Bayesian Computations Protect Sexually Violent Predator Evaluations from the Degrading Effects of Confirmatory Bias and Illusions of Certainty: A Reply to Doren and Levenson (2009)

Richard Wollert[1] & Jacqueline Waggoner[2]
[1]Independent Practice, Portland
[2]University of Portland

[Sexual Offender Treatment, Volume 4 (2009), Issue 1]

Abstract

Wollert (2007) published an article in Psychology, Public Policy, and Law on Poor Diagnostic Reliability, The Null-Bayes Logic Model, And Their Implications For Sexually Violent Predator Evaluations. The preceding article by Doren and Levenson (2009) criticizes this article. The article at hand answers their criticisms and also considers the importance of SVP selection systems to SVP evaluations, examines factors that degrade the evidentiary value of SVP evaluations, and suggests procedures that might be followed to preserve their value. It is concluded that many unresolved issues that pertain to SVP evaluations might be clarified by the application of Bayesian analyses and that evaluators could avoid problems associated with confirmatory bias, such as illusions of certainty, by using Bayes s Theorem (Bayes, 1764) to appraise the adequacy of their SVP selection systems.

Key words: Bayesian analysis, sexually violent predators, diagnostic assessment, risk assessment, confirmatory bias, illusions of certainty

The preceding article by Doren and Levenson (2009) criticizes an article by Wollert (2007) in *Psychology, Public Policy, and Law* (PPPL) on Poor Diagnostic Reliability, The Null-Bayes Logic Model, And Their Implications For Sexually Violent Predator Evaluations. This is not the first time, however, that Wollert has encountered these criticisms he first responded to them about 18 months ago when Dr. Doren was on the PPPL panel of expert reviewers who accepted Wollert s article for publication. Since then he has also read an almost identical version of Doren and Levenson s preceding article that they circulated after submitting it to PPPL, where it was rejected.

In addition to its objective content, Doren and Levenson s article is notable for its censorious tone. The authors insist, for example, that some procedures that Wollert followed were patently absurd (p. 13), that we can learn nothing from Wollert s computations (p. 15), and that attempts to apply the Wollert findings in sexually violent predator (SVP) civil commitment evaluations would be seriously misguided (p. 16). Furthermore, they observe that it is rather surprising that a data collection procedure that seemingly fails to meet conventional minimal standards of acceptable scientific rigor was published (p. 13). Finally, they warn practitioners and researchers that disagreement with sexually violent predator statutes (p. 16) will diminish the ability of experts to distinguish constructive science that promotes SVP commitment from destructive science that does not and that this, in turn, will restrict the access that decision-makers have to actuarial data.

Such statements could raise questions in the minds of some readers about the integrity of the peer-review process. Others, however, might see Doren and Levenson s article as a polemic that includes crude mathematical errors[1] and overlooks the major mathematical point of the Wollert article that is, that evaluators can estimate the probability that a diagnostic classification in a SVP evaluation is correct if they know the base rate with which it occurs and the likelihood ratio for the criteria associated with the diagnosis.

Considering these possibilities, we appreciated editor Mike Miner s invitation to add our thoughts to those of Doren and Levenson and have done so in the three remaining sections of the paper at hand. The first presents a brief summary of Wollert s (2007) article. The second consists of a discussion of ten different issues in which, for each issue, we summarize our understanding of Doren and Levenson s position and then lay out our own position. The final section considers the importance of SVP selection systems to SVP evaluations, factors that degrade the evidentiary value of SVP evaluations, and procedures that might be followed to preserve their scientific value.

A Summary of Poor Diagnostic Reliability, The Null-Bayes Logic Model, And Their Implications For Sexually Violent Predator Evaluations

Introduction

Forensic psychologists often determine whether the SVP construct may be applied to a respondent because he satisfies each of three prongs that define it. The first is that he has been convicted of a sexually violent crime. The second is that he suffers from a legally-defined mental abnormality or diagnosed mental disorder consisting of an acquired or congenital condition that impairs his volitional self-control and makes him a sexual danger to others. The third is that he will likely commit future acts of sexual violence because of his mental abnormality (MA).

The foregoing features form the Legal Theory of the SVP Construct (LT), which requires the presence of four elements linked by three causal relationships. Experts agree an acquired or congenital condition means a DSM-IV-TR (American Psychiatric Association, 2000) diagnosis and that likely to recidivate means a chance of re-offending in excess of a standard such as 50%. Terms like volitional impairment (VI) and sexual dangerousness are vague, however. Overall, science has not validated the LT.

The testimony of experts is hampered by the LT s unvalidated status. They have attempted to simplify this problem in two ways. First, they reach a diagnostic opinion and a risk opinion because some assessment criteria have been formulated for both. Then, if these opinions fit the SVP formula, they are combined with other information to infer where the respondent stands on the remaining elements and the MA subconstruct. This Applied Theory (AT) is illustrated in Figure 1.

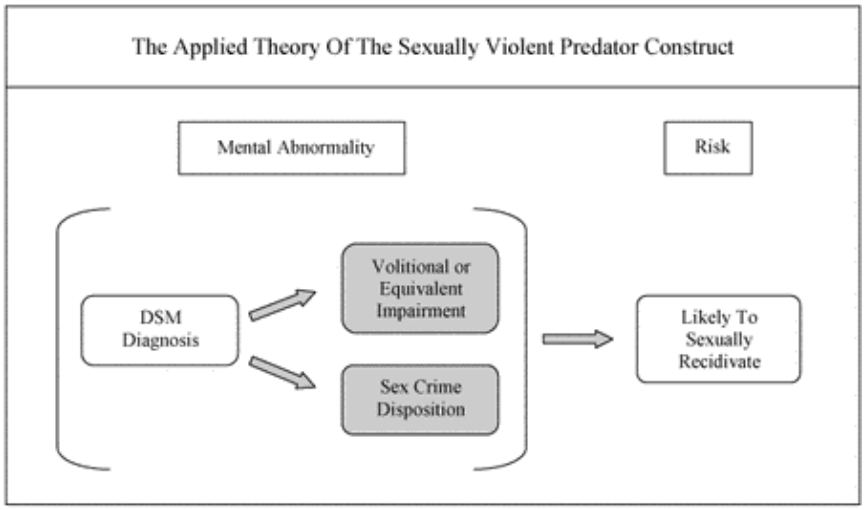

Figure 1: The Applied Theory of the Sexually Violent Predator Construct (from Wollert, 2007). Note. Shaded symbols in the bottom box represent concepts and relationships that are inferred when the Applied Theory is used. When a respondent is positive for all of the features in the figure it is assumed that the SVP construct provides a strong explanation of the respondent s sexual misconduct. By the same token, a respondent is not an SVP if even one of these features cannot be shown to be present.

