Psychology, Public Policy, and Law
2007, Vol. 13, No. 3, 167–203

1076-8971/08/$12.00 DOI: 10.1037/1076-8971.13.3.167

POOR DIAGNOSTIC RELIABILITY, THE NULL-BAYES LOGIC MODEL, AND THEIR IMPLICATIONS FOR SEXUALLY VIOLENT PREDATOR EVALUATIONS

Richard Wollert
Vancouver, Washington

Forensic psychologists have approached sexually violent predator (SVP) civil commitment evaluations from the position that respondents must be positive for a condition from the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) of the American Psychiatric Association to be classified as SVPs. The only research on DSM diagnostic reliability in SVP cases has been undertaken by J. S. Levenson (2004a) and R. L. Packard and J. Levenson (2006). Although Packard and Levenson claimed that diagnostic evaluations in SVP cases were highly reliable, a reanalysis of their data indicated otherwise. Further, high levels of diagnostic uncertainty were found for a proposed paraphilia referred to as paraphilia not otherwise specified–nonconsent. Diagnostic criteria used to identify paraphilias among SVP respondents are therefore characterized by poor reliability. Logic models that were previously used to determine diagnostic confidence are also obsolete. Recommendations for improving diagnostic reliability are discussed, and the Null-Bayes Logic Model (NBLM) is proposed as a method for reaching certainty opinions that is superior to past models based on unrestrained clinical judgment. The implications of the present results and the NBLM for future practice, research, and policy directions are discussed.

Keywords: Null-Bayes Logic Model, Bayes's Theorem, sexually violent predator evaluations, diagnostic reliability, DSM–IV–TR

Many states in the United States have enacted legislation allowing for the post-prison civil commitment of "a small but extremely dangerous group of sexually violent predators" (SVPs; Woodworth & Kadane, 2004, p. 221; see also Covington, 1997; Doren, 2002; Miller, Amenta, & Conroy, 2005). As the first stage of this process, incarcerated offenders thought to meet the commitment standards are referred by end-of-sentence review boards to prosecutors for commitment consideration. Prosecutors then decide whether commitment petitions should be filed. After this, the courts determine whether commitment proceedings should be instituted because probable cause exists that the respondents to these petitions are SVPs.

In connection with probable cause determinations and commitment trials, it is required that "potential SVPs undergo evaluation" (Jackson, Rogers, & Shuman, 2004, p. 116), so forensic psychologists are often hired by both defense and

Richard Wollert, Independent Practice, Vancouver, Washington.

I am indebted to Jacqueline Waggoner, Tom Zander, Fred Berlin, and Ted Donaldson for their comments on drafts of this article.

The address for my website is www.richardwollert.com.

Correspondence concerning this article should be addressed to Richard Wollert, 602 East 31st Street, Vancouver, WA 98663. E-mail: rwwollert@aol.com

prosecution attorneys in a typical commitment proceeding for the purpose of determining whether the SVP construct (Donaldson & Wollert, in press; Jackson et al., 2004) may be applied to a respondent because he satisfies each of the three prongs that define it. The first of these prongs is that he has been charged with or convicted of one or more sexually violent crimes. The second is that he suffers from a statutorily defined mental abnormality or diagnosed mental disorder consisting of an acquired or congenital condition, construed by most experts as a diagnosis from the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text revision; DSM–IV–TR; American Psychiatric Association, 2000),[1] that impairs his volitional self-control and makes him a sexual danger to others. The third is that he will likely commit more acts of predatory sexual violence in the future because of his mental abnormality. The uniformity of these or equivalent requirements across states with SVP laws has been discussed by Miller et al. (2005, p. 31), who observed that,

> The criteria commonly found in state SVP commitment laws closely resemble those set forth in *Kansas v. Hendricks* (1997). Such laws generally include four elements: (i) a history of sexual offenses, (ii) a mental abnormality, (iii) volitional impairment, and (iv) as a result of mental abnormality, the individual is likely to engage in acts of sexual violence. The reader is referred to Table 1 (pp. 32-36) for a review of statutory definitions across the 16 states with relevant civil commitment laws. Although variation exists among state definitions, the similarity of these definitions to those accepted by the Supreme Court as constitutional is clear.

This overview indicates that the second and third prongs of the legally specified formula for identifying SVPs contain both conjoint and causal components (Jackson et al., 2004; Schopp, 1998). Conjointly, they require that respondents cannot be classified as SVPs unless they are positive for a DSM diagnosis, volitionally impaired, disposed to being a sexual danger, and likely to sexually recidivate. Causally, they require that these conjoint elements be interrelated so that a DSM diagnosis produces both impaired control and sexual dangerousness and that the mental abnormality grounded in this triad induces a high risk of recidivism.

Taken together, the foregoing conjoint elements and causal mechanisms form the legal theory of the SVP construct that is illustrated in the top half of Figure 1. This is a very complex theory in that it requires seven tests for its confirmation

---

[1] Some states (e.g., IA) do not require that a mental abnormality be conditioned on a DSM diagnosis. The Supreme Court has also ruled that state legislatures "retain considerable leeway in defining the mental abnormalities . . . that make an individual eligible for commitment" (*Kansas v. Crane,* 2002, p. 6). Nonetheless, as this article emphasizes, the ethical guidelines that govern the practice of expert witnesses are often more restrictive than the letter of the law. Psychologists are expected to adhere to whatever ethical prescriptions are applicable when this is the case. Regarding the issue of diagnostic classification and other issues, in particular, the ethics code promulgated by the American Psychological Association (2002) states that "psychologists' work is based upon established scientific and professional knowledge of the discipline" (p. 5). By extension, experts in forensic psychology who classify offenders in the course of their work are obligated to use established diagnostic categories, such as those in the DSM or an alternative set of psychological concepts, that have been operationalized and validated. Faced with these options, the great majority of psychologists rely on DSM diagnoses (Prentky et al., 2006; Trowbridge & Adams, 2006–2007).

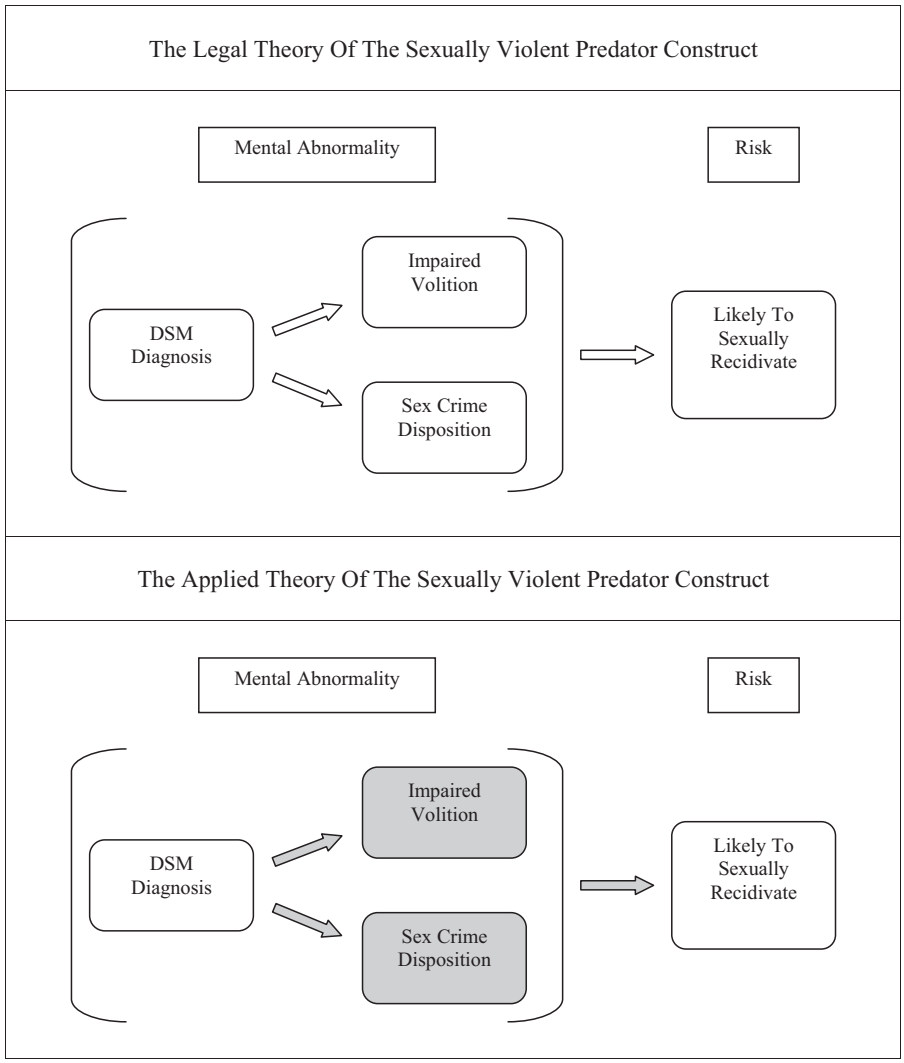

Figure 1. The legal and applied theories of the sexually violent predator construct. A full evaluation of all aspects of the legal theory in the top box would test for the presence of each concept in the rectangles with rounded edges and for the causal relationships specified by the arrows. Shaded symbols in the bottom box represent concepts and relationships that are inferred when the applied theory is used.

(one for each of the rectangular components with rounded edges and one for each of the causal arrows that connect the rectangles in Figure 1). Were it proposed by a group of psychologists, other psychologists would almost certainly insist that advocates for the theory define all of their concepts, develop methods for measuring them, and carry out research confirming that measurement methods were reliable and that the concepts of interest were causally related to one another (Prentky, Janus, Barbaree, Schwartz, & Kafka, 2006). Considering that decisions based on the theory might result in lengthy or life-long confinement (Janus &

Meehl, 1997; Wollert, 2002), accurate identification of SVPs would also seem to require the eventual validation of a taxonomic system specifying the presumably finite set of categories into which SVPs could be subdivided (Prentky et al., 2006).

Scant progress has been made on scientific validation of the SVP construct as a whole in the many years that have passed since the first SVP laws were passed in the early 1990s. Although widespread agreement exists that an *acquired or congenital condition* means a DSM diagnosis and that *likely to recidivate* means a chance of reoffending in excess of a standard, such as 50%,[2] terms like *volitional control* and *sexual dangerousness* remain shrouded in ambiguity (Jackson et al., 2004; Janus, 2001; Mercado, Schopp, & Bornstein, 2005; Miller et al., 2005; Prentky et al., 2006). No evidence has been published that verifies any of the causal linkages in the top box of Figure 1 and, contrary to this formulation, the developers of the DSM have explicitly warned those who use the diagnostic manual that "a diagnosis does not carry any necessary implications regarding the *causes* of the individual's mental disorder or its associated *impairments*" (American Psychiatric Association, 2000, p. xxxiii, emphasis added). Finally, in the absence of a taxonomic system that turns on normative comparison, SVP classification decisions are made largely on the basis of an ad hoc and "case-specific" approach (Hanson & Morton-Bourgon, 2007, p. 3) that is inconsistent with the unifying goal that characterizes scientific endeavor.

Forensic psychologists who offer expert witness testimony in SVP cases face a troubling dilemma. On the one hand, they are charged with providing a reliable perspective that may help the trier of fact reach a just verdict in relation to specific respondents. On the other, the construction of this perspective is hampered by the limited knowledge and conceptual clarity that have been discussed in relation to the legal theory. From reading hundreds of SVP evaluations and participating in over 70 SVP proceedings in six states, it is my contention that a large percentage of SVP experts have implicitly resolved this problem by simplifying the legal theory in two respects. First, they reach an opinion on a respondent's diagnostic status and his risk level because these concepts are defined most clearly and some assessment criteria have been formulated for both the former (American Psychiatric Association, 2000) and the latter (Barbaree, Seto, Langton, & Peacock, 2001; G. T. Harris et al., 2003; Jackson et al., 2004; Langton et al., 2007). Then, if the results of these operations fit the SVP formula, they are combined with other

[2] Although *likely* is defined as more likely than not in most states, some states have adopted a different terminology. In California, for example, likely to engage in acts of sexual violence is interpreted as meaning a serious and well-founded risk. Furthermore, an evaluator who concludes that a respondent meets this standard but recommends against commitment solely on the grounds that he or she cannot conclude that the "person is *more likely than not* to reoffend" is considered to have misapplied the California SVP statute (*People v. Superior Court*, 2002, p. 968; emphasis in the original). This decision does not rule out all quantitative definitions of the California standard, but it does rule out one that is frequently encountered elsewhere. However, the California standard might not be a significant aspect in many cases in that the third prong that defines an SVP is conditioned to a great extent on the presence of a mental abnormality, which, in turn, is conditioned on the presence of a DSM diagnosis. From this perspective, a respondent who is negative for a DSM diagnosis will be negative for a mental abnormality and will therefore be ineligible for classification as an SVP, regardless of whether he is a serious and well-founded risk or more likely than not to recidivate.

information to infer where the respondent stands on the remaining aspects specified in the legal theory. This decision-making strategy, which might be referred to in everyday language as "nail the first, nail the last, do your best with the rest," is consistent with Doren's (2002, p. 24) observation that "the basic referral questions for evaluators involve diagnostic and risk assessment considerations" and with the empirically based conclusion of Jackson et al. (2004, p. 122) that "for psychologists . . . lack of volitional control appeared less important in making their SVP recommendations than ratings of future sexual violence . . . or the presence of a mental abnormality."

