simply because many studies have not identified important acute factors does not mean that they are not relevant; rather, the study methodology may simply be flawed. However, keeping track of acute dynamic variables in an applied setting may be problematic from a logistical standpoint, given the large workloads of most parole officers and others who might realistically be involved in monitoring released offenders.

Identifying variables that predict sexual recidivism is just the beginning of the risk assessment process. After data on the important variables has been collected, these data must still
be aggregated in some way to make a prediction; data are not self-interpreting, and they very seldom speak unequivocally for one conclusion. As with any other type of prediction, aggregation will occur either through clinical or mechanical means. The exact methods of clinical prediction are by their very nature not well defined in the sexual recidivism literature, as elsewhere. When clinical prediction is studied, it is usually simply stated that clinicians were asked to judge an offender’s likelihood of recidivism (e.g., Hall, 1988; Hanson & Bussière, 1998). One of the few well-specified approaches to clinical prediction described in the literature is the decision chain model (Ward, Louden, Hudson, & Marshall, 1995). This involves analyzing the choices that led an offender to commit an offense, and the situations in which those choices were made. The purpose of the analysis is to identify the circumstances under which the offender is most likely to reoffend, so that the offender can learn to manage these circumstances through therapy. This method is therefore essentially a means of identifying acute dynamic risk factors. Besides the decision chain model, the literature details only a few, fairly similar approaches to clinical prediction (e.g., Finkelhor, 1984). Most methods of clinical prediction in this field are much less well described. Clinical predictions of risk are studied often enough (e.g. Hall, 1988; Hanson & Bussière, 1998; Quinsey & Maguire, 1986) to suggest that the clinical method is quite often used, but is unclear how heterogeneous clinicians’ procedures may be.

Actuarial approaches are covered in the research literature with much greater specificity. As legislatures have begun to require prediction of sex offenders’ future dangerousness for purposes of civil commitment as well as community notification (as community notification laws provided the original impetus for the development of the MnSOST-R; Epperson, Kaul, & Hesselton, 1997), increased research on risk prediction has led to the development of a number of actuarial instruments designed to make, or at least aid in making, such predictions (Beech, et
Among the more frequently cited actuarial instruments are the Rapid Risk Assessment for Sexual Offense Recidivism (RRASOR; Hanson, 1997), the Static-99 (Hanson & Thornton, 2000), the Violence Risk Appraisal Guide (VRAG; Harris, Rice, & Quinsey, 1993), the Sex Offense Risk Appraisal Guide (SORAG; Quinsey, Harris, Rice, & Cormier, 1998), and the Minnesota Sex Offender Screening Tool-Revised (MnSOST-R; Epperson, et al., 1997). The RRASOR, like many of the other actuarial schemes, is designed to predict any recidivism of a sexual nature, not necessarily violent sexual offending, though it has also been used to predict violent recidivism (which would appear to include all sexual offenses). The Static-99 is an extension of the RRASOR that also includes items from another instrument, the Structured Anchored Clinical Judgment (SACJ; Grubin, 1998). The Static-99 comprises only static factors and, like the RRASOR, is designed to predict any sexual recidivism, but can be used to predict violent recidivism. The VRAG was not designed for use exclusively with sex offenders, but rather was intended to predict violent recidivism among violent offenders in general. The definition of “violent” recidivism included all “hands on” sex offenses; 15% of the 618 offenders on whom the instrument was initially developed were sex offenders. The better to predict sex offender recidivism, the SORAG was developed from the VRAG. The SORAG still aims to predict violent recidivism including all contact sexual offenses, but it targets the specific sex offender population. However, it is claimed that the instrument can be used to predict any reoffenses with sexual components. Finally, the MnSOST-R is an actuarial instrument designed to predict sex reoffending among rapists and extrafamilial child molesters (the instrument was not predictive for intrafamilial child molesters). The MnSOST-R contains sections on historical/static variables and institutional/dynamic variables. The items contained on these instruments are shown in Table 2.
<table>
<thead>
<tr>
<th>Item</th>
<th>RRASOR</th>
<th>Static-99</th>
<th>VRAG</th>
<th>SORAG</th>
<th>MnSOST-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Past sexual offenses</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Age at release (younger)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any extrafamilial victims</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any male victims</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior sentencing dates</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noncontact sex offense convictions</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Violent, nonsexual index offense conviction</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior violent, nonsexual convictions</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any stranger victims</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Never married</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Injury to Victim</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Female Victims</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lived with both parents to age 16</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Elementary school behavior problems</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Alcohol problems (drugs for MnSOST)</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Prior nonviolent offenses</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Failure of prior conditional release</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Age at index offense (younger)</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Any personality disorder</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Schizophrenia*</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PCL-R</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Prior violent offenses (including all sex offenses with contact)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Sex offenses only against girls under 14*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Phallometric evidence of deviant arousal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Length of sex offending history**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Sex offense committed while under supervision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Sex offense in a public place</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
### Table 2. Content of Actuarial Risk Assessment Instruments.