Experts who claim they can use the AT to distinguish SVPs from the larger population of criminal sex offenders (SOs) must be able to do so with great certainty. DSM diagnoses must be reliably assigned to achieve this goal. Otherwise, it would be impossible to reject the null hypothesis (Donaldson & Wollert, 2008; Wollert, 2006; Wollert, 2007) that a respondent does not suffer from a MA. This, in turn, would make it impossible to reject the null hypothesis that he is not an SVP.

Complicating reliability issues, the product law of probability dictates that the probability of encountering any given diagnosis in concert with a VI will not exceed the probability that different raters will agree on the presence of the diagnosis itself. The reliability for identifying SVP-relevant diagnoses must therefore be very high in order to reliably identify MAs.

Levenson (2004) was the first to study the reliability with which experts could assign DSM diagnoses to SVP candidates. Identifying 295 cases that were examined by more than one forensic evaluator, she calculated a reliability index known as the kappa coefficient for various diagnoses. These calculations led her to conclude that the inter-rater reliability of 8 DSM-IV diagnoses was poor to fair (kappa = .23 to .70). She also observed that the recommendations made by evaluators regarding whether or not to refer a client for civil commitment demonstrated poor reliability (kappa = .54) (p. 357).

Packard and Levenson (2006) calculated new reliability indicia for Levenson s sample. In particular, they reported the level of certainty for the presence of each diagnosis when the same respondent was evaluated by two clinicians (positive predictive value, or PPV), the level of certainty for the absence of each diagnosis (negative predictive value, or NPV), the proportion of raters who agreed on the presence of each diagnosis (positive concordance rate, or PA^{+}), and the proportion who agreed on its absence (negative concordance rate, or PA^{-}). Reversing Levenson s 2004 conclusions because agreement on the diagnosed disorders (PA^{+}) was rather high, they claimed that civil commitment evaluation appears highly reliable (p. 14). They did

not analyze the MA concept, however.
Packard and Levenson s analysis was flawed because the diagnostic PPVs they calculated did not control for the pre-evaluation expectations that experts held for encountering these diagnoses. This is a problem because high PPVs do not necessarily reflect high levels of diagnostic reliability. Experts who assume they will frequently encounter a given diagnosis, for example, will have a high PPV for this diagnosis.

When diagnostic prevalence expectations [symbolized as P(D)] are high because of social or emotional reasons, PPVs may be high even when diagnostic reliability is low. Unfortunately, PPVs based on such circumstances convey only the illusion of diagnostic certainty (IDC).

The prevalence expectation of Levenson s experts for any diagnosis she reported [P(D)] may be calculated in two steps. The first involves calculating the likelihood ratio (LR) for the corresponding diagnostic criteria by dividing PA^+ by 1 PA^-. The second involves solving the following rearrangement of Bayes s Theorem (BT):

$$P(D) = \frac{\frac{PPV}{1-PPV} \; x \; \frac{1}{LR}}{1 + \left(\frac{PPV}{1-PPV} \; x \; \frac{1}{LR} \right)}$$

(1)

In the foregoing formula[2] LR reflects the reliability with which raters agree on the presence versus absence of a discrete diagnosis. PPV reflects the level of certainty that raters have that the diagnosis is accurate. Equation (1) therefore states that P(D) is simply the product of combining inter-rater reliability (as measured by LR) with the level of certainty that a diagnosis is correct (as measured by PPV).

Study 1

Using Packard and Levenson s data, Study 1 solved equation (1) for each of their diagnoses and described various ways to do this (see Table 1, Figure 3, and footnote 9). It was found that antisocial personality disorder, substance use disorder, paraphilia not otherwise specified (PNOS), and personality disorder not otherwise specified (PDNOS) shared psychometric characteristics indicative of IDC in their high PPVs (ranging from .45 to .72), low LRs (.64 to 1.64), and high values of P(D) (.55 to .68).

The LRs for sexual sadism, exhibitionism, pedophilia, and other specific personality disorders were larger. The psychometric characteristics for substance use disorder were compared with those for other mental illness (which included schizophrenia and depressive disorders) to illustrate the differences between illusory and non-illusory diagnostic certainty. 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, the LR for the first condition (1.24) was much smaller than the second (8.29).

Civil commitment recommendation also conveyed an IDC: With a PPV of .89 and a LR of 1.52, using the legally-specified criteria for identifying SVPs increased the confidence with which this

ultimate opinion was held by only 5 percentage points.

Study 2

Study 1 indicated that Levenson s experts held base rate expectations [P(D)] that were unreasonably high for several diagnoses. This produced inflated certainty levels (PPVs). One way to correct an inflated PPV is to collect new base rate [i.e. P(N)] information for the diagnosis in question and solve the following version of BT:

$$PPV = \frac{\frac{P(N)}{1-P(N)} \; x \; LR}{1 + \left(\frac{P(N)}{1-P(N)} \; x \; LR \right)}$$

(2)

This procedure was applied to a proposed diagnosis called Paraphilia Not Otherwise Specified Nonconsent (PNOSN). PNOSN was focused on because Packard and Levenson reported that its PPV was high (65%), and many SVP respondents have been diagnosed with PNOSN. PNOSN is also not in the DSM, lacks scientific validation, and has been derided by some experts as a mythical diagnosis (Zander, 2008). Finally, even experts who think it is a tenable diagnosis suggest different criteria for its identification (Zander, 2005).

A questionnaire based on the different PNOSN criteria was compiled, and staff members of a SO treatment clinic used it to estimate how frequently they encountered previously incarcerated SOs who met the PNOSN criteria. The average P(N) for PNOSN was found to be 5%, and the standard deviation was found to be 5%. When this P(N) was inserted into equation (2), a PPV of only 6% was obtained. Alternative ways of solving equation (2) were also discussed (see Table 3, Figure 3, and footnote 9).

Discussion

Study 1 suggested the high PPVs reported by Packard and Levenson for many diagnoses considered as prerequisites for a MA were not attributable to the reliable application of diagnostic criteria. Wollert s (2007) first manuscript submission theorized that evaluators held *untested beliefs* that a high percentage of the detainees they encountered would satisfy the criteria for one or more disorders. However, Dr. Doren pointed out as a reviewer of the initial manuscript 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 pointed to the conclusion that the high expectations that evaluators held for encountering sexual pathology were also due to a *halo effect* that was created when a finding of dangerousness by a screening evaluator was communicated to a mental health evaluator.

Study 2 suggested that experts would be less confident when they assigned PNOSN and other paraphilic diagnoses in non-SVP evaluations than in SVP evaluations.

Methods to control halo effects and improve diagnostic reliability were also discussed. It was

recommended that SVP evaluators should refrain from diagnosing any SVP respondent as suffering from PNOSN or PDNOS and should keep the number of diagnoses they assign to a minimum. Prescriptively, the adoption of more stringent diagnostic criteria for the identification of both paraphilias and VIs was suggested.