This "applied theory," which turns on the two decisions represented as the unshaded rectangles in the lower box of Figure 1, provides an explicit framework that judges and juries may use to track the logic behind the conclusions that experts present. It also points out three different classes of respondents who, from the standpoint of experts relying on the applied theory, clearly cannot be classified as SVPs. The first class, illustrated in the top box of Figure 2, includes those who cannot be assigned a relevant DSM diagnosis with a high degree of certainty and therefore cannot be positive for any of the characteristics that must, by statute, be caused by a diagnosis. The second, in the middle box, includes those who cannot be classified as likely recidivists and, by inference, are able to control their sexual behavior to a substantial extent and do not constitute a sexual danger because of this level of control. The third, in the bottom box, includes those who fall in both of the previous classes.

As this analysis suggests, the assumption that the applied theory can efficiently classify respondents as SVPs and non-SVPs is predicated on the further assumptions that the measurement processes underlying the assignment of DSM diagnoses are highly reliable and that risk assessment methods are valid for the prediction of sexual recidivism. Regarding the latter issue, many researchers have examined the validity of either clinical or actuarial approaches for the prediction of sexual recidivism (Abracen & Looman, 2006; Barbaree et al., 2001; Bartosh, Garby, Lewis, & Gray, 2003; Dix, 1976; Doren, 2004; Epperson, Kaul, Huot, Goldman, & Alexander, 2003; Hall, 1988; Hanson, 2006; Hanson & Morton-Bourgon, 2007; Hanson & Thornton, 2000; G. T. Harris et al., 2003; Kahn & Chambers, 1991; Langton et al., 2007; Mossman, 2006; Nunes, Firestone, Bradford, Greenberg, & Broom, 2002; Quinsey, Harris, Rice, & Cormier, 1998; Seto, 2005; Sjostedt & Langstrom, 2001; Smith & Monastersky, 1986; Sturgeon & Taylor, 1980; Wollert, 2006). As a result, experts who use these approaches should be able to inform the court of their "decision threshold operating characteristics" (Mossman, 2006; Swets, Dawes, & Monahan, 2000, pp. 6–8) and "error rates" in a manner that is sufficiently transparent for the court to determine the admissibility of opinions based on them (Prentky et al., 2006; Wollert, 2006).

In contrast to the large body of knowledge that has been compiled on risk estimation, almost no research exists on the extent to which DSM diagnoses may be reliably assigned in SVP cases. It is critically important, however, that experts retained to evaluate SVP candidates have the capacity to reliably diagnose the presence of DSM disorders associated with a statutorily defined mental abnormality (Marshall, 2006; Prentky et al., 2006; Trowbridge & Adams, 2006–2007; Zander, 2005). Otherwise, almost all respondents would fall in the top box of Figure 2. If this were the case, it would rarely be possible to confidently reject the

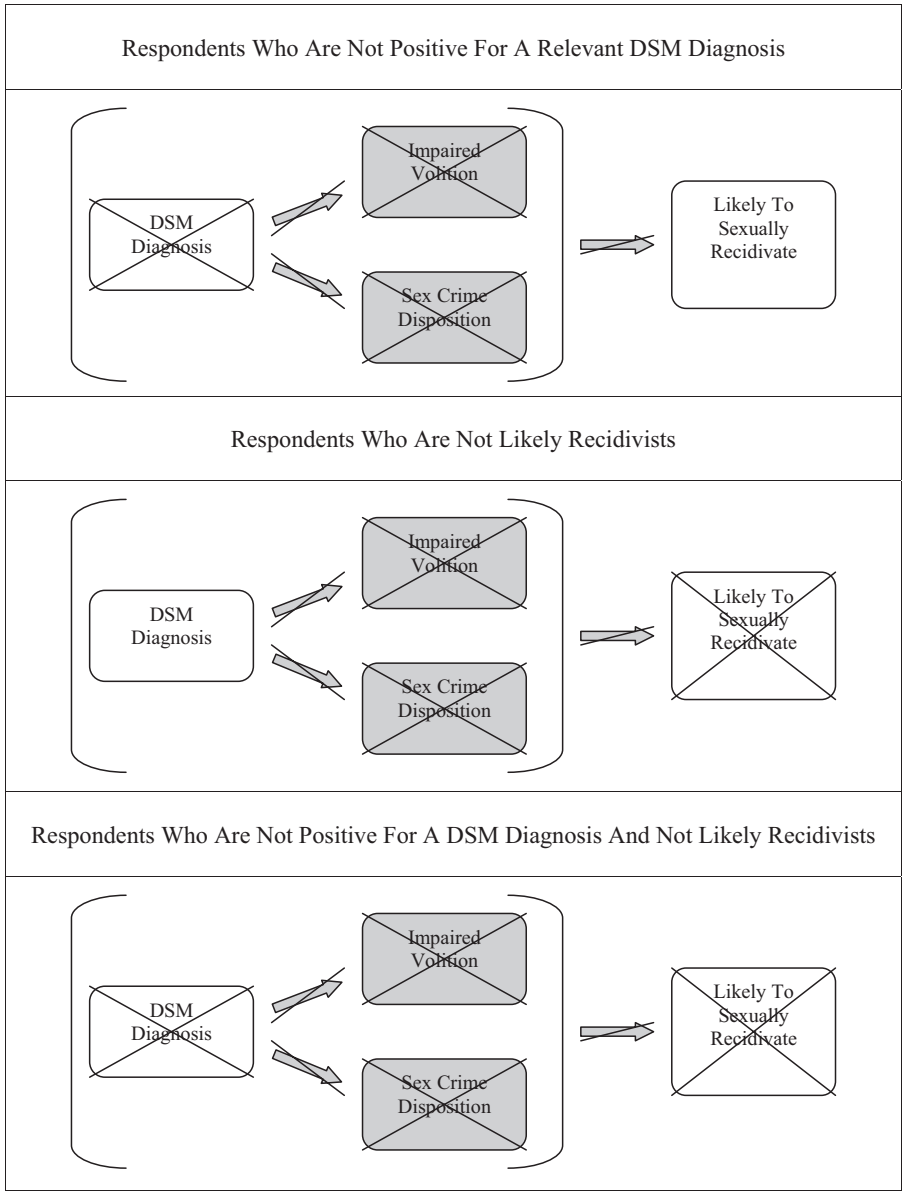

Figure 2. Three classes of respondents who would not qualify as sexually violent predators. Unshaded symbols that are crossed-out stand for an absence of positive measurements. Shaded symbols that are crossed-out stand for negative inferences based on the absence of positive measurements. DSM = *Diagnostic and Statistical Manual of Mental Disorders*.

null hypothesis (Donaldson & Wollert, in press; Wollert, 2006) that has, in so many words, been set forth by the United States Supreme Court as the point where evidence "is sufficient to distinguish the dangerous sexual offender whose serious mental illness, abnormality, or disorder subjects him to civil commitment from the

dangerous but typical recidivist convicted in an ordinary criminal case" (*Kansas v. Crane,* 2002, p. 5; also see *Kansas v. Hendricks,* 1996, p. 360).

Rejection of the null hypothesis in SVP proceedings is further complicated because the conjunctive probability of encountering a specific diagnosis in combination with such vague concepts as impaired self-control and/or sexual dangerousness is virtually certain to be lower than the probability that different raters will agree on the presence of the diagnosis by itself. This disquieting fact is deducible from the product law of probability. According to this law, the probability that two sentences, A and B, are both true is $P(A \cap B) = P(B \mid A) \times P(A)$, where P(A) is the probability or level of certainty (LOC) that Sentence A is true and P(B | A) is the probability or certainty that Sentence B is true when Sentence A is true (McCall, 1975; Woodworth, 2004). Because a probability may not exceed 1, the product law means that the conjoint probability that Sentence A and Sentence B are both true cannot exceed the probability that Sentence A is true. For example, suppose Sentence A is that a given "respondent is positive for paraphilic diagnosis D" and Sentence B is that the "respondent is positive for volitional impairment V." Under these circumstances, the probability that the respondent is positive for both a specific paraphilic diagnosis and a volitional impairment—a conjunction that is necessary to find that he suffers from a mental abnormality—cannot be greater than the probability that he is positive for the diagnosis itself. DSM reliability must therefore not only be high but very high if the conjunctive requirements set forth in SVP statutes are to be satisfied.

These considerations indicate that determining the diagnostic reliability of experts who undertake SVP evaluations is of great importance for taking the applied theory out of the realm of speculation and estimating the range of cases over which it might be useful for identifying SVPs. Noting that state-hired evaluators in Florida customarily relied on various DSM diagnoses that may be unreliable to determine whether civil commitment respondents are SVPs, Levenson (2004a) employed Cohen's (1960) kappa coefficient to measure the extent to which 25 evaluators agreed with respect to their diagnostic conclusions and their civil commitment recommendations concerning 295 detainees. In an article on her research published in *Law and Human Behavior*, she reported that "the inter-rater reliability of 8 DSM-IV diagnoses . . . was found to be poor to fair (kappa = .23 to .70)." She also indicated that "the recommendations made by evaluators regarding whether or not to refer a client for civil commitment demonstrated poor reliability (kappa = .54)" (p. 357).

The results of this line of research have not only cast doubt on the diagnostic reliability of almost all paraphilic categories in the DSM (Zander, 2005) but have caused one long-time leader in the field of sex offender treatment and research to conclude that "Levenson's (2004) findings are an indictment . . . of the SVP civil commitment process in so far as these processes rely on the accuracy of diagnoses" (Marshall, 2006, p. 35). In the wake of such mounting concerns, Packard and Levenson (2006) recently reanalyzed Levenson's (2004a) dataset on the grounds that kappa coefficients may be misleading under some conditions. In the course of doing so, they calculated such descriptive measures as the odds ratio, the relative risk ratio, the LOC for the presence of a diagnosis when the same respondent was evaluated by two clinicians (positive predictive value [PPV]), and the LOC for the absence of a diagnosis under the same condition (negative

predictive value [NPV]). Their statistical analysis, however, focused primarily on testing the "null hypothesis that . . . raters are independent" with respect to the proportion of the time they agreed on the presence (PA^+) or absence (PA^-) of a diagnostic condition (Packard & Levenson, 2006, p. 5).

Obtaining significant chi-square values for each dual-rater contingency table associated with each of the diagnostic categories they studied, Packard and Levenson (2006, p. 14) concluded that Levenson (2004a) was wrong in her original assessment and that "SVP civil commitment evaluation appears to be a highly reliable process." As descriptive evidence for this conclusion, they pointed to the values they calculated for PA^+, PA^-, PPV, and NPV. It was further observed that "forensic evaluators testifying in these matters should be aware of the concerns outlined here, and capable of fully informing the court about the complexities of interpreting diagnostic reliability" (Packard & Levenson, 2006, p. 14).

Forensic evaluators owe these researchers a debt of gratitude for placing valuable supplemental information in the public domain. Two flaws undermine their research, however. One is that rejection of the null hypothesis does not validate the alternative hypothesis that SVP evaluations are highly reliable (Denis, 2001; Wright, 2006). The other is that the measures of interrater agreement that Packard and Levenson (2006) calculated did not control for the expectations that experts held for encountering these conditions (Mossman, 1994a, 1994b). Neglecting to control for this possibility is a serious problem because high levels of interrater certainty, as captured by PPVs and NPVs, do not necessarily reflect high levels of diagnostic reliability. They may, instead, merely reflect the assumptions that evaluators hold about the prevalence rates, symbolized hereafter as P(D), for the disorders that they believe are associated with the SVP construct.

Wollert (2006, pp. 59–60) presented a couple of examples that illustrate the interrelationships between these concepts. In the first, a PPV (symbolized by Wollert as *E*) of 52% was obtained when experts applied a modestly accurate test for the prediction of sexual recidivism to high scoring offenders who supposedly came from a population with an anticipated prevalence or "base" rate for recidivism of 25%. In the second, a PPV of 31% was obtained for the same test and the same score group when it was assumed that they came from a population with a base rate of 12.5%.

High PPVs will therefore be found among experts who collectively assume that they will frequently encounter a given set of paraphilias. When their expectations are overstated, PPVs may be high even when diagnostic reliability is low. Unfortunately, PPVs that are based on this set of circumstances convey only the illusion of certainty.

If the assignment of DSM diagnoses were associated with poor reliability, the confidence in the diagnostic conclusions of state-hired experts, and thus the integrity of current SVP evaluations, would be called into question. Furthermore, if the presumptive base-rate expectations of these evaluators for encountering paraphilic diagnoses were found to be very high, it would be reasonable to collect additional base-rate information for some of these categories and reestimate the PPV LOCs associated with this new information. Then, the possibility that state evaluators were overconfident in their conclusions could be assessed by comparing the PPVs reported by Packard and Levenson (2006) with the reestimated

PPVs. Such comparisons would also be helpful for determining the extent to which subjectively estimated levels of expert certainty might differ from those that are estimated objectively.