<table>
<thead>
<tr>
<th>Item</th>
<th>RRASOR</th>
<th>Static-99</th>
<th>VRAG</th>
<th>SORAG</th>
<th>MnSOST-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Force or threat of force in sex offense</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex offense involving multiple acts on same victim</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of age groups victimized</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Offended against 13-15-year-old victim</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adolescent antisocial behavior</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack of stable employment</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discipline problems while incarcerated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completed chemical dependency treatment while incarcerated*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completed sex offender treatment while incarcerated*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Positive or larger values on items indicate greater risk of reoffense, unless otherwise noted. * - positive value indicates less risk. ** - item has a nonlinear relationship with risk.

Some researchers have recommended using clinical judgment to make the final determination of risk, at least in some circumstances, after considering the output of an actuarial instrument. Epperson, et al. (1997), in their final report on the MnSOST-R, suggest that clinicians can overrule the results of the MnSOST-R, but only when they have a clear rationale for doing so, such as when they note a factor operating in a particular case that has too low a base rate to have been included in an actuarial scheme. The particular “broken leg,” to return to Meehl’s (1954) term, that Epperson, et al. (1997) use as an example, is an offender making a declaration of his intent to reoffend. This is of course likely to be an important, but very rare, predictor of reoffending. The SPJ approach, as discussed above, can also be used in the prediction of recidivism (e.g. Boer, Hart, Kropp, & Webster, 1997), and follows the
recommendation of Quinsey, Lalumiere, Rice, and Harris (1995) that an actuarial instrument anchor the decision process, but differs from Quinsey, et al. in that the latter advise only modifying the statistical prediction under extraordinary circumstances. As it seems that clinicians generally lack skill at identifying the few instances that warrant overruling statistical predictions, there is little empirical warrant for the anchor-and-adjust procedure recommended by SPJ advocates, Epperson, et al., and others.

A number of authors who advocate adjusting statistical predictions recommend adding clinical consideration of dynamic factors to statistical predictions, because actuarial instruments primarily encode static factors. Quinsey, Rice, and Harris (1995) suggest beginning with an actuarial prediction of risk and then modifying its prediction based on clinical weighting of dynamic factors including situational factors, changes in mood, and treatment-induced changes. As noted before, this idea thoroughly confuses sources of data and methods of combining data, returning us to a pre-1954 muddle. From the broad prediction literature, one would expect that adding dynamic factors to actuarial instruments would be preferable to having clinicians make subjective adjustments to actuarial predictions.

Some investigators are beginning to incorporate dynamic factors into statistical predictions, such as the Sex Offender Need Assessment Rating (SONAR; Hanson & Harris, 2001). This includes five stable dynamic and four acute dynamic items. The existence of actuarial instruments that include dynamic factors may ultimately show that it is the dynamic factors themselves, and not the clinical modification of actuarial predictions, that adds to the validity of the actuarials.