It was also concluded that the clinical logic models that have previously been used to make SVP decisions are obsolete and should be replaced by mathematical models. One such model, called the Null-Bayes Logic Model, was described in detail.

Responses to Doren and Levenson

Issue One

Argument.[3] Packard and Levenson did not assess whether diagnoses were correctly or accurately applied, but rather the likelihood that two independent evaluators would agree that a particular SO met criteria for the same DSM-IV-TR diagnosis. Their research therefore focused on reliability rather than validity. The interpretation of the results of Wollert s Study 1 is irrelevant to Packard and Levenson s research because in Wollert s study more attention was given to validity than reliability.

Response. Doren and Levenson s assertion that Wollert concentrated only on validity is wrong in a specific sense in that the second input variable (i.e., LR) on the right side of the equality sign in equation (2) is a measure of reliability while the output variable (PPV) is a measure of validity. Equation (2) also embodies the consensually-accepted view that validity is a unitary concept (American Educational Research Association, American Psychological Association, National Council on Education, 1999, p. 11) that, among other things, relies on adequate reliability (p. 17). Doren and Levenson s more general thesis that reliability and validity should be dichotomized is therefore also erroneous. Levenson, in fact, endorsed the validity-reliability connection in her 2004(a) article, where she observed that validity implies reliability (p. 366) and suggested that DSM diagnoses were invalid in SVP cases by stating in the same sentence that the current study found an *unacceptable* degree of inconsistency among evaluators, particularly related to diagnostic decisions and civil commitment selection (emphasis added).

While we disagree with Doren and Levenson on this point, we concur with their insistence that scientific rigor (p. 15) should characterize any procedures that combine psychometric measurement with hypothesis-testing in SVP evaluations. Such a rigorous scientific strategy would ideally envision the following steps in addition to the collection of data.

1. The elements of the favored theory (Nickerson, 1998, p. 177) that an evaluee may be classified as a SVP are specified. Figure 1 presents a graphic illustration of such a specification.
2. The elements of a thoughtfully crafted plausible alternative theory under which the evaluee does not meet the SVP criteria are specified. For example, it might be theorized that his sexual misconduct was attributable to criminal impulsivity (Montaldi, 2007) or overwhelming stress (Groth, 1985) rather than the predatory characteristics portrayed in Figure 1.
3. Current scientific knowledge is applied in a competent and unbiased way to develop the optimum psychometric system for classifying the evaluee as an SVP on the basis of observations about him that are thought to be relevant to the SVP construct.

The classification system that is developed in connection with these procedures may distinguish the evaluee from non-SVPs at a high level of certainty and thus confirm the favored SVP hypothesis. Figure 2 depicts this result. The large square on the left represents a population of 200 incarcerated SOs, 5% of whom those in the small shaded squares are SVPs. The rectangle on the right represents a selection system that, when ten SVP selections are made, results in 9 accurate classifications. This is an adequate system because it is right 90% of the time and errs only 10% of the time.

The optimum SVP selection system for the evaluee and others like him may be unable to distinguish him from non-SVPs at a high level of certainty, however. Figure 3 depicts this result. Like Figure 2, the large square on the left represents a population of 200 incarcerated SOs that contains 10 SVPs. The rectangle on the right represents a much less accurate selection system, however, where only 5 SVP classifications are correct out of every 10 that are made.

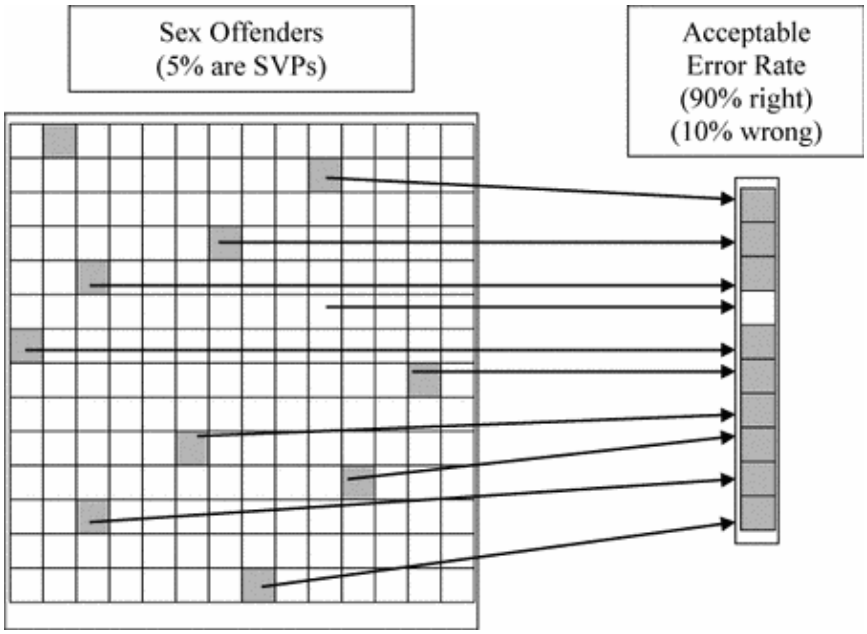

Figure 2: An Accurate System for Identifying SVPs

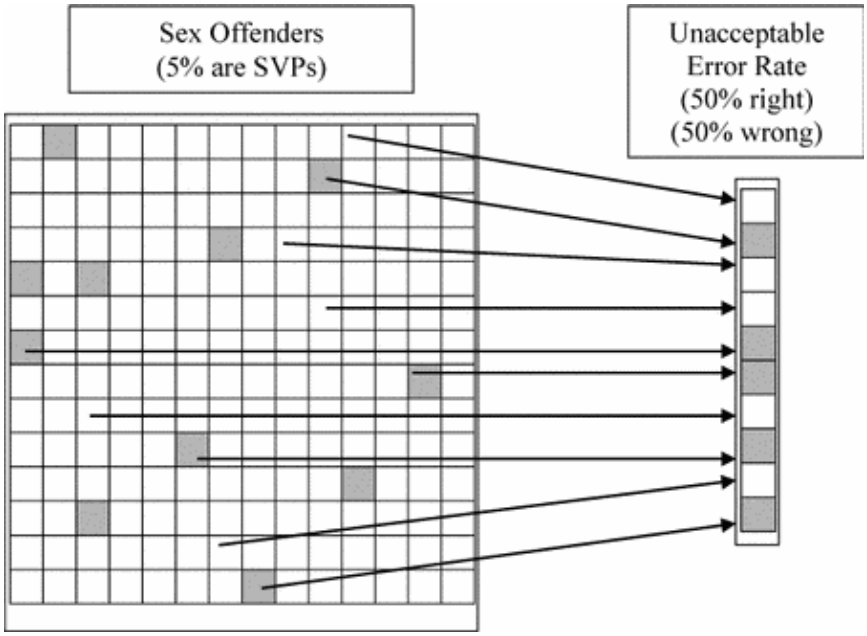

Figure 3: An Inadequate System for Identifying SVPs

After implementing the foregoing steps an evaluator who is working on a specific case must discharge one more obligation disclosing the case-specific adequacy of his or her SVP selection process (see Figures 2 and 3) to the referral source that requested it. Although normative hostility towards SOs (Levenson, Brannon, Fortney, & Baker, 2007) and cognitive similarity heuristics[4] (Tversky & Kahneman, 1974) may bias experts towards confirmation of the SVP theory, the role that such confirmatory biases (Campbell, 2007; Nickerson, 1998) play in the development of the IDCs discussed by Wollert might be moderated if evaluators classified respondents as SVPs only when these classifications were supported by a selection system that operated like the one in Figure 2. By the same token, an expert would not use or recommend the use of the system portrayed in Figure 3 in a civil commitment case, because it is unacceptable to be right only half the time when a wrong decision may result in lifelong incarceration.