Both the presence of unjustified certainty and the extent of its presence may be studied by two applications of a formula for determining conditional probabilities that is known as the odds ratio version of Bayes's Theorem (Dawid, 2002). The "illusion of certainty hypothesis," for example, may be evaluated in three steps. The first is to calculate the positive likelihood ratio (LR), which is the quotient of $PA^{+}/(1 - PA^{-})$, for each diagnosis studied by Packard and Levenson (2006).[3] The second is to determine the preevaluation probabilities—that is, P(D)—that evaluators held for encountering each diagnosis. The third is to inspect the values of LR, P(D), and PPV for each diagnostic option.

A different formulation is required to evaluate the extent to which the illusion of certainty might apply to any given diagnosis. The first step in this application is to collect information that specifies the base-rate percentage of sex offenders who are positive for the diagnosis being evaluated. The second is to reestimate the diagnostic PPV by combining this new information with the diagnostic LR calculated in the preceding paragraph. The third is to compare the original PPV with the reestimated PPV.

Carrying out the first analysis, I discovered that many diagnoses assigned in SVP evaluations were characterized by low reliability coefficients, elevated P(D)s, and PPVs that approximated the P(D)s. Because this pattern appeared so often, I carried out a second study in which a panel of sex offender treatment providers were surveyed to appraise the possibility that the presumptive base rate for one of the supposed paraphilic diagnoses—referred to as paraphilia not otherwise specified–nonconsent (PNOSN)—might be overstated. The PPV reported by Packard and Levenson (2006) for paraphilia not otherwise specified (PNOS) was reestimated as part of this study after it was found that the base rate from the survey was 11 standard deviations below the presumptive base rate.

Both of these studies are presented in the next two sections of the article. In light of their results, it is concluded that the current range of application of the applied theory must be very narrow and that its expansion will require significant improvements in diagnostic reliability. Recommendations for improving diagnostic reliability are also discussed, as is a Bayesian logic model that is more objective and internally consistent for evaluating a respondent's fit with the applied theory than are the unspecified logic models that have been relied on in the past.

---

[3] As Packard and Levenson (2006) indicated, PA^{+} and PA^{-} are analogous to the concepts of sensitivity and specificity, respectively. Sensitivity, within the context of their research, was equal to the proportion of all offenders who, having been classified by one evaluator as being positive for a specified diagnosis, were assigned the same diagnosis by a second evaluator. Specificity was equal to the proportion of all offenders who, having been classified by one evaluator as being negative for a specified diagnosis, were not assigned this diagnosis by a second evaluator.

Study 1

Sample

The following passage from Packard and Levenson (2006, p. 7) describes the sample on which data were originally collected by Levenson (2004a):

> The sampling frame included all 450 male, adult, competent, convicted sex offenders in Florida prisons who received face-to-face evaluations by psychologists or psychiatrists for SVP civil commitment between July 1, 2000 and June 30, 2001. In 295 of those cases, the sex offender was assessed by two independent forensic evaluators, generating the current purposive sample. A total of 25 evaluators were involved in the examination of the subjects, and 88% of the evaluators were male. The evaluators were all licensed psychologists or psychiatrists, possessing a Ph.D. (76%), Psy.D. (16%), or M.D. (8%). Their experience assessing and/or treating sex offenders prior to be being hired to conduct SVP evaluations under Florida's Jimmy Ryce Act ranged from zero to eighteen years ($M = 6$ years). The evaluators "worked for" neither the state nor the defense; they were hired by a private agency independently contracted by the Florida Department of Children and Families.
>
> The mean age of the sex offenders who were evaluated was 41 years, and they had, on average, an 11th grade education. Nearly half (47%) belonged to a racial or ethnic minority group. Only 14% of the sample was currently married. About 5% of subjects had no diagnosis. About 21% had one diagnosis, 37% had two diagnoses, and 36% had three diagnoses or more. Rapists, who were defined as having victims who were all over the age of eighteen, comprised 23% of the sample. Child molesters whose victims were under age 18 comprised slightly less than half of the sample (45%). Mixed offenders (30%) had both adult and minor victims.

Data Analysis and Examples of Calculating LRs and Base-Rate Expectations

LRs bearing on the presence of each diagnosis studied by Packard and Levenson (2006) were calculated as the first step of the reanalysis.[4] As Swets (1988) has observed, calculation of the LR controls for conflation between expected base rates and diagnostic accuracy. In addition, when the PPV and the LR are known for a diagnosis, the subjective preevaluation base-rate expectation that one or both evaluators hold for encountering the diagnosis of concern, symbolized as P(D), may be determined in a few simple operations derived from the odds version of Bayes's Theorem (Dawid, 2002).[5] Finally, when construed

[4] Although the methods set forth in this article may be used to determine the reliability with which SVP evaluators are able to diagnose the absence of DSM disorders, this issue is not the article's focus. The reason for this is that, per the null hypothesis, commitment candidates are regarded as being condition free until evidence proves otherwise.

[5] In general, Bayes's Theorem (Bayes, 1764) is a tool for assessing the probability that a theory—for example, that a person with heart disease will die in 5 years—is true when considered in light of the diagnostic accuracy (i.e., LR) of some piece of evidence, such as a disease criterion or a test score, and what is known about the overall, or base rate, probability of the focal outcome P(D) (D stands for death in this one example; in all other sections of this article it stands for diagnosis). A great number of texts and treatises have been written on Bayesian analysis (Fienberg,

as reliability coefficients, LRs are readily interpretable and hold implications that are easily quantified. Regarding the first issue, for example, an LR that is greater than 1 shows that interrater reliability exceeds chance for the presence of a diagnosis, and diagnostic procedures associated with large LRs (e.g., in the range of 6 or higher) are much more reliable than are diagnostic procedures associated with small (e.g., 1 to 2) LRs (Biggerstaff, 2000). Regarding the second, a PPV associated with an inaccurate P(D) may be revised by combining a more accurate estimate of P(D) with an LR that may be assumed to be stable.

Table 1 provides a worksheet that calculates (a) the LRs for the diagnostic criteria studied by Levenson (2004a) and Packard and Levenson (2006) on the basis of how their subjects were classified by each of two evaluators, and (b) the probabilities that reflect the P(D)s of the evaluators who participated in that research. It also applies these procedures to two diagnostic examples, substance use disorder and other mental illness.

Although Table 4 from Packard and Levenson (2006, p. 12) indicated that evaluators were about 73% certain that their diagnoses for each of these two disorders were correct (see row 1 of Table 1 from this article), the LR for the first condition (1.24, from row 5) was much smaller than the LR for the second condition (8.29, also from row 5). Furthermore, the P(D) anticipated by evaluators for substance use disorder was almost the same as their postevaluation LOC (67% vs. 72%, from rows 1 and 8). This pre–post difference was much larger, however, for other mental illness (25% vs. 73%). Substance use disorder therefore fit the pattern of illusory diagnostic certainty because it lacked reliability (in the sense of being associated with a low LR), put evaluators in the position of having to assume a base rate for the diagnosis that was so high as to be implausible, and added little information beyond the assumed base rate when criteria for the condition were applied to a respondent. In contrast, clinicians were able to use the criteria for other mental illness in a way that was more reliable and resulted in a

---

2006), and many statisticians (e.g., de Finetti, 1964; Jaynes & Bretthorst, 2003; Jeffreys, 1939; Ramsey, 1931; Savage, 1954) have elaborated a mathematical foundation that supports its use as a robust method of scientific inquiry. Several issues related to SVP evaluations have also been analyzed from a Bayesian perspective (Donaldson & Wollert, in press; Janus & Meehl, 1997; Mossman, 2006; Wollert, 2006). The odds version of Bayes's Theorem consists of three formulas that combine LRs with the odds ratios that correspond to P(D) and PPV. The basic formula for determining the LOC that may be placed in a theory on the basis of a given set of evidence is as follows: PPV odds (i.e., PPV when it is converted into an odds ratio) = base-rate odds (i.e., P(D) when it is converted into an odds ratio) multiplied by LR (which needs no conversion because it is already an odds ratio). Two other formulas follow from this basic formula: LR = PPV odds/base rate odds, and P(D) odds = PPV odds/LR. PPV and P(D), when they are used for the purpose of calculation, must be converted into odds ratios before these formulas may be used. In addition, when the PPV odds and P(D) odds formulas above have been used, the final ratio should be converted back to a probability statement. These conversions require a few extra steps, but the mathematical operations entailed by this version of Bayes's Theorem are simple compared to other versions. Another important advantage of this approach is that it not only estimates the LOC with which any legal or scientific theory may be held on the basis of a well-defined set of evidence (i.e., PPV) but offers an easily accessible method for determining other important indicia, such as base-rate probabilities (P(D) may be calculated as long as the corresponding LRs and PPVs have been estimated) and LRs (which may be calculated as long as the corresponding base-rate probabilities and corresponding PPVs have been estimated).

Table 1
Steps Involved in Calculating Dual-Rater Diagnostic Reliabilities (LR) and Preevaluation Expectations (P(D))

Procedures	Substance Use Disorder	Other Mental Illness
1. Record the PPV for the diagnosis.	0.72	0.73
2. Record PA^+ for the diagnosis.	0.57	0.58
3. Record PA^- for the diagnosis.	0.54	0.93
4. Subtract PA^- (row 3) from 1.	0.46	0.07
5. Calculate the LR for the diagnosis by dividing the PA^+ (row 2) by the term from row 4.	1.24	8.29
6. Convert the PPV (row 1) into an odds ratio by dividing it by 1 minus PPV.	2.57	2.70
7. Divide the term from row 6 by the LR (row 5) to calculate the base-rate odds for encountering the diagnosis in question.	2.07	0.33
8. Divide the term in row 7 by the term plus 1 to calculate the dual-rater probability (symbolized as P(D)) for encountering the diagnosis prior to evaluating a respondent.	0.67	0.25

Note. Data in the first three rows are from "Revisiting the Reliability of Diagnostic Decisions in Sex Offender Civil Commitment," by R. L. Packard and J. Levenson, 2006, *Sexual Offender Treatment, 1,* pp. 11–12, which represent the values reported in their studies of evaluators in Florida. The positive predictive value (PPV) represents the probability that a specific diagnosis D assigned by one evaluator to a respondent on the basis of diagnostic criteria C is assigned to the same respondent by a second evaluator. PA^+ is equal to the proportion of all offenders who, having been classified by one evaluator as being positive for a specified diagnosis, were assigned the same diagnosis by a second evaluator. PA^- is equal to the proportion of all offenders who, having been classified by one evaluator as being negative for a specified diagnosis, were not assigned this diagnosis by a second evaluator.

greater reduction of uncertainty, as reflected in the 48% difference between P(D) in row 1 and the PPV in row 8.

Results

Table 2 presents the PA^+ and PA^- values reported by Packard and Levenson (2006) for each diagnosis they studied, the LRs for estimating the presence of these diagnostic conditions, the ultimate probabilistic LOCs (PPVs) that were calculated for state evaluators when they classified respondents on the basis of various diagnostic criteria, the subjective preevaluation expectations that evaluators held for encountering each diagnosis (P(D)), and the kappa coefficients reported by Levenson (2004a) and Packard and Levenson. The following conclusions are justified by the contents of Table 2:

1. Four diagnoses—antisocial personality disorder, substance use disorder, PNOS, and personality disorder not otherwise specified—shared psychometric characteristics indicative of illusory diagnostic certainty. This deficiency was reflected in high PPVs (ranging from .45 to .72), low LRs (0.64 to 1.64), and high values of P(D) (.55 to .68). In addition, the PPV was not much larger than the P(D) for three diagnoses (from 2 to 12 percentage points) and, for one diagnosis

Table 2
Psychometric Properties Associated With the Diagnoses Studied by Packard and Levenson (2006)

Diagnostic categories	A. PA^+	B. PA^-	C. LR	D. PPV	E. P(D)	F. kappa
1. Other mental illness	.58	.93	8.29	.73	.25	.70
2. Sexual sadism	.18	.97	6.00	.20	.04	.30
3. Exhibitionism	.33	.93	4.71	.56	.21	.47
4. Other personality disorder	.19	.95	3.80	.43	.16	.29
5. Pedophilia	.62	.80	3.10	.76	.51	.65
6. Antisocial personality disorder	.54	.67	1.64	.67	.55	.51
7. Civil commitment recommendation	.79	.48	1.52	.89	.84	.54
8. Substance use disorder	.57	.54	1.24	.72	.68	.43
9. Paraphilia not otherwise specified	.47	.56	1.07	.65	.63	.36
10. Personality disorder not otherwise specified	.23	.64	0.64	.45	.56	.23

Note. Data in columns A, B, and D are from "Revisiting the Reliability of Diagnostic Decisions in Sex Offender Civil Commitment," by R. L. Packard and J. Levenson, 2006, *Sexual Offender Treatment, 1,* pp. 11–12. PA^+ is equal to the proportion of all offenders who, having been classified by one evaluator as being positive for a specified diagnosis, were assigned the same diagnosis by a second evaluator. PA^- is equal to the proportion of all offenders who, having been classified by one evaluator as being negative for a specified diagnosis, were not assigned this diagnosis by a second evaluator. The positive predictive value (PPV) represents the probability that a specific diagnosis D assigned by one evaluator to a respondent on the basis of diagnostic criteria C is assigned to the same respondent by a second evaluator. P(D) is the subjective preevaluation base-rate expectation that evaluators held for encountering a given diagnosis. LR = likelihood ratio.