Thus far, we have reviewed factors hypothesized or demonstrated to significantly predict recidivism, as well as several instruments designed to combine factors to yield recidivism
predictions. We now examine the validity of such predictions. The few studies comparing
clinical to statistical prediction of recidivism favor the latter. Hall (1988) studied 342 offenders
and found no significant association between clinical prediction of dangerous and commission of
a sexual felony against an adult \( R^2 = .0003, F<1 \), but did find a significant association
between a linear combination of actuarial predictors and this outcome (Wilks’s \( \lambda = 0.590 \), \( F(19
,130)=4.76, R^2 = .41, p<.0001 \)). Hall (1988) found no significant predictors of either sort for the
commission of a sexual felony against a child. The general finding is that actuarial sex reoffense
predictions are more accurate than clinical ones (e.g., Hanson, Morton, & Harris, 2003; Quinsey
& Maguire, 1986). Hanson and Bussière (1998) found clinical risk assessments to be very
modestly correlated with recidivism (average \( r = .10 \) for sexual recidivism, \( .06 \) for nonsexual
violent recidivism, and \( .14 \) for general recidivism). Statistical prediction schemes, on the other
hand, showed more robust associations recidivism (average \( r = .46 \) for any sexual recidivism, \( .46
\) for nonsexual violent recidivism, and \( .42 \) for general recidivism). Hanson, et al.’s (2003) meta-
analysis reported actuarial instruments to be significantly more predictive of sexual recidivism
than unstructured clinical assessments (average Cohen’s \( d = .68 \) for actuarial instruments, \( .28 \) for
clinical judgments).

Empirically-guided clinical judgments – structured clinical judgments in which the set of
factors to be considered is determined in advance, identical for every case, and supported by
empirical literature, but where the clinician makes the final judgment of risk – were intermediate
between the other two types of prediction \( d = .52, .42 \) when outlier [Dempster, 1998; cited in
Hanson, et al., 2003] removed). At least one study shows clinical prediction using an SPJ
instrument, the Sexual Violence Risk-20 (SVR-20; Boer, et al., 1997), outperforming the
numerical scores on the instrument, itself (de Vogel, et al., 2004). This one study, while certainly
an exception to the rule that actuarial schemes are superior or at least equivalent to clinical predictions, does not in and of itself negate the overall finding of actuarial superiority, especially given that the structured professional judgments were less predictive than actuarials on average in Hanson et al.’s (2003) meta-analysis.

There is considerable information on the validity of the actuarial instruments discussed above. A comparison of AUCs for the various instruments, broken down by study, is shown in Table 3. One comparison (Nunes, Firestone, Bradford, Greenberg, & Broom, 2002) found no significant difference between the Static-99 and the SORAG in predicting new sexual or violent charges. The authors of this study also noted that the SORAG, with more varied item content, requires more time to complete than does the Static-99; economic considerations would therefore favor the Static-99, which is easier to complete and demonstrates equivalent or greater validity.

In another comparison, Hanson, et al. (2003) averaged findings from five studies and reported that the SORAG generally outperformed the RRASOR, but was less accurate than the Static-99 in predicting “violent” recidivism. The effect varied by sample, however. The seeming contradiction in Hanson, et al.’s and Nunes, et al.’s results may be at least partially explained by the near total lack of adult rapists in Nunes’ sample (rapists made up 5% of the sample).

Another study (Harris, et al., 2003) compared the VRAG, the SORAG, the RRASOR, and the Static-99. It was reported that the VRAG and SORAG were significantly better at predicting new charges for violent offenses, including sexually motivated offenses, than the RRASOR or the Static-99 among both rapists and child molesters. All instruments predicted specifically sexual reoffending less well, but the relative performances of instruments were stable. Scores on all four instruments were associated in a survival analysis with time until first reoffense ($r = .33$ for VRAG, .33 for SORAG, .23 for the RRASOR, .36 for Static-99).
Moreover, the SORAG and VRAG scores were significantly associated both with severity of a new offense, as measured by a 20-point scale ranging from property offenses to first degree murder, and severity of harm to the victim, as measured by a seven-point scale ranging from no harm to death and mutilation, among those who committed new offenses. Harris, et al. (2003) also examined the predictive validity of the PCL-R, which is included on the SORAG and VRAG, and at phallometric results (included on the SORAG); both were associated with recidivism, and the PCL-R was associated in a survival analysis with time from release to time of first new charge.