This line of reasoning also indicates that the SVP construct cannot provide the only explanation for a respondent s misconduct when the best available SVP selection system does not match Figure 2. In this situation the plausible alternative theory should be regarded as rivaling or even surpassing the SVP theory in providing motivational insights and suggesting clinical interventions.[5]

Within the foregoing context, equation (2) provides the most readily accessible mathematical method that the average evaluator can use to determine the accuracy of an SVP selection system. Evaluators should therefore use it for this purpose and should also report the results they obtain in their commitment evaluations and court testimony.

Issue Two

Argument. Wollert s (2007) AT holds that experts conduct diagnostic assessments in SVP evaluations. Many SVP-relevant diagnoses were studied by Levenson (2004a) and Packard and Levenson (2006). Study 1 of Wollert s article reported that the dual rater expectations of Florida experts for encountering most diagnoses were over 50% and that the confidence that face-to-face evaluators placed in many diagnostic opinions was only slightly higher than their expectations. He argued that evaluator expectations were inflated due to the effects of a halo process in which the

tentative hunches of initial screeners were converted into illusions of certainty within the context of an interview. It is inconceivable that the conclusions of clinical interviewers might be influenced by contextual factors such as the conclusions of initial screeners.

Response. It is helpful to go beyond the information Wollert provided about Florida s SVP nomination system to appraise the plausibility of this argument.[6] According to a governmental report (OPPAGA, February 2000), the Florida legislature passed the SVP Involuntary Civil Commitment Act in 1998. This act required the Department of Children and Families (DCF) to assess persons who have committed sexually violent crimes to determine whether they are likely to commit further sexually violent acts (p. 2). DCF subsequently established a Sexually Violent Predator Program (SVPP) to discharge this obligation by carrying out a three-step process which identified SOs for assessment, completed an initial file review to determine which inmates appear to be most likely to reoffend (Levenson & Morin, p. 610), and referred these inmates for face-to-face evaluation with a licensed psychologist or psychiatrist (p. 610).

Regarding the first and second steps, the Department of Corrections sent 2,418 files of SOs to the SVPP between July 1, 2000 and June 30, 2001 (Levenson, 2004b), where they were reviewed by one of a number of master s level psychological specialists who were part of a Multidisciplinary Team (MDT) (Lucken & Bales, 2008). Many files were also reviewed by licensed psychiatrists or psychologists who held contracts with DCF to participate on the MDT and help administer the SVP nomination system (OPPAGA, February 2000).

Reviewers did not use a standardized protocol for the collection and interpretation of data, but were charged with coding each SO on selected risk factors, compiling both a criminal and clinical history, and entering this information into a state-wide SVP database (Lucken & Bales, 2008). Their recommendations therefore reflected the application of clinical judgment to criminal history information and whatever knowledge they had about SO research in general and recidivism research in particular.

Although reviewers did not meet their evaluees, each reviewer who concluded a SO met the SVP criteria gave him a DSM diagnosis. Only 5% of the evaluees had a prior clinical history that included a sexual disorder diagnosis, however (Lucken & Bales, 2008, p. 116). Actuarial and other risk factors were related to referral decisions but diagnostic factors were not (Lucken & Bales, 2008, p. 117; OPPAGA, February 2000).

Many adult male SOs (N=450) were referred for face to face evaluations over Levenson s study period. At this point doctoral level clinicians on the MDT considered each screening review to determine whether the individual meets the sexually violent predator criteria (OPPAGA, February 2000, p. 3). Although the actual standards remain undisclosed, the MDT arranged for each SO who met criteria to have a clinical evaluation with one of 25 licensed mental health professionals (LPs) who belonged to an SVP Evaluation Panel organized under the auspices of DCF.

Prior to conducting a SO s evaluation, a LP received a file of police reports, correctional records, pre-sentence investigations, previous psychological evaluations, internal correspondence by DCF staffers, correspondence between the MDT and the Attorney General s office, and notes on factors that were aggravating in nature. A record of positive accomplishments was not compiled, and time constraints for the completion of evaluations made it difficult for evaluators to locate such data.

DCF provided two trainings per year because some LPs had little forensic training. At one such training it was suggested that LPs assign the diagnosis of PNOSN to SOs convicted of rape and Pedophilia to child molesters. Such recommendations emphasized diagnostic decision-making

on the basis of offense-specific behaviors, so evaluators rarely conducted a separate test for the presence of a VI. Furthermore, evaluators were not trained in the limitations of actuarial testing even though they were required to score actuarials such as the MnSOST-R (Epperson, Kaul, & Hesselton, 1999) and Static-99 (Hanson & Thornton, 2000) and another test called the Psychopathy Checklist (PCL-R; Hare, 1991).

Although the Florida LPs have been described in previous articles by Levenson (2004a) as objective and by Packard and Levenson (2006) as independent, our investigation of the Florida system indicated that some LPs perceived that the rates with which they referred respondents for civil commitment proceedings were correlated with whether they were viewed in a positive or negative light by members of the MDT. Such perceptions may have been unintended, but they placed LPs in the position of having to choose between their personal and professional interests.

Wollert (2007) argued that Levenson s and Doren s descriptions indicated that the opinions of LPs were affected by negative halo processes that arose because they were aware that all evaluees had been classified as potential SVPs. Additional information, presented above, suggests that it was more biased than this. In particular, LPs were encouraged as loyal members of a team to confirm theories that were passed on to them. This bias was reinforced by offense-specific diagnostic practices, rigid risk assessment procedures, ready access to confirmatory information, and restricted access to disconfirmatory information. These factors suggest that LPs were overly confident in their diagnostic decisions.

Issue Three

Argument. A recent article by Lucken and Bales (2008) reported that every Florida SO who was referred for a clinical evaluation was assigned an SVP-relevant diagnosis at the screening stage. This proves that the first stage of the two-stage nomination process used in Florida did not inflate the diagnostic expectations of the LPs.