(personality disorder not otherwise specified), a negative difference of 11 percentage points indicated that evaluators were more certain of this diagnosis before they had an opportunity to apply its criteria.

2. Civil commitment recommendation shared the same psychometric characteristics as the foregoing diagnoses: With a PPV of .89 and an LR of 1.52, the application of the legally specified criteria for identifying SVPs increased the confidence with which this diagnosis was held by only 5 percentage points. This was probably due to the fact that, prior to even evaluating a respondent, one or both evaluators thought that the chances he would be an SVP (i.e., P(D)) were about 84%. An LR of 1.52 for civil commitment recommendation is also concerning in light of two considerations. First, civil commitment recommendation is based on the presence of a mental abnormality in conjunction with a high risk of recidivism. Second, the LR associated with the optimum actuarial method of identifying high risk offenders for recidivating without mistaking nonrecidivists for recidivists is 3.1 (Wollert, 2006), which reflects a moderate level of detection power. Within this context, an LR of 1.52 means that the reliability for the statutorily defined concept of mental abnormality would probably have been below this level had the mental abnormality coefficient been determined in the absence of actuarial information.

3. LRs in excess of 3 were found for other mental illness, sexual sadism, exhibitionism, other personality disorder, and pedophilia. These results indicated that clinicians were clearly able to use the criteria for these diagnoses more effectively than they were able to use the criteria for civil commitment recommendation and the four diagnoses referenced above in the first subsection. Nonetheless, the postevaluation levels of confidence/certainty in the accuracy of these diagnoses (with PPVs ranging from .20 for sexual sadism to .76 for pedophilia) were still not high enough to dispel reasonable uncertainty as I have defined it.

4. Regarding the value of kappa as a measure of diagnostic reliability in SVP evaluations, the rank order of the kappa coefficients reported by Levenson (2004a) paralleled the rank order of the LRs in Table 2, with the exception of three diagnoses (sexual sadism, exhibitionism, other personality disorder). Packard and Levenson (2006) concluded, however, that kappa was spuriously deflated for these particular diagnoses because the marginals associated with their contingency tables were not homogeneous. Therefore, when these three diagnoses are removed from the kappa analysis because of distortion effects, the remaining kappa coefficients are perfectly aligned with the LRs.

Study 2

Subjects

The entire six-person staff of a clinic that specializes in the outpatient treatment of sex offenders completed a questionnaire that (a) presented a case history and diagnostic criteria pertaining to PNOSN (described below) and (b) asked each clinician to estimate the percentage of once-incarcerated sex offenders he or she had treated whose case history or symptoms matched the referents in the questionnaire. Four clinicians were men and two were women.[6] Three had treated from 40 to 100 sex offenders who had been incarcerated, and the other three had treated from 1,000 to 2,000 incarcerated sex offenders. Five had master's degrees and one had a doctorate. Some had up to 12 years of experience treating sex offenders, but two had provided treatment for only a year or two. On the average, they had 8 years of clinical experience with this population. Although none had ever completed a multipage psychosexual evaluation that was submitted to the court for the purpose of sentencing, the evaluations they completed at the clinic before an offender entered treatment included a life history, diagnostic opinions, the results of actuarial testing, and a treatment amenability assessment. Therefore, as key informants, they constituted an informed sample whose views were not contaminated by a vested interest in reporting high or low estimates of the diagnostic base rate for PNOSN. Furthermore, they had a level of experience in working with sex offenders that was comparable to that of the group of evaluators who participated in Levenson's (2004a) and Packard and Levenson's (2006) research.

---

[6] I am indebted to Casey Weber, Dan Minkel, Kathe Mansfield, Cameron DeYoung, Helen Weber, and Drew Caesar for their participation in this study. Copies of the questionnaire they completed may be obtained from Richard Wollert.

Paraphilias in the DSM and PNOSN

The DSM–IV–TR, which was developed by the American Psychiatric Association (2000), includes a 10-page section on paraphilias that elaborates diagnostic criteria for eight specific disorders (exhibitionism, fetishism, frotteurism, pedophilia, sexual masochism, sexual sadism, transvestic fetishism, and voyeurism). Each set of these standards turns on two components. One is an A criterion that specifies the essential or defining symptoms and features of the disorder in question. For each paraphilia, the A criterion specifies the paraphilic stimulus that leads to sexual arousal on the part of the person with the disorder (e.g., suffering on the part of the victim is the paraphilic stimulus for sexual sadism), the modes for the disorder's expression (e.g., via urges or fantasies), critical experiential elements of the disorder (e.g., repetitive and intense symptoms), and the minimum timeframe over which the previous elements must be manifested (e.g., 6 months). The second component is a B or clinical significance criterion that "helps establish the threshold for the diagnosis . . . in those situations in which the symptomatic presentation . . . is not inherently pathological" (American Psychiatric Association, 2000, p. 8). Examples of conditions that may satisfy this criterion for a paraphilia include having acted on urges related to a paraphilic stimulus, feeling distressed over having paraphilic symptoms, being faced with interpersonal difficulties because of paraphilic symptoms, or experiencing reduced efficacy in social functioning because of paraphilic symptoms.

At the end of the section describing specific paraphilias, a two-sentence paragraph constitutes the entirety of the space in the DSM–IV–TR allotted to discussing the category of PNOS. Explaining only that "this category is included for coding Paraphilias that do not meet the criteria for any of the specific categories," it states that "examples include, but are not limited to, telephone scatologia (obscene phone calls), necrophilia (corpses), partialism (exclusive focus on a part of the body), zoophilia (animals), coprophilia (feces), klismaphilia (enemas), and urophilia (urine)" (American Psychiatric Association, 2000, p. 576). Although specific criteria for these disorders are not considered, it seems only reasonable to assume that clinicians who invoke them must have a set of diagnostic requirements in mind that are equivalent to the A criterion and the B criterion that are required to assign any of the specific paraphilias to a respondent. Otherwise, in the absence of meaningful diagnostic criteria, it is possible that virtually any overt, covert, or inferred sexual behavior that is seen as even slightly problematic could be classified as a paraphilia.

Although PNOSN has never been validated or listed as a diagnosis in the DSM–IV–TR, it has frequently been assigned to SVP respondents (Miller et al., 2005; Prentky et al., 2006; Trowbridge & Adams, 2006–2007; Zander, 2005), and a nine-item checklist of characteristics of sex offenders who are positive for this category has been formulated by Doren (2002). An analogous rape-based diagnostic conception called paraphilic coercive disorder that specified the paraphilic stimulus of the A criterion as "forcing sexual contact . . . on a nonconsenting person" (Fuller, Fuller, & Blashfield, 1990, p. 165) was also unsuccessfully proposed for inclusion in the DSM about 20 years ago (Fuller et al., 1990; Prentky et al., 2006; Zander, 2005). Finally, a case history of a sex offender who allegedly met the criteria for this disorder because his "erotic arousal depended on having

a nonconsenting partner" has been disseminated by a publisher whose works do not represent the policies of the American Psychiatric Association (Spitzer, Gibbon, Skodol, Williams, & First, 1994, p. 173).

Questionnaire

Michael First, the editor of the final versions of both the DSM–IV (American Psychiatric Association, 1994) and the DSM–IV–TR (American Psychiatric Association, 2000) has testified that there "are no accepted diagnostic criteria" that define PNOSN (First, 2006, p. 100). In the absence of a standardized definition, a questionnaire based on the plausible criteria described in the foregoing section was administered to elicit base-rate information about PNOSN. As part of this questionnaire, clinicians who served as key informants were asked to read the case history from the book by Spitzer et al. (1994) and the checklist of characteristics proposed by Doren (2002). Then they were asked to estimate the number of sex offenders they had treated who had been imprisoned for the commission of a sex offense. After this they were asked to estimate the number of these offenders who (a) had case histories that matched the one presented to them, (b) were like the offender depicted in the case history in the sense that their "erotic arousal *depended* on having a nonconsensual partner – that is they couldn't get aroused unless their partner was *resistant* to the sexual advances," and (c) satisfied at least six of Doren's nine items. Lastly, they were asked to divide the number of offenders who fell in each of these three categories by the total number of sex offenders they had treated who had been imprisoned and to adjust any of the percentages that they felt were inaccurate.

Analysis and Results

The mean and standard deviation were calculated for the average base-rate opinion reported by each clinician. A mean of 5% and a standard deviation of 5% were obtained. Although estimates based on recall may result in greater imprecision than other estimation methods, data from key informants are recognized as useful for conducting Bayesian analyses (Grove & Meehl, 1996). Furthermore, the obtained results were consistent with previous reviews of the literature and opinions, suggesting that only a very small number of incarcerated sex offenders would be positive for PNOSN if it were ever found to be a meaningful diagnostic category (First, 2006; Prentky et al., 2006; Trowbridge & Adams, 2006–2007).

Reestimation of PPV

Reestimation of the PPV for PNOS from Packard and Levenson (2006) was justified by the fact that the P(D) for PNOSN reported by the surveyed clinicians was 11 standard deviations below the corresponding P(D) of 63% reported in Study 1 for the evaluators in Packard and Levenson's research. In addition, because the LR for PNOSN had not been previously defined, it seemed reasonable to rely on the LR for PNOS based on Packard and Levenson's research as the most plausible estimate of the LR for PNOSN. Table 3 provides a worksheet of steps involved in the reestimation process. In contrast to an originally estimated 65% LOC (see row 9 and column D in Table 2), the entries in Table 3 indicate that only

Table 3
Steps in Calculating the Level of Certainty for Diagnostic Assignments, Applied Specifically to Paraphilia Not Otherwise Specified–Nonconsent (PNOSN)

Procedures	Example
1. Select a diagnosis for analysis.	Presence of PNOSN
2. Specify an evidence variable thought to predict the presence of the selected diagnosis.	Various criteria proposed to define PNOSN
3. Estimate the base-rate probability of encountering PNOSN in a relevant and clearly defined group of offenders (from Study 2).	.05
4. Estimate the likelihood ratio (LR) of the category of PNOSN for predicting the presence or absence of the diagnostic criteria for this category (from row 9 and column C of Table 2).	1.07 to 1
5. Estimate the base rate odds of encountering PNOSN in the group from which offenders are referred for evaluation. Divide .05 (row 3) by one minus this term (i.e., .05/.95 = .053).	.053 to 1
6. Calculate the final odds that an offender who is positive for the evidence in row 2 has PNOSN. Multiply the term from row 4 by the term from row 5.	.06 to 1
7. Calculate the probability a person who is positive for the evidence under row 2 has PNOSN. Sum the left and right terms from row 6 (.06 + 1 = 1.06). Divide the left term by the sum (.06/1.06 = .06).	.06

Note. The underlined steps show that probabilities are initially converted to odds ratios when the odds ratio version of Bayes's Theorem is used, that mathematical operations are subsequently performed on odds ratios, and that the results are then converted back to probabilities.

a 6% LOC is justified (see row 7) when calculations are based on an estimated P(D) of 5% and an LR of 1.07 (from row 9 and column C in Table 2).

Discussion

Concluding that SVP commitment evaluations were "a highly reliable process," Packard and Levenson (2006, p. 14) implied that high levels of confidence may be placed in the diagnostic assignments that are made by state-hired experts in SVP evaluations. Although significant chi-square values were interpreted as supporting this view, the present research indicates that the high LOC endorsed by Packard and Levenson is simply an illusion. The reason why this is the case is that the high levels of PPV certainty associated with most of the diagnoses typically considered as prerequisites for a mental abnormality were not attributable to the reliable application of diagnostic

criteria. On the contrary, elevated PPVs stemmed from untested beliefs on the part of evaluators that a high percentage of the detainees they encountered would satisfy the criteria for not only one disorder but for multiple disorders. When one of these beliefs was tested in Study 2 of this article through the collection of additional base-rate information, the elicited base rate was found to be considerably below the presumptive base rate. More research would be useful for identifying the dimensions of this discrepancy in light of the fact that certainty opinions will vary as a function of base-rate levels (Donaldson & Wollert, in press). Nonetheless, the results of the present research suggest that some of the expectations evaluators held for encountering other diagnostic categories, including civil commitment recommendation, may have been inflated.

This position gains further credence from an observation by an anonymous reviewer of the present article that "for [Florida] evaluators to get the cases to assess, there was first a screening process by other personnel to determine who would go that far. This has pertinence to the idea that 'prior to evaluating a respondent, one or both evaluators thought that the chances he would be an SVP' were high." This new information, taken together with the obtained results, suggests that greater efforts should be made within the context of SVP evaluations to control for "halo effects" (Thorndike, 1920, p. 25) that are likely to affect the perceptions of evaluators who are exposed to the positive or negative results of previous evaluations of respondents before they have had a chance to reach their own conclusions independently.