The MnSOST-R has been studied less than other actuarial instruments, yielding fewer comparisons to other instruments. In its initial development study (never peer-reviewed), the MnSOST-R reportedly evidenced similar predictive ability to what has been demonstrated for other instruments (Epperson, et al., 1997). An addendum to the technical report outlining the MnSOST-R development briefly describes a 95-subject “cross-validation” (AUC = .76); however, the cross-validation sample was available to Epperson, et al. at the same time as the development sample, and the authors were generally familiar with its nature and characteristics when constructing the MnSOST-R (D.L. Epperson, personal communication to W.M. Grove, June 15, 2001). In one study comparing the MnSOST-R to other actuarial instruments (Barbaree, Seto, Langton, & Peacock, 2001), the MnSOST-R, like the other instruments, was able significantly to predict general recidivism but not “violent” recidivism or sexual reoffending. This was not due to having an unpredictable sample; the VRAG, SORAG, RRASOR, and Static-99 were all able significantly to predict “violent” recidivism and sexual reoffending. Of the other instruments, the SORAG was best at predicting violent recidivism (which included sexual recidivism), while the RRASOR performed better than other instruments in predicting sexual
recidivism. The VRAG best predicted general recidivism. One reason the MnSOST-R did not show as many significant results may have been that it was scored on a smaller sample than other instruments (150 offenders versus 215 for the others), mainly because 59 participants were incest offenders, for whom the MnSOST-R authors recommend the instrument not be used. Six other offenders without MnSOST-R scores were missing four or more items. Beech, et al. (2003) note that the MnSOST-R is harder to score than other instruments, and some of its items presuppose Minnesota conditions, e.g., the ready availability of sex offender and chemical dependency treatment (cooperation with such programs is measured by MnSOST-R items). Barbaree et al. (2001) note the same scoring problems, and had six subjects with unscoreable MnSOST-R protocols and many more with unscoreable items.

<table>
<thead>
<tr>
<th>Study &amp; Criterion</th>
<th>Instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RRASOR</td>
</tr>
<tr>
<td>Hanson &amp; Thornton (2000)</td>
<td></td>
</tr>
<tr>
<td>Sex offenses</td>
<td>.68</td>
</tr>
<tr>
<td>Rapists</td>
<td>.68</td>
</tr>
<tr>
<td>Child molesters</td>
<td>.69</td>
</tr>
<tr>
<td>Violent offenses</td>
<td>.64</td>
</tr>
<tr>
<td>(including sex offenses)</td>
<td></td>
</tr>
<tr>
<td>Rapists</td>
<td>.64</td>
</tr>
<tr>
<td>Child molesters</td>
<td>.66</td>
</tr>
<tr>
<td>Sjostedt &amp; Langstrom (2001)</td>
<td></td>
</tr>
<tr>
<td>Sex offenses</td>
<td></td>
</tr>
<tr>
<td>Rapists</td>
<td>.69</td>
</tr>
<tr>
<td>Child molesters</td>
<td>.76</td>
</tr>
<tr>
<td>Violent offenses</td>
<td></td>
</tr>
<tr>
<td>(including sex offenses)</td>
<td></td>
</tr>
<tr>
<td>Rapists</td>
<td>.59</td>
</tr>
<tr>
<td>Child molesters</td>
<td>.63</td>
</tr>
<tr>
<td>Rice &amp; Harris (1995)</td>
<td></td>
</tr>
<tr>
<td>Violent Offenses</td>
<td></td>
</tr>
<tr>
<td>(including Sex Offenses)</td>
<td></td>
</tr>
<tr>
<td>Rice &amp; Harris (1997)</td>
<td></td>
</tr>
<tr>
<td>Violent Offenses</td>
<td></td>
</tr>
<tr>
<td>(including Sex Offenses)</td>
<td></td>
</tr>
<tr>
<td>Nunes, et al. (2002)</td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Predictive Validity (Areas Under the Curve) of Actuarial Instruments, by Study