Response. Two facts from the foregoing section are relevant to this argument. First, specialists who referred SOs for further evaluation were required to assign an SVP-relevant diagnosis to each case. Second, 5% of those who were referred for evaluations had a clinical history that included a sexual disorder diagnosis. The first fact indicates that all nominees were assigned a diagnosis only because this was required by an administrative rule. The low diagnostic base rate reflected in the second fact points to the conclusion that the percentage of referred SOs who were actually positive for a diagnosis was probably far less than 100%.

Issue Four

Arguments. Levenson and Packard s results were valid. In contrast, Wollert incorrectly claimed that the LR (1.52) associated with civil commitment recommendations was a measure of certainty. He also did not collect data on the MA subconstruct. Doren and Levenson did so after Wollert s paper was published, however, and several results from this effort point to the conclusion that it is highly reliable. For example, one indicator of its reliability was that the PPV for MA was 90.5% (210 of 232 raters agreed on its presence) while its NPV was 40% (18 of 45 raters agreed on its absence). A second was that the dual rater expectation [P(D)] that evaluators held for encountering a MA prior to evaluating a respondent, which was 85.5%, did not differ from its overall concordance rate of 82% (228 of 277 raters). A third was that evaluators achieved a 34% gain over what would be expected per chance [i.e., P(D)] by using the MA criteria. Finally, the concordance rate for the absence of a MA (18 of 45 raters, or 40%) was lower than the concordance rate for its presence

(210 of 232 raters, or 90.5%). This pattern attests to the forensic value of the MA subconstruct because by definition, positive findings about MA are what take cases into court -- not negative findings.

Responses. Although numerous, the foregoing allegations are hollow. Regarding the assertion in the first sentence, for example, Doren and Levenson repeatedly claim that Levenson obtained high reliability coefficients. They do not, however, offer any new research that calls Wollert s basic conclusion that paraphilic diagnoses are characterized by poor reliability into question. Regarding the second sentence, page 179 of Wollert s (2007) article indicated that the certainty level for the civil commitment recommendation was a function of the base rate *and* the LR, not just the LR by itself. Regarding the third, summary data on the MA subconstruct could not have been reported in the Wollert article, because Levenson did not report any data on it. Regarding the fourth and fifth sentences, the PPV and NPV that Doren and Levenson report for the MA subconstruct indicate that the LR associated with it is only 1.62 [i.e., (210/237) divided by (22/40)]. Rather than disputing Wollert s point, this very poor LR confirms it. Regarding the sixth sentence, a significant statistical test for the difference between the dual-rater expectation for encountering MAs and the overall concordance rate for the MA subconstruct might provide some evidence for its validity. A non-significant test does not, however. Regarding the seventh sentence, it is inappropriate to characterize the difference between P(D) and PPV for the MA subconstruct as relative improvement over chance, because P(D) in this case is a measure of subjective expectations and does not reflect the true base rate for MA among the evaluees. Regarding the last two sentences, a pattern which shows that the concordance rate for the presence of a legal construct exceeds the concordance rate for its absence does not validate the construct for use in a forensic setting (Nickerson, 1998). The reason for this is that the only cases that should be brought to court are those where the base rate of the construct in question and its LR (which is the ratio of the positive concordance rate for the characteristic to one minus its negative concordance rate) are sufficient, per equation (2), to exceed a reasonable degree of certainty.

Issue Five

Argument. The results of Wollert s research are irrelevant for assessing SVP candidates because the MA subconstruct includes elements other than a DSM diagnosis, and Wollert incorrectly equated the two. Consequently, Wollert also mistakenly assumed that diagnostic inter-rater reliability is synonymous with the inter-rater reliability of evaluator opinions concerning the presence of a MA.

Response. Wollert consistently indicated in his introduction and figures (e.g., Figure 1) that the SVP construct consists of the MA subconstruct and the risk element. He further stressed that the MA subconstruct requires the presence of three elements: a DSM diagnosis, a VI (or its equivalent), and a predisposition for engaging in sexually dangerous crimes. Doren and Levenson are therefore wrong in asserting that he equated a MA with a DSM diagnosis.

Practically, however, the predisposition element of the MA subconstruct loads most heavily on the risk element. The larger portion of the MA subconstruct, in contrast, revolves around a diagnostic condition and its associated degree of VI. Wollert s article, echoing the views of others (First & Halon, 2008; Jackson & Richards, 2007; Janus, 2001; Mercado et al., 2005; Montaldi, 2007), encouraged evaluators to expand the AT by giving more thought to defining the VI element and testing for its presence without relying exclusively on concepts from the DSM.

Contrary to this approach, Florida evaluators tested only for the presence of a diagnosis (see Issue Two). Assuming that a VI assessment was covered by a diagnostic assessment, they not Wollert

equated the LR for a specific diagnosis with the LR for a MA based on that diagnosis.

Evaluators may eventually reach a consensus as to what is meant by a VI. It may then be possible to break down each of the LRs based on Packard and Levenson s data into one component that loads on the concept of VI and another that loads on diagnostic features that have little to do with it.

For now, however, the LRs reported by Wollert (2007) provide a useful anchor for gauging whether crossing the MA threshold in a SVP evaluation is within the realm of possibility. For example, suppose an evaluator believes that a SVP nominee might possibly be positive for antisocial personality disorder (ASPD). The LR for ASPD reported by Wollert is 1.64 (p. 179). Roberts, Doren, & Thornton (2002) also reported that the base rate for ASPD in a sample of SVP nominees is 35% (p. 576). Inserting these figures into equation (2), it may be determined that, all other things being equal, there is about a 47% chance that an evaluator s theory that a nominee is positive for ASPD is actually correct. Obviously, such a high level of doubt rules out the assignment of ASPD in the absence of a consistent record of strong evidence to the contrary. This, in turn, rules out the assignment of a MA based on ASPD and indicates that the evaluator needs to (a) consider other diagnostic possibilities or (b) conclude that the evaluee does not meet the criteria that define a SVP.

Issue Six

Argument. A MA that rests on a specific diagnostic condition will be unreliable if experts cannot reliably apply the criteria that define the condition. However, specific SVP-relevant conditions that do not overlap non-SVP conditions may be combined into a compound set of conditions. Following this type of disjunctive approach, a respondent who is positive for any one of the conditions in the compound set would be classified as having a MA. In contrast, a respondent who was negative for all of the diagnoses in the compound set would be regarded as normal. The reliability of the MA subconstruct associated with this type of disjunctive model would be greater than the reliability of any specific diagnostic condition subsumed by the model.

Response. The disjunctive model does not hold up to practical or empirical scrutiny. Practically, isolating a discrete set of SVP conditions such as those envisioned by Doren and Levenson would be impossible because of the limitations of diagnostic reliability. Furthermore, the architects of the DSM have never adopted a disjunctive diagnostic model and have clearly asserted that a DSM disorder does not establish a legally-defined MA (First & Halon, 2008; Frances, Sreenivasan, & Weinberger, 2008). The chances are therefore low that a disjunctiv e model for the identification of a MA would ever be included in the DSM solely on the basis of Doren and Levenson s recommendation.