One method of injecting more controls into SVP evaluations would be to surreptitiously include nonpredatory offenders who fall just below the commitment standard among the group of offenders whom the end-of-sentence review boards or prosecutors believe are true SVPs. Simultaneously, the flow of potentially prejudicial or biasing information to evaluators could be restricted so that evaluators would not know which offenders scheduled for evaluation had been "prescreened" as SVP candidates or assigned high actuarial scores and diagnosed with a paraphilia by previous evaluators.

In addition to offsetting halo effects, these procedures would constitute a useful pilot study for testing the null hypothesis, suggested by the civil commitment recommendation LR of 1.52 obtained in Study 1, that state-hired experts are unable to differentiate SVPs from non-SVPs to a reasonable degree of certainty. Conducting SVP evaluations under these conditions on a representative sample of incarcerated sex offenders would provide an even stronger test of this hypothesis.

Two further steps should be taken to address other problems with the SVP evaluation process. One is to attempt to improve diagnostic reliability for the group of diagnoses that are relevant for SVP evaluations. The other is to replace the subjective logic model that experts currently use to reach diagnostic decisions with a probabilistic model that generates more accurate certainty estimates.

Recommendations for promoting each of these endeavors are considered in the next two sections. Because the contents of these sections and the results of the foregoing studies hold implications for the future direction of practice, research, and policies related to SVP evaluations, these issues are discussed in the concluding section.

Recommendations for Improving Diagnostic Reliability

Expanding on the earlier comments of Marshall (2006), the results of Levenson's (2004a) research really do constitute an indictment of the SVP civil commitment process, the lack of diagnostic accuracy that characterizes the DSM criteria for the paraphilias when they are used to evaluate SVP respondents, and the way that evaluators use these criteria. They are also consistent with the conclusion of Prentky et al. (2006, p. 382) that "the mental disorder prong [which] plays a central role in legitimizing SVP commitments . . . lacks legitimacy" at the present time.

Of relevance to these issues, Packard and Levenson (2006, p. 13) opined that "diagnoses with specific, behaviorally anchored criteria have better agreement. . . . Those that have confusing, vague criteria . . . have poorer agreement. . . . Categories that are 'not otherwise specified' also provide less specific diagnostic criteria, leading to a certain degree of subjectivity in decision-making." I concur with all of these conclusions except the last, where it seems that a high—rather than nondescript—degree of subjectivity is involved in the assignment of most diagnoses that are relevant to the SVP construct. In contrast to the views of Packard and Levenson, however, I believe a number of steps should be taken to halt poor diagnostic practices and to improve the reliability of the criteria that are most frequently referenced in SVP cases.

Regarding proscriptive action, two recommendations stand out as being particularly important. The first is that psychologists who undertake SVP evaluations should no longer diagnose any SVP respondent as suffering from PNOSN or personality disorder not otherwise specified. Two reasons support this recommendation. One is that both of these diagnoses are so unreliable, with LRs of 1 or less, that it is impossible to attain a reasonable degree of certainty as to their presence under virtually any base rate. The other is that, in light of this circumstance, the only function that a not otherwise specified diagnosis can serve is to provide what Prentky et al. (2006, p. 361) have referred to as a "pretext" for "allowing the state to cast" what may be "the constitutionally doubtful preventive detention of dangerous individuals as constitutionally safe civil commitment."

The second proscriptive recommendation is that psychologists should keep the number of diagnoses they assign in SVP cases to a minimum. As Packard and Levenson (2006) documented, over 70% of those in their cohort who underwent SVP evaluations were assigned multiple diagnoses. This is consistent with my experience, where I periodically encounter assertions to the effect that "this cluster of diagnoses impairs the respondent's volitional capacity and causes him to be a high recidivism risk."

Perhaps evaluators who make such claims anticipate that multiple diagnoses will strengthen their opinions. According to the product law of probability, however, the theory that a respondent is simultaneously positive for several diagnoses will be associated with a lower LOC than the alternative theory that he is positive for whatever single diagnosis is associated with the highest LOC. Furthermore, the upper limit of certainty for the cluster as a whole will be equal to the lowest LOC for the presence of any single condition in the cluster. For example, if the chance of finding a tall partner is 40%, and the chance of finding a dark partner is 30%, and the chance of finding a handsome partner is 10%, the

chance of filling the "tall, dark, and handsome" bill will never be more than 10%. Evaluators therefore should avoid assigning multiple paraphilic diagnoses to respondents unless they can prove that the selected diagnoses are so strongly intercorrelated that the LOC for the presence of the entire set is very high. Currently, this seems like an impossible task in that the highest LOC reported by Packard and Levenson (2006) for a single paraphilic diagnosis (pedophilia) is only 76% (see Table 2).

Regarding prescriptive action, the method and theories used in the development of risk assessment instruments constitute one of the most authoritative bodies of knowledge that might be drawn upon to improve diagnostic reliability (Swets et al., 2000). An important principle derived from this research is that reaching a high LOC as to the presence of a condition that has a relatively low base rate is virtually impossible in the absence of stringent standards of evidence (Wollert, 2006). The developers of risk assessment instruments consequently design their tests so that they include many different scores that reflect low, moderate, and high levels of evidence stringency (Mossman, 2006; Rice & Harris, 1995).

Although the latest manual of DSM disorders, the DSM–IV–TR (American Psychiatric Association, 2000), indicates that maladaptive "recurrent, intense sexually arousing fantasies, sexual urges, or behaviors" are "essential features of a Paraphilia" (p. 566), a continuum for operationalizing these elements at different stringency levels has never been formulated (Marshall, 2006). It has also recently become apparent that the "essential features of a Paraphilia" were misspecified in both versions of the DSM (DSM–IV and DSM–IV–TR) that have been used for SVP evaluations. In particular, the DSM–IV work group on the paraphilias, the DSM–IV *Options Book* (American Psychiatric Association, 1991), and the DSM–IV *Draft Criteria* (American Psychiatric Association, 1993) all conceptualized the A criterion as it was conceptualized in the DSM–III–R, which required the presence of "recurrent intense sexual urges and sexually arousing fantasies" to a particular deviant stimulus "to which the person is attracted . . ." (American Psychiatric Association, 1987, p. 279) ". . . over a six-month period" (p. 282). The B criterion was also conceptualized per the DSM–III–R, which required the person to "have either acted on the urges" or be "markedly distressed by them" (p. 282). The chair of the DSM–IV Task Force and the editor of the DSM–IV–TR and *Draft Criteria* changed the criteria just before DSM–IV went to press, however, so that paraphilic acts and behaviors came to be referenced under the A criterion rather than the B criterion (First, 2006; First & Halon, in press).

This change has had a serious and unintended impact on diagnostic decision making. As explained by DSM–IV–TR editor First, "It's [now] being used by individuals . . . to mean that all you need to do is focus on behaviors in order to meet the criteria" when "the construct [requires] a deviant pattern of sexual arousal" (First, 2006, p. 72) and "behavior itself is [therefore] not enough to make the diagnosis" of paraphilia (Aldhous, 2007, p. 7, citing a portion of his interview with First).

From a psychometric point of view, First's (2006) comments suggest that diagnostic reliability has been undermined by the confusion the foregoing change in criteria sets has created in the minds of evaluators. Further dissemination, recognition, and acceptance of his clarifications would obviously be useful for

enhancing the evidence stringency underlying all paraphilic diagnoses. A new reliability study of SVP evaluators using more stringent diagnostic criteria might conceivably obtain LRs that exceed those reported in column C of Table 2. Larger LRs, in turn, would translate into higher LOCs regarding the presence of the relevant DSM conditions.

The introduction to the use of the DSM–IV–TR manual is another general source of guidelines and observations that might be expected to improve diagnostic reliability. For example, it defines a mental disorder as a "clinically significant behavioral or psychological syndrome . . . that is associated with present distress . . . or disability . . . or with significantly increased risk of suffering death, pain, disability, or an important loss of freedom. . . . [This syndrome] must currently be considered a manifestation of a . . . dysfunction in the individual" (American Psychiatric Association, 2000, p. xxxi). In a subsection titled "Use of Clinical Judgment," the DSM developers remind users that "the specific diagnostic criteria included in the DSM-IV are . . . not meant to be used in a cookbook fashion. . . . Excessively flexible and idiosyncratic application of DSM-IV criteria and conventions reduces its utility as a common language for communication" (p. xxxii).

These guidelines suggest that experts would probably increase the reliability of the diagnoses they assign to SVP respondents by concentrating on present-day information rather than extrapolations from the past, by relying primarily on florid (Monahan, 1992) and observable signs and symptoms rather than inferential speculations, and by interpreting criteria conservatively rather than stretching them to fit cases that fall in gray areas. As Miller et al. (2005) have suggested, "The development of structured or semi-structured interviews for the sexual disorders (like we have for most diagnostic categories) would allow for increased reliability" (p. 48) as well. Furthermore, in the interest of avoiding pretextuality and the use of idiosyncratic categories, A and B criteria that have been discussed above under the DSM subsection of Study 2 should be specified whenever an evaluator decides to invoke a PNOS diagnosis. It would also be important to simultaneously present an LR or other evidence indicating that the proposed category may be used in a reliable way.

It may be tempting, in light of the reliability findings reported in Study 1, to fault the developers of the DSM for not building greater reliability into the criteria sets that are used for reaching opinions in SVP cases. Some criticism, on the one hand, is probably warranted because the current DSM (e.g., American Psychiatric Association, 2000, p. xxxiii) promises that "the use of an established system of diagnosis (such as the DSM) enhances the reliability" with which the "presence of a mental disorder . . . for a subsequent legal determination (e.g., involuntary civil commitment)" might be made. On the other, extensive criticism is unjustified because (a) the foregoing passage was meant to apply to mental health commitments rather than SVP commitments (R. Halon, personal communication, December 4, 2007), and (b) it has always been the case that the "highest priority" of the developers of the DSM has been to "provide a helpful guide to clinical practice" (American Psychiatric Association, 2000, p. xxiii) and to compile a resource that would enhance the access that mentally ill patients have to the benefits of treatment (Fauman, 2002; Frances & Ross, 2001; Spitzer, 2001). Within the latter context, a somewhat elevated rate of false positives—and thus

decreased reliability—is understandable in light of the importance of maximizing patient access to healthcare resources. Psychologists and others who evaluate SVP respondents as forensic experts, however, deplore a high false positive rate because of their investment in minimizing the number of respondents who may be unjustifiably deprived of their freedom for the rest of their lives. It therefore stands to reason that a classification instrument designed with the first set of values in mind would be less than satisfactory when applied to a context where the second set of values prevails. Recognizing this problem, which suggests that research needs to be undertaken for the development of a classification system that applies specifically to SVP respondents, the editors of the DSM have explicitly indicated (American Psychiatric Association, 2000, pp. xxxii, xxxiii, and xxxvii) that,

> when the DSM-IV categories, criteria, and textual descriptions are employed for forensic purposes, there are significant risks that diagnostic information will be misused or misunderstood. These dangers arise because of the imperfect fit between the questions of ultimate concern to the law and the information contained in a clinical diagnosis. . . . A (DSM) diagnosis does not carry any necessary implications regarding the causes of the individual's mental disorder or its associated impairments. . . . The fact that an individual's presentation meets the criteria for a DSM-IV diagnosis does not carry any necessary implication regarding the individual's degree of control over the behaviors that may be associated with the disorder. It is to be understood that inclusion here, for clinical and research purposes, of a diagnostic category such as . . . Pedophilia . . . does not imply that the condition meets legal or other non-medical criteria for what constitutes mental disease, mental disorder, or mental disability.

Recommendations for the Development of a Probabilistic Logic Model for Making Diagnostic Decisions

The present results suggest that a large number of respondents have been committed when the diagnoses they were assigned could not possibly have been given to a reasonable degree of objective certainty. This, in turn, reflects the primitive state of the logic model that evaluators have used for determining the extent to which respondents meet the criteria outlined in Figure 1.

Regarding this point, experts who conduct SVP evaluations have access to many sources of information and guidance (Amenta, Guy, & Edens, 2003; Beech, Fisher, & Thornton, 2003; Boer, Wilson, Gauthier, & Hart, 1997; Craig, Browne, & Stringer, 2003; Doren, 2002; Hanson, 2006; Hanson & Morton-Bourgon, 2007; A. Harris, Phenix, Hanson, & Thornton, 2003; Lanyon, 2001; Miller et al., 2005; Prentky et al., 2006; Quinsey et al., 1998; Seto, 2005; Sinclair, 2000; Trowbridge & Adams, 2006–2007; Wakefield & Underwager, 1998; Wollert, 2006). These resources relate to such issues as the details of SVP laws, expert qualifications, materials that should be reviewed in the course of an evaluation, diagnostic options, ethical concerns, mental abnormality determinations, and, most of all, risk assessment. Although they have encouraged evaluators to develop at least partially quantified certainty opinions to address the issue of recidivism risk, they have not offered any options other than clinical judgment for addressing issues related to diagnosis, the existence of a mental abnormality, and whether or not a respondent meets all of the criteria for being classified as an SVP.