<table>
<thead>
<tr>
<th>Study &amp; Criterion</th>
<th>RRASOR</th>
<th>Static-99</th>
<th>VRAG</th>
<th>SORAG</th>
<th>MnSOST-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex Offenses</td>
<td></td>
<td>.70</td>
<td></td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>Violent offenses</td>
<td></td>
<td>.69</td>
<td></td>
<td>.69</td>
<td></td>
</tr>
<tr>
<td>(including sex offenses) Harris, et al. (2003)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex offenses</td>
<td>.59</td>
<td>.62</td>
<td>.65</td>
<td>.66</td>
<td></td>
</tr>
<tr>
<td>Rapists</td>
<td>.56</td>
<td>.59</td>
<td>.64</td>
<td>.62</td>
<td></td>
</tr>
<tr>
<td>Child molesters</td>
<td>.61</td>
<td>.65</td>
<td>.70</td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td>Violent offenses</td>
<td>.56</td>
<td>.63</td>
<td>.73</td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>(including sex offenses)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rapists</td>
<td>.50</td>
<td>.58</td>
<td>.73</td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td>Child molesters</td>
<td>.61</td>
<td>.64</td>
<td>.70</td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>Epperson, et al. (1997)</td>
<td></td>
<td></td>
<td></td>
<td>.77</td>
<td></td>
</tr>
<tr>
<td>Sex offenses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Langton, et al. (2002; cited in Beech, et al., 2003)</td>
<td></td>
<td>.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex offenses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barbarine, et al. (2001)</td>
<td></td>
<td>.77</td>
<td>.70</td>
<td>.61</td>
<td>.70</td>
</tr>
<tr>
<td>Sex offenses</td>
<td>.77</td>
<td>.70</td>
<td>.61</td>
<td>.70</td>
<td>.65</td>
</tr>
<tr>
<td>Violent offenses</td>
<td>.65</td>
<td>.70</td>
<td>.69</td>
<td>.73</td>
<td>.58</td>
</tr>
<tr>
<td>(including sexual offenses)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General criminal offenses</td>
<td>.60</td>
<td>.71</td>
<td>.77</td>
<td>.76</td>
<td>.65</td>
</tr>
</tbody>
</table>

The actuarial risk assessments generally seem to demonstrate moderate predictive validity. Given the varied item content, it is certainly plausible that several of these instruments could be combined into a new mechanical predictive scheme, resulting in improved accuracy.

Indeed, it appears that many clinicians do attempt to combine the actuarials, as clinicians in 15 of the 17 states with civil commitment laws report using multiple instruments (Doren, 2004; cited in Seto, 2005). Seto (2005) attempted to find the optimal combination of four of the most common instruments (VRAG, SORAG, RRASOR, Static-99) in a sample of 215 offenders, using a variety of decision rules to combine the instruments. Ultimately, no combination of instruments was able to improve significantly upon the single best instrument (SORAG for broadly defined
violent recidivism; RRASOR for specifically sexual recidivism). This study, at least, shows little
benefit in attempting to combine actuarial risk assessments.

While the discussion to this point may seem to have shown almost no use for clinical
judgment, it is important to note that clinical judgment is, given the current state of knowledge,
an inevitable component of sex offender risk assessments. There are currently no actuarial
instruments that can identify those factors that may alter an offender’s risk of recidivism. Thus,
the actuarials do not identify the most promising targets for therapeutic intervention. Given that
one of the objects of civil commitment is, ostensibly, to treat dangerous offenders so that they
may safely re-enter the community, clinical judgment continues to serve a vital role in deciding
on appropriate courses of treatment.