Empirically, the overall LR of 1.62 for the MA subconstruct that was calculated in the response to Issue Four does not exceed the LRs for such SVP-relevant paraphilias as Sexual Sadism (6.0) and Pedophilia (3.1). This result not only disconfirms the disjunctive model but means that the PPVs for Sexual Sadism and Pedophilia would be underestimated if the overall LR for MA were to be inserted in equation (2) in place of the more precise and accurate LRs that Wollert reported for these conditions.

Issue Seven

Argument. The results of Wollert s second study are invalid because Lucken and Bales (2008) reported that 100% of the SOs who were referred for a face to face interview were classified as having a MA while Wollert estimated that the base rate for PNOSN was only 5% for SOs who were released from prison.

Response. The difference in rates cited by Doren and Levenson has no meaningful bearing on how Wollert s results should be interpreted, because reviewers classified all referred SOs as having a MA as a matter of policy rather than diagnostic judgment. Furthermore, Lucken and Bales randomly sampled 773 records from the SVP data base files that were compiled by reviewers. They found that 1% of those who were released (*N*=5,931) and that 5% of those who were referred (*N*=415) had a clinical history of being given a sexual disorder diagnosis (Lucken & Bales, 2008, p. 116). Overall, this means that only about 1.3% of all incarcerated SOs in Florida were diagnosed as having a sexual disorder prior to being screened. It also indicates that Wollert s estimate that only 5% of all formerly incarcerated SOs are positive for PNOSN is not unreasonable.

Issue Eight

Argument. One of Wollert s analyses indicated that the diagnostic certainty for PNOSN is so low that no one who conducts an SVP evaluation should diagnose any SVP respondent as suffering from it. This conclusion is based on the results of a written survey he administered to the staff of a SO treatment clinic for the purpose of estimating the base rate of PNOSN among SOs in general. The estimate Wollert obtained may not have been accurate because he (1) used idiosyncratic criteria for identifying PNOSN; (2) did not provide enough information about the clinicians he surveyed; (3) did not collect data on the reliability of his estimate; and (4) surveyed clinicians when the only reliable method of obtaining mental health information is through a records review or direct interviews. The diagnostic certainty for PNOSN could be very high if Wollert s estimate is wrong.

Response. Several considerations undermine this argument. One is that Wollert addressed the first three of the foregoing reservations in his 2007 article by providing information that the clinicians he surveyed had a level of experience that was comparable to that of Levenson s evaluators (p. 180), by defining PNOSN using a variety of plausible criteria that were described in detail (pp. 181-182), and by calculating the standard deviation for the estimated rate of PNOSN (p. 182). Another is that dismissing survey research would be misguided in that it is not unusual to survey clinicians for the purpose of conducting diagnostic research on rapists (Fuller, Fuller, & Blashfield, 1990; Marshall, 2006; McLawsen, Jackson, Vannoy, Gagliardi, & Scalora, 2008), and survey research represents a valuable complement to records research when results from each method converge on the same conclusion. From this perspective, the results of the records research conducted by Lucken and Bales (2008) that was reported in the previous section strengthens the empirical foundation of Wollert s estimate. It also points up the inadequacy of relying solely on speculation to validate a set of results.

This line of reasoning suggests that the most constructive scientific option that Doren and Levenson might exercise would be to launch an empirical study of their own that informs this issue. However, we do not believe that the results of future research are likely to change our recommendation that SVP evaluators should discontinue the diagnostic use of PNOSN. The reason for this is that our prohibition recommendation is based primarily on the LR for PNOSN, which is so abysmally low (1.06) that PPV [see equation (2)] would not exceed 50% even if the base rate were 40%. PNOSN in its current undeveloped state will therefore never, regardless of its base rate, warrant a high level of expert confidence.

Issue Nine

Argument. We believe that about 5% of all incarcerated SOs may be identified to a reasonable degree of certainty as SVPs. Wollert is so hostile to SVP laws on constitutional grounds, however, that he does not accept scientific evidence that confirms our theory, and he never will. He would

also deprive decision-makers of actuarial tools that could help them make diagnostic assessments and identify recidivists.

Response. Regarding the last assertion of this argument, Wollert has encouraged the refinement of actuarials (Wollert, 2002; Wollert, 2003), explained how actuarial mathematics might be used in SVP evaluations for diagnostic (Wollert, 2007) and risk assessment purposes (Donaldson & Wollert, 2008; Wollert, 2006), and disseminated and validated actuarial tests (Waggoner, Wollert, & Cramer, 2008; Wollert, August 2007). This is not a record that would deprive decision-makers of actuarial tables that, on the basis of compelling evidence and theory (American Educational Research Association, 1999, p. 17), are valid for identifying SVPs.

Regarding the first assertion, Wollert has also opined that respondents are SVPs in several cases. This obviously means that he is not hostile to SVP laws on the grounds that preventive detention is an unconstitutional practice.

The remaining issue for discussion revolves around the scientific status of the SVP construct. Unlike the legislative arena, the scientific arena imposes an exacting standard on the proponents of a theory. In particular, it is their responsibility to prove their point and not the responsibility of skeptics to disprove it. The scientific landscape would otherwise be littered with the baggage of useless and potentially harmful theories.

Within this context, it is incorrect for Doren and Levenson to imply that strong evidence attests to the validity of the SVP theory and the various psychiatric and psychological tools that are commonly-used to identify SVPs . On the contrary, there has been an upsurge of facts and opinions, published in reputable sources, that challenge the validity of the SVP construct at almost every turn (Wollert, 2007, p. 197). Its status has also been undermined by recent research on psychological tests. For example, the PCL-R (Boccaccini, Turner, & Murrie, 2008) and MnSOST-R (Murrie, Boccaccini, Turner, Meeks, & Woods, 2009), each of which were used extensively by Levenson s LPs, have been found to be unreliable in the hands of SVP evaluators. Furthermore, the developers of Static-99 have acknowledged that the 2000 actuarial table relied upon by Levenson s experts overestimates sexual recidivism risk, because a significant drop in recidivism rates has occurred in the last 15 years (Harris, Helmus, Hanson, & Thornton, October, 2008). They have consequently instructed evaluators to discontinue its use.

Doren and Levenson s assumption that their position will be dismissed regardless of the quality of the evidence they might provide on its behalf is also incorrect, because all they need to do to make their point is to produce an array of SVP selection systems such as the one depicted in Figure 2 (see Issue One) that apply to SOs in states that have SVP laws. Although they have not tried to do so, the feasibility of reaching this goal could be assessed if they were to implement two streams of research. The first of these would focus on diagnostic reliability and would calculate PPVs for SVP-relevant diagnoses using equation (2). Since current SVP-relevant diagnoses are largely unreliable, the criteria sets for each diagnostic condition would have to be modified in the course of this research so that PA+ would be increased, perhaps by further elaboration of the VI concept, and PA would be increased by specifying criteria associated with the condition s absence. The prevalence rates of each diagnostic condition among incarcerated populations of SOs in states with SVP laws would also need to be determined prior to applying equation (2).