This is a serious problem because it encourages the formulation of certainty opinions or ultimate probabilities about the presence of these conditions on the basis of an ambiguous logic model that is uninformed by explicit decision rules or an adequate consideration of the probabilities—analogous to P(D), PA^{+}, and $1-PA^{-}$—that constitute the mathematical foundation for the calculation of ultimate probabilities. This type of decision making, which I refer to as unrestrained clinical judgment, has been shown to be notoriously inaccurate when it comes to risk prediction (Grove & Meehl, 1996; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Hanson & Morton-Bourgon, 2007; Prentky et al., 2006; Quinsey et al., 1998; Wollert, 2006; Woodworth & Kadane, 2004). As the present results show, it is similarly inaccurate when it comes to the assignment of DSM diagnoses.

Although there may not have been a compelling alternative to clinical logic models in the past, the foundational probabilities identified through the present research and the development of actuarial tests make these models obsolete. In their place, the probabilistic approach that was applied in the studies at hand may easily be transformed into a logic model for determining whether the null hypothesis that a respondent is a non-SVP may be rejected when the criteria in Figure 1 are considered. Referred to in another article (Donaldson & Wollert, in press) as the Null-Bayes Logic Model (NBLM), I contend that psychologists should, to the greatest extent possible, apply this or an equivalent mathematical model to each element of the SVP construct they appraise in an SVP evaluation. In addition, they should inform the court of the criteria they used to reach each opinion and the foundational probabilities from which these opinions were derived. Finally, in cases where this information is not provided, I believe that it should be elicited on cross-examination or by the court in the course of exercising its duty to distinguish science from speculations that "any verbally fluent and somewhat intelligent person can come up with" (Underwager & Wakefield, 1993, p. 5).

The steps that make up the NBLM are summarized below, together with explanatory comments and references to examples where they were applied in Study 1 and Study 2.

Step 1: Analyze the simplest statutory prerequisites (symbolized as R) for classifying a respondent as an SVP first. Other prerequisites may be analyzed after this in order of their complexity and ambiguity. A forensic psychologist using the applied theory to screen a referral, for example, would concentrate first on the offender's diagnostic status to determine whether R is present (R^{+}) or absent (R^{-}). The offender's risk level would be considered after this. If one or both did not fit the legal theory, it would be appropriate and cost effective to terminate the evaluation as soon as it was clear that this was the case and to subsequently inform the referral source of the results. If a fit was obtained with the legal theory, the evaluator would infer the status of those components (i.e., volitional control and disposition to the commission of sexually violent crimes) that have remained relatively undefined. Step 1 is exemplified in row 1 of Table 3.

Step 2: Specify an evidence variable or a set of evidence variables that are thought to predict R^{+} when they are present and to predict R^{-} when they are not present. A specific set of diagnostic criteria may, for example, be thought to predict the presence of a specific DSM diagnosis. Alternatively, as part of a

conjunctive hypothesis for evaluating the presence of a mental abnormality, a DSM diagnosis may be thought to predict the presence of volitional impairment. From the perspective of psychometric theory, a variable that is predicted is called a criterion variable and a variable that is used for the purpose of prediction is called a predictor variable (Cronbach, 1970). Keeping in mind the distinction between criterion and predictor variables, evidence is a predictor variable and R is a criterion variable. Step 2 is exemplified in row 2 of Table 3, which is based on the *Paraphilias in the DSM and PNOSN* subsection of Study 2.

Step 3: Proceed from the null hypothesis that the respondent is R^{-}. As the introduction pointed out, the Supreme Court has set forth a concept that is analogous to the null hypothesis by requiring that SVPs be distinguished from typical recidivists, clearly indicating that a respondent should be considered a non-SVP until it is possible to reject this assumption on the basis of evidence. Furthermore, it is widely recognized that both statistical and legal perspectives draw on the null hypothesis (B. Johnson & Christensen, 2004; R. A. Johnson & Bhattacharyya, 1996; Kidder & Judd, 1986). The excerpt below from a text on statistics (R. A. Johnson & Bhattacharyya, 1996, p. 328) bears this out.

> In the language of statistics, the claim or research hypothesis that we wish to establish is called the alternative hypothesis. . . . The opposite statement, one that nullifies the research hypothesis, is called the null hypothesis. . . . The word "null" in this context means that the assertion we are seeking to establish is actually void. . . . Before claiming that a statement is established statistically, adequate evidence from data must be produced to support it. A close analogy can be made to a court trial where the jury clings to the null hypothesis of "not guilty" unless there is convincing evidence of guilt. The intent of the hearing is to establish that the accused is guilty, rather than to prove that he or she is innocent.

Step 4: Specify the alternative hypothesis that stands in contrast to the null hypothesis. In an SVP evaluation, the alternative hypothesis to the null hypothesis is that the respondent is positive for whatever SVP prerequisite is being considered.

Step 5: Specify the subjective LOC that needs to be reached with respect to the alternative hypothesis in order to reject the null hypothesis. As Donaldson and Wollert (in press) have pointed out, the standard of certainty an evaluator adopts is critical for determining the status of any given null hypothesis having to do with SVP proceedings. With a low standard, the null hypothesis would be rejected for almost all SVP candidates. With a high standard, the null hypothesis would be rejected for a much smaller number. Regarding future risk of recidivism, many laws define the standard as likely or more likely than not, suggesting a certainty threshold in excess of 50% (Woodworth & Kadane, 2004). Statutes leave it to evaluators to define other certainty standards, however. With respect to this issue, I believe that experts should entertain a certainty standard on the order of 90% to 99% for the following reasons.

a. As part of exercising beneficence and nonmaleficence, which is the first principle of the *Ethical Principles of Psychologists* (American Psychological Association, 2002, p. 3), psychologists are charged with striving "to benefit those with whom they work and take care to do no harm. . . . Because psychologists' scientific and professional judgments and actions may affect the lives of others,

they . . . guard against personal, financial, social, organizational, or political factors that might lead to the misuse of their influence." As this passage indicates, conducting one's practice in accordance with nonmaleficence is of the utmost importance in SVP cases (Mercado et al., 2005). A high LOC provides psychologists with a safeguard for doing so.

b. A high LOC also closely approximates the legal standard of proof of beyond a reasonable doubt as this standard has been discussed by various jurists, statisticians, and journalists (Dawid, 2002; Dean, 2006; *United States v. Fatico,* 1978; *United States v. Schipani,* 1968). Although some states require the less demanding standard of clear and convincing evidence (Woodworth & Kadane, 2004), the ethical principle of nonmaleficence suggests that the safest legal reference point for evaluators is beyond a reasonable doubt rather than clear and convincing evidence.

c. Research on the quantification of probabilistic expressions has shown that the terms *very high probability* and *certain* approximate a 90% to 99% LOC (Kadane, 1990; Kent, 1994; Mosteller & Youtz, 1990).

d. A high LOC, on the order of 95%, has traditionally been adopted in psychological research for the purpose of interpreting which findings are significant (Fisher, 1926) and in psychological assessment for the purpose of differentiating those who have extreme scores on intelligence and personality tests from those who have average scores (American Psychiatric Association, 2000; Graham, 1977).

Step 6: Estimate the base-rate probability of encountering R^+ ($P(R^+)$) in the most clearly defined reference group that includes the offender under evaluation. Study 2's key informant survey exemplifies one method for deriving an estimate of the base rate. A variety of other data sources may be accessed for this purpose (e.g., published research, representative sampling of the reference group, individual clinical experience, or quantification of qualitative findings). One minus $P(R^+)$ equals $P(R^-)$.

Step 7: Estimate the LR for the evidence variable. As indicated in Step 2, a variable that is predicted is called a criterion variable and a variable that is used for the purpose of prediction is called a predictor variable. The LR calculated in Step 7 inverts the relationship between criterion and predictor variables, however, because it reflects the power of the criterion variable for indicating the presence or absence of the predictor variable. An odds statement, the LR is the quotient of dividing the probability of finding the evidence selected under Step 2 in a group of offenders who have been classified as R^+ by the probability of finding the same evidence in a group of offenders classified as R^- . In Table 2, for example, the numerator of the diagnostic LR was PA^+ and the denominator was 1 minus PA^-. This step is also exemplified in rows 2 through 5 of Table 1.

Percentiles for the calculation of an LR may also be derived from frequency data (Donaldson & Wollert, in press). If a research project were to focus on the mental-abnormality hypothesis described in Step 2, for example, the numerator of the relevant LR would be the number of volitionally impaired offenders who are positive for a specified DSM disorder divided by the total number of impaired offenders without regard to their diagnostic status. The denominator would be the number of unimpaired offenders who are positive for the same diagnosis divided by the total number of unimpaired offenders.

To fully appreciate the significance of the LR for evaluating the presence of all or part of the SVP construct, it may be helpful to remember that, in accordance with the product law of probability, a construct that is subsumed by another construct will never predict the evidence for the broader construct more accurately than the broader construct. The relatively exclusive category of tall and dark, for example, will never predict the evidence for tall any better than tall by itself. Similarly, the conjoint category of pedophilic and volitionally impaired will never predict the diagnostic criteria for pedophilic better than 3.1, which is the LR for pedophilia. A diagnostic LR may therefore be treated as the maximum possible LR for a conjoint set of conditions which includes the diagnosis in question, such as the elements of a mental abnormality (see Step 2), when the precise value of the conjunctive LR is unclear or unknown. This, in turn, means that reasonable certainty for the presence of a mental abnormality is unattainable whenever reasonable certainty at a diagnostic level is unattainable because of a weak LR.

Step 8: Use $P(R^+)$ and $P(R^-)$ from Step 6 to determine the base-rate odds of encountering R^+. Divide $P(R^+)$ by $P(R^-)$. The quotient will be the left-hand side of the relevant odds statement, and the right-hand side should always be set to a value of 1. Step 8 is also exemplified in row 5 of Table 3.

Step 9: Calculate the ultimate odds that a person who is positive for the evidence under Step 2 may be classified as R^+. Multiply the odds in Step 7 by the odds in Step 8. For example, if the LR from Step 7 is 1.07 to 1, and if the base-rate odds from Step 8 are .053 to 1, the product of multiplying these ratios will be .06 to 1. Step 9 is exemplified in row 6 of Table 3.

Step 10: Calculate the ultimate probability, or PPV, that a person who is positive for the evidence under Step 2 is R^+. Sum the left and right terms from Step 9 (e.g., .06 plus 1 equals 1.06). Then divide the left term by the sum to obtain the final LOC (.06 divided by 1.06 equals 6%). Step 10 is also exemplified in row 7 of Table 3.

Step 11: Compare the PPV from Step 10 with the LOC from Step 5. Uncertainty is dispelled, the null hypothesis may be rejected, and support exists for the alternative hypothesis if the PPV equals or exceeds the LOC.

The critical terms in the computational algorithm that includes Steps 6 through 10 are $P(R^+)$, LR, and PPV. When the values of any two of these terms are known, it is possible to solve for the third by using the odds ratio version of Bayes's Theorem. Charts, called nomograms, have also been developed that table these solutions so that users do not need to make any calculations (Fagan, 1975; Grove & Meehl, 1996; Page & Attia, 2003; Schwartz, 2006; Sonis, 1999). The empty Bayesian nomogram presented in the first panel of Figure 3, for example, consists of three upright axes. The vertical line on the left reflects all possible values of $P(R^+)$. The middle line reflects all possible values of LR. The right line reflects all possible values of PPV.[7] When the values of the first two terms are known, and the goal is to find PPV, hash marks are inserted at the points reflecting the values of $P(R^+)$ and LR on their respective axes. The marks are then

[7] I am indebted to Alan Schwartz for relabeling the axes of his nomogram (Schwartz, 2006) to reflect the terms used in the present article and also for giving me permission to reprint the result in Figure 3.