The Role of Incremental Validity

The research just reviewed establishes that sex offender recidivism, variously defined,
can be predicted with non-negligible (if far from perfect) validity. In order to improve upon
existing levels of validity, we need to find variables that are (1) not too redundant with existing
measures and (2) as strongly related to the criteria as possible. This neatly describes the domain
of incremental validity (IV).

Incremental validity is the ability to add to predictive validity by adding information
(Wiggins, 1973). Both clinical and statistical predictions can demonstrate incremental validity.
Worthwhile clinical IV is demonstrated when a clinician, already possessed of certain pieces of
data allowing predictions possessing a certain level of validity, improves significantly and to a
pragmatically important degree when a new piece of data is acquired. Worthwhile statistical IV
is demonstrated when a statistical prediction scheme, having a certain level of validity based on
weighing a given set of variables, significantly and practically increases its criterion validity when a new variable is added to the prediction scheme.

A major gap in our knowledge is that we do not fully understand how clinicians performing risk assessments in day-to-day practice actually go about integrating the information available to them to arrive at recidivism risk estimates. We know a fair amount about their outputs in research settings, but precious little about their deliberative processes in real life. Ordinarily, clinicians have access to a good deal more information than enters into statistical prediction instruments. We do not know whether clinicians with a statistical instrument in hand, plus other information, base their risk judgments on the statistical prediction, use an anchor-and-adjust heuristic, largely ignore the statistical prediction, or employ a mix of strategies that varies from case to case. Further, we do not know much about what factors, not represented in the statistical instruments, they routinely (or even more than just occasionally) consider in their deliberations. Finally, we do not know which variables clinicians routinely weight heavily, which they regularly consider but do not give pride of place, which variables they pay attention to but assign only very modest weight, and which ones they essentially ignore. To help dismantle complex clinician decision processes, the field of sexual risk assessment would be greatly aided by conducting appropriate IV studies.

Since the validity of an additional piece of information depends on what information is already in hand, there is no single IV coefficient between a variable and a criterion, comparable to the unique criterion validity coefficient. If numerous pieces of information are considered, the number of distinct IV coefficients explodes combinatorially. However, often there will be one or a few plausible orderings for information acquisition, dictated by cost, risk, or convenience. For example, one does not ask what is the IV of taking a headache patient’s medical history after one
has done an MRI. One does not perform an amniocentesis (with its 0.5% risk of miscarriage; U.S. Department of Health and Human Services Public Health Services Centers for Disease Control and Prevention, 1995) before finding out whether a previously born offspring showed a chromosomal abnormality; instead, one decides from history (and maternal age) whether an amniocentesis is warranted. Therefore, it is sometimes possible to design IV studies involving reasonable numbers of clinical relevant data, without having to present clinicians with myriad shuffled orders of data presentation.

Because clinical IV involves human judgment, it is possible that a clinician or group of clinicians will show negative IV; a certain piece of information may prove attractively misleading, resulting in lowered validity when added to previously available, valid information (Hunsley, 2002). Hence, the study of clinical IV can identify and help root out such red herrings. Another advantage of clinical IV studies is that it can identify the “point of vanishing returns,” i.e., an order and number of variables past which there is little or no gain obtained from doing more tests, poring over more reports, or conducting more interviews. The relevance of such studies to maximizing the cost effectiveness of clinical assessments should be crystal clear.

On the other hand, statistically optimized prediction schemes can never perform worse by adding an additional variable; if the variable does not correlate with the criterion and does not interact with other predictors (e.g., a suppressor variable; Horst, 1941), it will automatically get zero weight. Statistical IV essentially determines how clinicians should be weighting various measures, if they were to mimic the actuary.

Despite the potential contribution IV studies can make to prediction research, such studies are seldom conducted (Hayes & Lench, 2003). IV studies have been quite useful in understanding and streamlining assessment procedures, as well as furnishing valuable
admonitions. In an early classic study, Sines (1959) studied IV in the description of personality and psychopathology in psychiatric outpatients. His results showed that commonly-used measures, including interview data, the MMPI, and the Rorschach, offered no or very limited incremental validity over quick, cheap biographical data sheets and even over stereotypes clinicians held about typical patients at the clinic. Indeed, the Rorschach sometimes showed negative IV.