The second stream of research that Doren and Levenson could implement for studying the validity of the SVP construct would focus on determining the chances of sexual recidivism and include the following steps.

1. Compilation of a test from a list of items thought to predict sexual recidivism.
2. Formulation of a system for assigning a score to each test item.
3. Formulation of a system for obtaining a total test score (symbolized as S_j) for each SO scored on all test items.
4. Administration of the test to a large (e,g, N=3,000) unbiased sample of U.S. SOs upon their release from prison.
5. Compilation of the criminal activities of the sample of released SOs to identify sexual recidivists (R+) and non-recidivists (R).
6. Calculation of the sexual recidivism [P(R+)] and nonrecidivism [P(R)] rates for the entire sample.
7. Calculation of the relative frequency of recidivists assigned each total test score S_j [symbolized as $P(S_j|R+)$].
8. Calculation of the relative frequency of non-recidivists assigned each S_j [symbolized as $P(S_j|R)$].
9. Solution of the following equation from Donaldson and Wollert (2008), which is a version of equation (2), to determine the recidivism rate for SOs with the highest S_j:

$$\mathrm{PPV} = \frac{P(S_j|R+) \times P(R+)}{P(S_j|R+) \times P(R+) + P(S_j|R-) \times P(R-)} \quad (3)$$

Strong evidence for the feasibility of developing an adequate SVP selection system would be established if the first line of research were to show that some diagnostic PPVs exceeded 90% and if the second line of research were to show that the PPV for the highest score on the risk assessment test [see equation (3)] exceeded 50%.

Far from depriving decision-makers of actuarial tools for making diagnostic assessments and identifying recidivists, the foregoing research programs would give them excellent selection devices. We therefore encourage Doren and Levenson to take the Bayes challenge if they are truly interested in attempting to validate the SVP theory.

Issue Ten

Argument. Wollert used flawed methodologies that invalidated his findings. There is therefore nothing of value that can be learned from Wollert s computations, and experts should not rely on them when they evaluate respondents in SVP cases.

Response. It is unreasonable to criticize Wollert s use of Bayesian statistical analysis, because this method holds out many advantages (Wollert, 2007), has long been accepted by statisticians (Fienberg, 2006), is widely-used in scientific endeavors (Woodworth, 2004), and has accurately predicted the relationship between age and sexual recidivism (Wollert, 2006; Hanson, 2006; Wollert, August 2007). There is therefore no reason to expect that it will produce invalid results as long as the values of Bayesian variables [e.g., P(R+), $P(S_j|R+)$, and $P(S_j|R)$ in equation (3)] can be estimated accurately, and the results are not applied to irrelevant populations.

Several considerations also suggest that SVP experts should reject Doren and Levenson s view that nothing of value can be learned from Wollert s computations or Bayesian research. For

example, Bayesian analyses have already produced a substantial body of knowledge pertaining to the SVP construct (see, in particular, Donaldson & Wollert, 2008; Janus & Meehl, 1997; Mossman, 2006; Waggoner, Wollert, & Cramer, 2008; Wollert, 2007; Wollert, 2006). Furthermore, solutions to equations (2) and (3) constitute the most accessible methodology for discharging the duty that evaluators have to describe the adequacy of their SVP selection systems (see Figures 2 and 3). Finally, Principle C of the *Ethical Principles of Psychologists and Code of Conduct* (American Psychological Association, 2000) states that psychologists seek to promote accuracy and Section 9.01 states that psychologists base the opinions contained in their recommendations, reports, and diagnostic or evaluative statements, including forensic testimony, on information and techniques sufficient to substantiate their findings. Within this context, experts who are unfamiliar with Bayesian research on SVPs and do not cite it in their reports because nothing can be learned from it may convey the ethically precarious impression that they are uninformed.

If case-building and confirmation of a favored hypothesis were the goals of expert testimony, it would make sense for Doren and Levenson to advise experts who favor the SVP construct against including Bayesian computations in their evaluations, because some analyses have asserted that the SVP risk standard may be unattainable (Janus & Meehl, 1997), others have shown that older offenders with high actuarial scores are unlikely to recidivate (Wollert, 2006), and still others suggest that many SVP-relevant diagnoses cannot be assigned to a reasonable degree of certainty (Wollert, 2007).

Case-building and confirmation of a favored hypothesis are not the goals of expert testimony, however. On the contrary, an expert conducting an evaluation must maintain professional integrity by examining the issue at hand from all reasonable perspectives, actively seeking information which will differentially test rival hypotheses (American Psychological Association, March 9, 1991, Section VI.C.) When adequate data are available, as in SVP evaluations, Bayes s Theorem [Bayes, 1764; see equations (2) and (3)] is an optimal method for testing the two rival hypotheses one of which is that a respondent is an SVP, and the other of which is that he is not that are of concern. Therefore, contrary to Doren and Levenson s advice, evaluators who are invested in maintaining their professional integrity should emphasize Bayesian computations.

We also believe that the truth value of SVP evaluations would gradually be increased if experts kept an eye open for issues related to the SVP construct that could be clarified through Bayesian computations. For example, a Static-99 experience table that estimates the sexual recidivism rates of SOs released from U.S. prisons has never been compiled. Such a table would be useful, however, because information on Static-99 is often cited in civil commitment probable cause petitions that are filed against U.S. SOs who have been detained after completing their prison terms.

This particular gap in the SVP knowledge base can be filled quickly and economically by estimating the values of P(R+), P(R), $P(S_j|R+)$, and $P(S_j|R)$ from current data that are readily available and then solving equation (3). Tables 1 and 2 show these computations. Sexual recidivism rate data for 17,697 SOs released from prisons in 14 states between 1989 and 2000 are presented in Table 1. This table, which was compiled primarily from exhaustive SO samples tracked by state correctional departments, indicated that the weighted average follow-up period for the entire cohort spanned five years and that the weighted average rate of sexual recidivism (P(R+) was 7%. P(R) was therefore 93%.

The first five columns of Table 2 estimate each $P(S_j|R+)$ and $P(S_j|R)$ from Static-99 data on 2,638 offenders that Doren and others[7] provided us in 2005 (Wollert, Lytton, Waggoner, & Goulet, November 2005) so we could verify that the LRs underlying one of his papers (Doren, 2004)

approximated the LRs obtained in earlier Static-99 research (Hanson & Thornton, 2000). The sixth column presents the PPV, or sexual recidivism rate, that was obtained for each Sj by applying equation (3) to the Bayesian estimates.