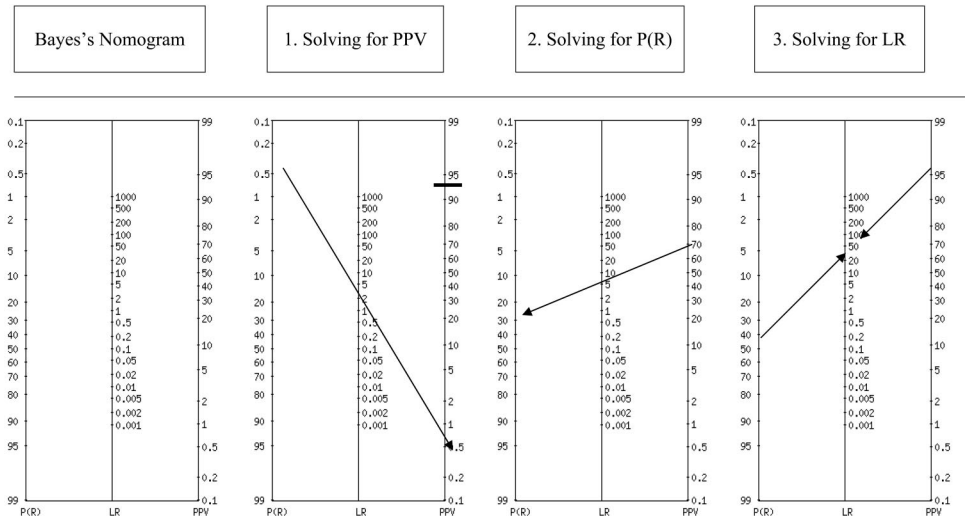

Figure 3. Bayes's nomogram and three examples for using it to solve for positive predictive value (PPV), the base-rate probability of encountering the statutory prerequisites (P(R)), or likelihood ratio (LR) when two other values are known. Arrows point to various solutions. P(R) and LR are known in Solution 1, PPV and LR in Solution 2, and P(R) and PPV in Solution 3. The heavy horizontal line near the top of the PPV axis for the nomogram in the second panel represents the reasonable level of certainty standard discussed under *Step 5* of the article. Adapted from *Diagnostic Test Calculator*, by A. Schwartz, 2006, http://araw.mede.uic.edu/cgi-bin/testcalc.pl. Copyright 2007 by Alan Schwartz. Adapted with permission.

connected by drawing a straight line that cuts across all three axes. The value of PPV is the point where this line intersects the last axis. Therefore, when the values of $P(R^+)$ and LR for PNOSN are inserted into the second panel of Figure 3, the same PPV that was found in Study 2 (i.e., .06) is identified.

More generally, when the values of any two of the three terms in Bayes's nomogram are known, the nomogram may be used to find the unknown value of the remaining term by inserting the known values on their axes and running a line through these points so that it intersects the axis of the remaining term. Inserting the PPV and LR for other mental illness from rows 1 and 5 of Table 1 in the third panel of Figure 3 therefore points to a P(R) of about 25%, which is identical to the value presented in the 8th row of that table.

It is also possible to use Bayes's nomogram to estimate the LR for a paraphilic diagnosis and compare it with other LRs to appraise whether or not an expert's professed LOC falls in a range that is reasonable. For example, suppose that during direct testimony, an expert claims to be reasonably certain that a respondent is a pedophile. On cross-examination, opposing counsel might ask the expert to quantify what a reasonable LOC means and what percentage of respondents in the pool from which referrals are received are "true" pedophiles. If the answer to the first question is "over 95 percent," and the answer to the second is "40 percent," counsel can insert these points in Bayes's nomogram and determine, as shown in the fourth panel of Figure 3, that the expert is presuming that the LR for

pedophilia is about 29. This is clearly an unreasonable assumption in that the LR reported for pedophilia in Table 2 is about 3.1, and the LR for actuarial tests that predict sexual recidivism, which have been the subject of much more reliability research than the diagnosis of pedophilia, is also about 3.1 (Wollert, 2006). Finally, inserting a presumed LR of 3.1 for pedophilia in Bayes's nomogram with a 40% base rate points to a level of diagnostic certainty of slightly over 60%, which is unpersuasive in light of the fact that the chances of a misdiagnosis are close to 40%.

Steps 8 through 10 of the NBLM may therefore be solved by carrying out a few mathematical operations or relying on a nomogram that summarizes the results of these operations. Although use of the nomogram may require some interpolation when it comes to the insertion of hash marks, it reduces the chance of calculation errors when great mathematical precision is not really necessary. Nomograms are also valuable in that they serve as a rough check on the accuracy of mathematical calculations.[8]

Future Directions for Practice, Research, and Policies Related to SVP Evaluations

For a number of years after the initiation of SVP legislation, science had not yet reached the point where it was possible to specify the base rate, predictors, and LRs for any of the variables associated with the SVP construct. Experts did not, as a result, have a feasible alternative to clinical judgment for forming the certainty opinions they conveyed to the court in SVP cases.

The status of science regarding SVP assessment methods has changed, however, over the last 10 years. One critical change has been that actuarial tests based on probabilistic reasoning (Donaldson & Wollert, in press) have come to be commonly used for the purpose of risk estimation. Another is that, as the present article and other supplementary resources (e.g., Levenson, 2004b; Packard & Levenson, 2006; Roberts, Doren, & Thornton, 2002) have indicated, it is now possible to specify the base rates, predictors, and LRs for the diagnostic categories that experts treat as congenital or acquired conditions essential to the definition of a mental abnormality. Taken together, these changes have provided the founda-

[8] Evaluators, attorneys, and judges may also use a set of formulas rather than Bayes's nomogram to solve for PPV, P(R), or LR. These formulas are

$$\mathrm{PPV} = \frac{\frac{\mathrm{P(R)}}{1-\mathrm{P(R)}} \times \mathrm{LR}}{1 + \left(\frac{\mathrm{P(R)}}{1-\mathrm{P(R)}} \times \mathrm{LR}\right)}$$

$$\mathrm{P(R)} = \frac{\frac{\mathrm{PPV}}{1-\mathrm{PPV}} \times \frac{1}{\mathrm{LR}}}{1 + \left(\frac{\mathrm{PPV}}{1-\mathrm{PPV}} \times \frac{1}{\mathrm{LR}}\right)}$$

$$\mathrm{LR} = \frac{(\mathrm{PPV} \times \mathrm{P(R)}) - \mathrm{PPV}}{(\mathrm{PPV} \times \mathrm{P(R)}) - \mathrm{P(R)}}$$

tion and motivation for the formulation of a method of SVP assessment based on probabilistic reasoning called the NBLM.

The NBLM points to new directions that should be taken in both the research and legal arenas. From the standpoint of research, it suggests an agenda that, if completed, could further inform the process by which diagnostic decisions are made in SVP cases. In the legal arena, it holds out the promise of refining the practice of evaluation by forensic psychologists, the examination of expert witnesses by attorneys for both the defense and prosecution, and the administration of justice by the courts as a result of stressing that any respondent who is classified as an SVP must have a very high probability of being positive for at least one discrete and specifiable mental abnormality. By the same token, it would be illogical and irresponsible to classify any respondent as an SVP simply because he has been assigned multiple diagnoses that fail the requisite probability test both separately and collectively.

Compared to the clinical judgment model that evaluators have relied on in the past, the NBLM rests on a solid scientific footing and is quantified, internally consistent, standardized, and generalizable to forensic evaluations other than SVP evaluations. Based on simple calculations, it provides evaluators with a rationale for maintaining their objectivity and professional autonomy, and it empowers them to undertake probabilistic analyses addressing situational or geographic factors that may not have been previously considered in a single expectancy table.

In addition, the NBLM is transparent and economical with respect to specifying the content of evaluations that, because of the prejudicial effects of some types of content (Jackson et al., 2004), should be confined whenever possible to delineating SVP prerequisites, evidence that has a meaningful bearing on these prerequisites, base-rate information, LR information, and tolerable standards of uncertainty. Furthermore, it has the capacity to integrate different types of data from different sources and articulates with other quantified assessment methods. Regarding the first issue, Study 2 exemplified how data based on clinical experience could be combined with data derived from more traditional methods of sampling. Regarding the second, Step 7 of the NBLM calculates the decision threshold operating points that are elemental to receiver operating characteristics theory (Rice & Harris, 1995; Swets et al., 2000); actuarial tables or curves for the prediction of sexual recidivism (Epperson et al., 2003; Hanson, 2006; Prentky et al., 2006, p. 377) are also subsumed by the model in that each of the risk percentages they present are the results of Steps 8 through 10 (Donaldson & Wollert, in press). Finally, the NBLM encourages collegial discourse and the resolution of central evaluation issues even when disagreement might exist on more peripheral issues. Experts who work within this framework may therefore hold somewhat different foundational probabilities regarding an SVP prerequisite for a given respondent but still agree that none of the combinations of probabilities exceeds a reasonable LOC.

Use of the NBLM requires certain accommodations and concessions, however. Some of these apply to the primary figures in the SVP arena in that forensic psychologists, attorneys, and judges who regard the model as more informative and less misleading than clinical judgment will need to acquire an adequate understanding of the concepts, operations, and tools discussed in this article. For example, in the formulation of diagnostic opinions, I believe it will be crucial for

forensic psychologists who use the NBLM to specify and justify (a) the diagnostic criteria they used, (b) an LOC that comports with whatever legal standards of proof (e.g., beyond a reasonable doubt or clear and convincing evidence) are relevant, and (c) the foundational probabilities (i.e., base-rate information and LRs) that informed their final opinions. Opinions that fall below the LOC should also be brought to the court's attention. In the absence of new reliability evidence, the most reasonable expectation on the basis of the results of Study 1 is that expert opinions will rarely fall above a defensible LOC.

Some psychologists may, of course, refuse to accept the NBLM in favor of continuing to rely on clinical judgment. Section 9.01 of the *Ethical Principles of Psychologists* (American Psychological Association, 2002, p. 13) requires, however, that "psychologists base the opinions contained in their . . . evaluative statements, including forensic testimony, on . . . techniques sufficient to substantiate their findings," and Section 9.08 states that "psychologists do not base their assessment . . . decisions . . . on tests and measures that are obsolete" (p. 14). From my perspective, the advances that have been made over the last 10 years in isolating variables that predict risk and diagnostic status, in collecting information on the base rates for encountering sexual recidivism and diagnostic conditions, and in calculating predictor–criterion LRs indicate that it is poor practice to substantiate opinions in SVP cases by relying primarily on clinical judgment or other methods that do not specify foundational probabilities. Consequently, psychologists who continue to follow such outmoded practices may be exposing themselves unnecessarily to the risk of being charged with ethical misconduct.

In spite of the NBLM's advantages, some criticism should be anticipated because of the challenge it represents to clinical judgment. One objection that might be raised, for example, is that it is not accepted by the relevant professional community. Another is that it does not take qualitative evidence into account. A third is that it may be manipulated to obtain whatever results are desired. A fourth is that probabilistic reasoning based on aggregating data at a group level is irrelevant for dealing with the unique nature of each civil commitment respondent. Finally, the NBLM might be criticized on the grounds that it sets a standard for civil commitment that is unreachable except in a very small number of cases.

None of these objections are sufficiently compelling to rule out the use of the NBLM. The first is illogical because the NBLM is directly derived from the statistical training that psychologists receive as undergraduate and graduate students, and any psychologist rejecting the NBLM would therefore be in the position of rejecting his or her own training. The second overlooks the fact that experts always have the option of introducing qualitative evidence and discussing it in such a way as to help a judge or jury understand why some probabilities are likely to be high and others are likely to be low. The third is misleading because it implies that experts may assign whatever values they wish to foundational probabilities when it is much more likely that experts will entertain only those values that are derived through the application of credible methods. The fourth minimizes the importance that humans naturally place on aggregated data when they are faced with very serious decisions. The personal significance that prob-

abilistic knowledge holds in such cases is vividly portrayed in the following scenario by Grove and Meehl (1996, p. 305):

> Suppose you are suffering from a distressing illness . . . and your physician says it would be a good idea to have . . . a certain radical operation. . . . You would naturally inquire . . . how risky it is. The physician might say, ". . . There are people who die on the operating table, but not usually." You would ask, ". . . What percentage of the time does it work? . . . Half . . . 90%, or what? . . . How many people die under the knife? One in a thousand? If it were five in a hundred, I don't know that I'd want to take the chance . . ." How would you react if your physician replied, "Why are you asking me about statistics? We are talking about *you* – an individual patient. You are unique. Do you want to be a mere statistic? What differences do these percentages make, anyway?" We do not think a person should be pleased if the physician replied in that evasive fashion. Why not? Because . . . probability is the guide of life.[9]

Of all the objections raised against the NBLM, the last is probably most accurate. It does not point to a flaw in the model, however, because it is consistent with the legislative intent behind all SVP statutes to identify a small group of offenders. Furthermore, implying that a method is invalid because it does not lend itself well to committing a large group of persons shows a bias in favor of commitment when the first responsibility of forensic psychologists is to scientific impartiality.

Therefore, rather than reflexively attempting to impeach the NBLM because it is a disappointing reminder that the legal theory of the SVP construct might be invalid, it would be more constructive for critics to consider other questions. How small is the group defined by the legal theory? How many committed individuals may have been deprived of their liberty in the past on the basis of evidence that is now obsolete? What is the best way "to seek new trials for these individuals" in the future "to determine whether they actually qualify as SVPs" (Wollert, 2006, p. 79)? Focusing on these alternative questions seems particularly important in light of the upsurge of facts and opinions, published in reputable sources, that challenge the validity of the SVP construct at almost every turn (Davey & Goodenough, 2007; Jackson et al., 2004; Janus, 2000; Janus & Meehl, 1997; Levenson, 2004a; Marshall, 2006; Miller et al., 2005; Morse, 1998; Prentky et al., 2006; Schopp, 1998; Wollert, 2006; Woodworth & Kadane, 2004; Zander, 2005).