Other studies have demonstrated positive IV for certain measures. Measures of perfectionism added validity to stress indices in predicting suicidal behavior (Chang & Rand, 2000). Additionally, adding normal-range personality measures led to increased validity over usual assessments in predicting alcoholics’ relapses (Cannon, et al., 1997). These are just a few contributions of IV studies to the assessment literature.

We have found no clinically representative IV studies of sexual recidivism risk assessments. That is, there appears never to have been a risk assessment that employed IV methods to determine the IV of each piece of information contributing to the risk judgment. However, there have been limited reports regarding the IV of some measures used in predicting sexual recidivism. For example, Rice and Harris (1997) showed that Psychopathy and sexual deviance not only added incrementally to each other, but also interacted to predict recidivism more accurately than would be predicted by the sum of their zero-order criterion correlations. On the minus side, offenders’ performance in treatment did not improve prediction from Static-99 scores in predicting recidivism (Peacock & Barbaree, 2000; cited in Sjostedt & Langstrom, 2001).

A number of studies have shown that dynamic factors can add incrementally to the predictive ability of actuarial instruments, although the increments are often only a few percent
of variance accounted for. Batteries designed to measure dynamic risk factors, such as the Sex Offender Treatment Evaluation Project (STEP; Beech, Fisher, & Beckett, 1999) test battery, which looks at level of deviancy in child molesters, have been shown to add incremental validity over the Static-99. The Initial Deviance Assessment (IDA) Structured Risk Assessment model (Thornton, 2002), which measures four dynamic factors, has also shown IV over the Static-99. Hanson and Harris’s (2001) SONAR was able to add incrementally to both the VRAG and Static-99 in distinguishing recidivists from non-recidivists. Hanson and Harris (2000) also showed that dynamic variables, especially stable dynamic variables, could improve predictive accuracy beyond VRAG scores and other static variables. Certainly, a number of variables have been shown to add incrementally to measures already in use for sexual risk assessments, but a more thorough IV study remains to be done to determine which of the many elements used in risk assessments contribute unique incremental validity.

The field of sexual recidivism risk assessment has much to benefit from the application of IV methodology. As stated by Hunsley (2003), incremental validity is essentially a measure of clinical utility. Thus, IV methodology can determine which measures demonstrate worthwhile amounts of incremental validity, and these measures can be kept, studied, and improved upon, while eliminating other measures that take time and money while delivering either misleading (inaccurate) or redundant (not new) information. With the time and money saved, more assessment measures that do work well can be added to the risk assessments, or the dividends can be reinvested elsewhere, e.g., on treatment or prevention efforts. Clinicians are being called on to perform risk assessments on a great number of sex offenders, and improving the efficiency of these assessments—without sacrificing quality—would be real financial boon to the state. An even more vexing problem, in the long run, which could be ameliorated by good incremental
validity studies, is the all-too-high error rates of risk assessments. IV methodology can help determine whether some measures, when employed jointly with other information, are deterring from the accuracy of risk assessments. These findings can form the basis for decisions on whether to remove such measures from assessments, or to try to improve the measure and test its incremental validity again later on if the negative IV is small, based on the content or validation history of the measure. Given the very high cost of inaccurate predictions, either in decreased public safety or unnecessary deprivations of liberty (with attendant long-term very expensive incarceration and intensive treatment, taking up beds that perforce are not being used for someone who really is dangerous to the community), any research that can provide data to lessen the inaccuracy of risk assessments as they are commonly conducted—in a context of multiple pieces of information being integrated by a clinician is very important indeed. This methodology can begin to fill in some of the holes in our knowledge of the risk assessments used to predict sexual recidivism.

References


*Frye v. United States*, 293 F. 1013, 1014, D.C. Cir. 1923.