Regarding the issue of the actual number of persons who merit SVP status, the present research found that the reliability coefficients for most paraphilias were very low. To correct this problem, and to open the door to research that addresses the prevalence question, evaluators need to rely primarily on present-day behaviors and symptoms that are clearly interpretable when they diagnose SVP respondents and to avoid applying the criteria from the DSM in an over-inclusive way. The American Psychiatric Association will also need to clarify and elaborate the

[9] It is beyond the scope of this article to consider every possible objection that might be raised to predictive approaches based on probabilistic reasoning. Grove and Meehl (1996), however, counter 17 such objections in detail. Readers interested in a broader discussion of this issue are referred to this source.

DSM so that more stringent diagnostic criteria are developed for the paraphilias. Whether or not the American Psychiatric Association decides to pursue this goal, the American Psychological Association should consider striking up a committee of forensic psychologists to develop a system for classifying these disorders that is more adequate than the one that psychologists rely upon now.

The foregoing results and analyses suggest that the percentage of SVP respondents classified by evaluators as meeting the criteria in Figure 1 would shrink dramatically if the recommended corrections were made and evaluators were to base their opinions on the NBLM. In the short run, this might lead many to question the value of relying on science for reaching decisions in SVP cases when its most defensible application makes it difficult to "lock up all the bad guys" with a clear conscience. In the longer run, however, the NBLM would offer reassurance that a value fundamental to the moral soul of our country—the right to due process—has been protected in that bad guys are not civilly committed just because they have been bad but because we are also confident on the basis of meaningful evidence that they are SVPs.

References

Abracen, J., & Looman, J. (2006). Evaluation of civil commitment criteria in a high risk sample of sexual offenders. *Journal of Sexual Offender Civil Commitment, 1,* 124–139.

Aldhous, P. (2007, February 24). Sex offenders: Throwing away the key. *New Scientist, 2592,* 6–9.

Amenta, A. E., Guy, L., & Edens, J. (2003). Sex offender risk assessment: A cautionary note regarding measures attempting to quantify risk. *Journal of Forensic Psychology Practice, 3*(1), 39–50.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., revised). Washington, DC: Author.

American Psychiatric Association. (1991). *Diagnostic and statistical manual of mental disorders options book.* Washington, DC: Author.

American Psychiatric Association. (1993). *Diagnostic and statistical manual of mental disorders* (4th ed., draft criteria). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct* [PDF format]. Retrieved February 8, 2008, from http://www.apa.org/ethics/code2002.html

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

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

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

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

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

Boer, D. P., Wilson, R., Gauthier, C., & Hart, S. (1997). Assessing risk for sexual violence. In C. D. Webster & M. A. Jackson (Eds.), *Impulsivity: Theory, assessment, and treatment* (pp. 326–342). New York: Guilford.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

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

Craig, L. A., Browne, K., & Stringer, I. (2003). Risk scales and factors predictive of sexual offence recidivism. *Trauma, Violence, & Abuse, 4,* 45–69.

Cronbach, L. J. (1970). *Essentials of psychological testing*. New York: Harper & Row.

Davey, M., & Goodenough, A. (2007, March 4). Doubts rise as states hold sex offenders after prison. *The New York Times,* pp. 1, 18–19.

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

Dean, C. (2006, December 5). When questions of science come to the courtroom, truth has many faces. *The New York Times.* Retrieved December 7, 2006, from http://www.nytimes.com/2006/12/05/science/05law.html?n=Top%2fReference%2fTimes%20TOP

de Finetti, B. (1964). Foresight: Its logical laws, its subjective sources. In H. E. Kyburg & H. Smokler (Eds.), *Studies in subjective probability* (pp. 91–158). New York: Wiley.

Denis, D. J. (2001). Inferring the alternative hypothesis: Risky business. *Theory & Science, 1.* Retrieved June 7, 2006, from http://theoryandscience.icaap.org/content/vol002.001/03denis.html

Dix, G. E. (1976). Differential processing of abnormal sex offenders: Utilization of California's mentally disordered sex offender program. *The Journal of Criminal Law & Criminology, 67,* 233–243.

Donaldson, T., & Wollert, R. (in press). A mathematical proof and example that Bayes's Theorem is fundamental to actuarial estimates of sexual recidivism risk. *Sexual Abuse: A Journal of Research and Treatment.*

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

Doren, D. (2004). Stability of the interpretative risk percentages for the RRASOR and Static-99. *Sexual Abuse, 16,* 25–36.

Epperson, D., Kaul, J., Huot, S., Goldman, R., & Alexander, W. (2003, December). *Minnesota Sex Offender Screening Tool—Revised (MnSOST-R): Development, validation, and recommended risk level cut scores*. Retrieved February 18, 2007, from http://www.psychology.iastate.edu/~dle/mnsost_download.htm

Fagan, T. J. (1975). Nomogram for Bayes's Theorem. *New England Journal of Medicine, 293*(5), 257.

Fauman, M. A. (2002). *Study guide to DSM-IV-R.* Washington, DC: American Psychiatric Publishing, Inc.

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian." *Bayesian Analysis, 1*(1), 1–40.

First, M. B. (2006, December 11). Deposition in re the detention of William Davenport, Volume I. In the Superior Court of the State of Washington, in and for the County of Franklin, Case number 99–2-50349–2.

First, M. B., & Halon, R. (in press). Use of DSM paraphilia diagnoses in sexually violent predator commitment cases. *Journal of the American Academy of Psychiatry and Law*.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33,* 503–513.

Frances, A., & Ross, M. (2001). *DSM-IV-TR case studies.* Washington, DC: American Psychiatric Publishing, Inc.

Fuller, A. K., Fuller, A. E., & Blashfield, R. K. (1990). Paraphilic coercive disorder. *Journal of Sex Education & Therapy, 16,* 164–171.

Graham, J. R. (1977). *The MMPI: A practical guide.* New York: Oxford University Press.

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

Grove, W. M., Zald, D., Lebow, B., Snitz, B., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30.

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

Hanson, R. K. (2006). Does Static-99 predict recidivism among older sexual offenders? *Sexual Abuse, 18,* 343–355.

Hanson, R. K., & Morton-Bourgon, K. (2007). *The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis*. Ottawa, Canada: Public Safety and Emergency Preparedness Canada.

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

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

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

Jackson, R. L., Rogers, R., & Shuman, D. (2004). The adequacy and accuracy of sexually violent predator evaluations: Contextualized risk assessment in clinical practice. *International Journal of Forensic Mental Health, 3,* 115–129.

Janus, E. (2000). Sexual predator commitment laws: Lessons for law and the behavioral sciences. *Behavioral Sciences and the Law, 18,* 5–21.

Janus, E. S. (2001). Sex offender commitments and the "inability to control": Developing legal standards and a behavioral vocabulary for an elusive concept. In A. Schlank & F. Cohen (Eds.), *The sexual predator*: *Vol. II* (pp. 1/1–1/30). Kingston, NJ: Civic Research Institute.

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

Jaynes, E. T., & Bretthorst, G. (2003). *Probability theory: The logic of science.* New York: Cambridge University Press.

Jeffreys, H. (1939). *The theory of probability.* Oxford, England: Oxford University Press.

Johnson, B., & Christensen, L. (2004). *Educational research.* Boston: Allyn and Bacon.

Johnson, R. A., & Bhattacharyya, G. (1996). *Statistics: Principles and methods.* New York: Wiley.

Kadane, J. B. (1990). Comment: Codifying chance. *Statistical Science, 5*(1), 18–20.

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

Kansas v. Crane, 534 U.S. 407 (2002).

Kansas v. Hendricks, 521 U.S. 346 (1996).

Kent, S. (1994). *Sherman Kent and the Board of National Estimates: Collected essays.* Washington, DC: Center for the Study of Intelligence.

Kidder, L. H., & Judd, C. (1986). *Research methods in social relations.* New York: Holt, Rinehart, & Winston.

Langton, C. M., Barbaree, H., Seto, M., Peacock, E., Harkins, L., & Hansen, K. (2007). Actuarial assessment of risk for reoffense among adult sex offenders. *Criminal Justice and Behavior, 34,* 37–59.

Lanyon, R. I. (2001). Psychological assessment procedures in sex offending. *Professional Psychology: Research and Practice, 3,* 253–260.

Levenson, J. S. (2004a). Reliability of sexually violent predator civil commitment criteria. *Law and Human Behavior, 28,* 357–369.

Levenson, J. S. (2004b). Sexual predator civil commitment: A comparison of selected and released groups. *International Journal of Offender Therapy and Comparative Criminology, 48,* 638–648.

Marshall, W. L. (2006). Diagnostic problems with sexual offenders. In W. Marshall, Y. Fernandez, & L. Marshall (Eds.), *Sexual offender treatment* (pp. 33–43). New York: Wiley.

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

Mercado, C. C., Schopp, R., & Bornstein, B. (2005). Evaluating sex offenders under sexually violent predator laws: How might health professionals conceptualize the notion of volitional impairment? *Aggression and Violent Behavior, 10,* 289–309.

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

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

Morse, S. J. (1998). Fear of danger, flight from culpability. *Psychology, Public Policy, and Law, 4,* 250–267.

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

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

Mossman, D. (2006). Another look at interpreting risk categories. *Sexual Abuse, 18,* 41–63.

Mosteller, F., & Youtz, C. (1990). Quantifying probabilistic expressions. *Statistical Science, 5,* 2–34.

Nunes, K., Firestone, P., Bradford, J., Greenberg, D., & Broom, I. (2002). A comparison of modified versions of the Static-99 and the sex offender risk appraisal guide. *Sexual Abuse, 14*(3), 253–269.

Packard, R. L., & Levenson, J. (2006). Revisiting the reliability of diagnostic decisions in sex offender civil commitment. *Sexual Offender Treatment, 1*. Retrieved December 19, 2006, from http://www.sexual-offender-treatment.org/50.0.html

Page, J., & Attia, J. (2003). Using Bayes' nomogram to help interpret odds ratios. *Evidence-Based Medicine, 8,* 132–134.

People v. Superior Court, 27 Cal. 4th 888 (2002).

Prentky, R. A., Janus, E., Barbaree, H., Schwartz, B., & Kafka, M. (2006). Sexually violent predators in the courtroom: Science on trial. *Psychology, Public Policy, and Law, 12,* 357–393.

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

Ramsey, F. P. (1931). *The foundations of mathematics.* London: Routledge and Kegan Paul.

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

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

Savage, L. J. (1954). *The foundations of statistics.* New York: Wiley.

Schopp, R. F. (1998). Civil commitment and sexual predators: Competence and condemnation. *Psychology, Public Policy, and Law, 4,* 323–376.

Schwartz, A. (2006). *Diagnostic test calculator*. Retrieved February 20, 2007, from http://araw.mede.uic.edu/cgi-bin/testcalc.pl

Seto, M. (2005). Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. *Psychological Assessment, 17,* 156–167.

Sinclair, L. (Producer). (2000). *Sex offender re-offense risk assessment program* [videotape]. (Available from Sinclair Seminars, 3630 Lake Mendota Dr., Madison, WI 53705)

Sjostedt, G., & Langstrom, N. (2001). Actuarial assessment of sex offender recidivism risk: A cross-validation of the RRASOR and the Static-99 in Sweden. *Law and Human Behavior, 25*(6), 629–645.

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

Sonis, J. (1999). How to use and interpret interval likelihood ratios. *Family Medicine, 31,* 432–437.

Spitzer, R. L. (2001). Values and assumptions in the development of DSM-III and DSM-III-R. *The Journal of Nervous and Mental Disease, 189,* 351–359.

Spitzer, R. L., Gibbon, M., Skodol, A., Williams, J., & First, M. (1994). Perfect relationship. In *DSM-IV-TR casebook* (pp. 171–174). Washington, DC: American Psychiatric Press.

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

Swets, J. A. (1988, June 3). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285–1293.

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

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25–29.

Trowbridge, B. C., & Adams, J. (Fall 2006/Winter 2007). Sexually violent predator assessment issues. *The Trowbridge Foundation Report, 8*(4)/*9*(1), 1–15.

Underwager, R., & Wakefield, H. (1993). A paradigm shift for expert witnesses. *Institute for Psychological Therapies Journal, 5,* 1–18.

United States v. Fatico, 458 F. Supp. 388 (E.D. NY 1978).

United States v. Schipani, 289 F. Supp. 44 (E.D. NY 1968).

Wakefield, H., & Underwager, R. (1998). Assessing violent recidivism in sexual offenders. *Issues in Child Abuse Accusations, 10*. Retrieved December 7, 2006, from http://www.ipt-forensics.com/journal/volume10/j10_6.htm

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

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

Woodworth, G. G. (2004). *Biostatistics: A Bayesian introduction*. Hoboken, NJ: Wiley.

Woodworth, G. G., & Kadane, J. (2004). Expert testimony supporting post-sentence civil incarceration of violent sex offenders. *Law, Probability and Risk, 3,* 221–241.

Wright, D. B. (2006). Causal and associative hypotheses in psychology. *Psychology, Public Policy, and Law, 12,* 190–213.

Zander, T. (2005). Civil commitment without psychosis: The law's reliance on the weakest links in psychodiagnosis. *Journal of Sexual Offender Civil Commitment, 1,* 17–82.

Received March 22, 2007
Revision received August 22, 2007
Accepted December 20, 2007 ■