Table 1: Sexual Recidivism Rates For Exhaustive Samples Of Sex Offenders Reported By U. S. Correctional Agencies From 2000 to 2007

Source, Year	Base Rate	Sample Size	Groups In The Sample	% Prison Releasees	Follow-Up Period (In Years)	When Follow-Up Began	Definition Of Sexual Recidivism: RA=Rearrest RC=Re-conviction
State of Minnesota, 2007[a]	.120	3,166	1	All	8.3	1990	RA for a sex crime
U.S. Dept. of Justice, 2003[b]	.052g	9,466g	14	All	3.0	1994	RA for a sex crime
State of Washington, 2005[c]	.034	1,939	1	All	5.0	1994	RC for a sex crime
State of Iowa, 2000[d]	.045	202	2	All	4.3	1995	RC for a sex crime
State of Ohio, 2001[e]	.093	879	1	All	10	1989	RC or violation for a sex crime
State of New York, 2007[f]	.060	2,045	4	.55	5.0	2000	RA for a registerable sex crime
Summary[h]	**.065[i]**	**17,697[j]**			**4.8[k]**		

Notes. Data are from [a] MN Department of Corrections (April 2007), [b] Langan, Schmitt, & DuRose (2003), in which data were compiled data from OR, CA, AZ, TX, IL, MI, MN, OH, NY, NJ, MD, DL, VA, NC, and FL, [c] Barnoski (2005), [d] Adkins, Huff, & Stageberg (December 2000), [e] OH Department of Rehabilitation and Correction (April 2001), and [f] NY State Division of Probation and Correctional Alternatives (May 2007). [g] 225 offenders were removed from this study because they were also reported in the MN Department of Correctons study. [h] Data came from 19 different states. [i] This is the average of the base rates when each is weighted by its sample size. [j] This is the total number of offenders. [k] This is the average length of the follow-up periods when each is weighted by its sample size.

Table 2: PPVs (Sexual Recidivism Rates) Obtained When The $P(S_j|R+)$s And $P(S_j|R-)$s For Each Static-99 Score Are Combined With A Five-Year Base

Recidivation Rate Of 7% Per Bayes's Theorem

S_j	R_j+	$P(S_j\|R+)$	R_j-	$P(S_j\|R-)$	PPV
H	140	.399	278	.122	.199
MH	102	.290	465	.203	.098
ML	76	.217	764	.334	.047
L	33	.094	780	.341	.020
Totals	**351**	**100%**	**2,287**	**100%**	

Notes. The abbreviations in the first 5 header columns stand for the following variables: S_j = a score of j on Static-99 [6+ = high (H); 4-5 = moderately high (MH); 2-3 = moderately low (ML); 0-1 = low (L)]; R_j+ = number of recidivists with a score of j; $P(S_j|R+)$ = the relative frequency of S_j among recidivists; R_j- = number of nonrecidivists with a score of j; $P(S_j|R-)$ = the relative frequency of S_j among nonrecidivists. PPV is the positive predictive value that was obtained when the 7% base rate from the second column of the Summary Row in Table 1 was combined with the relative frequencies from each row of this table according to the sequence of operations presented in equation (3).

Experts whose reports are appended to petitions arguing that probable cause exists to believe a SO is an SVP should find these computations relevant. We say this because the fact to be determined with respect to the risk element of the SVP construct at a probable cause hearing in many states is whether the probability of the offender s reoffending exceeds 50% (In re the detention of Brooks, 2001). The PPV/sexual recidivism rate for the highest score in Table 2 does not exceed 20%, however. It would seem prudent, in light of this discrepancy, for experts to reference this 20% figure in any evaluations submitted to the court. Otherwise, they might leave themselves open to ethical proceedings on the grounds that they violated Section VII.D. of the *Specialty Guidelines for Forensic Psychologists* (American Psychological Association, March 9, 1991), which states that forensic psychologists do not, by either commission or omission, participate in a misrepresentation of their evidence, nor do they participate in partisan attempts to avoid the presentation of evidence contrary to their position.

Discussion

Several Bayesian analyses bearing on the SVP construct have now been published. Rejoinders to four Bayesian articles (Janus & Meehl, 1997; Mossman, 2006; Wollert, 2006; Wollert, 2007) have also been published (Doren, 2006; Doren & Epperson, 2001; Doren & Levenson, 2009; Harris & Rice, 2007). Overall, the first set of articles would have to be judged as more meritorious than the second on the grounds that they were published in stronger journals. Nonetheless, experts and attorneys occasionally attempt to argue that the results reported in one of the Bayesian papers has

been refuted by one of the anti-Bayesian rejoinders.

How does the scientific community get beyond this situation, where it sometimes seems that weak and scattered arguments are formulated for the primary purpose of giving evaluators an excuse even a far-fetched one for not using a powerful analytical tool that might undermine case-building by calling attention to the limitations of the SVP construct? Obviously, advocates of Bayesian methods need to be given more opportunities to respond in a timely way to the criticisms of anti-Bayesians. To our knowledge, the present exchange of views is the first occasion where such a dialogue has been facilitated by a journal editor.

We are grateful to Dr. Miner for taking this step for a couple of reasons. First, we were able to discuss a diverse array of topics. In particular, we had a chance to address the deficiencies in Doren and Levenson s methodological criticisms (see Issues Three through Eight), extend Wollert s 2007 reliability analysis to the MA subsconstruct (Issue Four), highlight the value of Bayesian computations for appraising diagnostic certainty (Issue Five) and estimating sexual recidivism risk (Issue Ten), analyze Doren and Levenson s polemical stance (Issue Nine), further document the unvalidated status of the SVP construct (Issues Nine and Ten), and recommend a research program that advocates of the SVP construct could adopt in an attempt to validate it (Issue Nine).

We also appreciate Dr. Miner s invitation because it gave us a chance to consider the relationship of SVP selection systems to SVP evaluations, factors that undermine the evidentiary value of SVP evaluations, and procedures that might be followed to preserve their value. Regarding the first topic (see Tables 2 and 3 from Issue One), we conceptualized the challenge of developing adequate SVP selection systems as a problem in applied mathematics (Meehl, 1996, p. 266). Regarding the second, we elucidated a number of normative, cognitive, and contextual factors that should be controlled by state SVP nomination systems because they instill biases favoring the confirmation of the SVP hypothesis by evaluators (Issues One and Three). Regarding the third, we pointed out that evaluators might avoid the trap of confirmatory bias, which leads to illusions of certainty, by using Bayes s Theorem to appraise the adequacy of their SVP selection systems whenever this is possible (Issues One, Nine, and Ten).

Overall, many unresolved issues that pertain to SVP evaluations might be clarified by the application of Bayesian analyses. Our conversations with colleagues and attorneys lead us to believe that Bayesian concepts are now more well-understood than was the case even three years ago and that many evaluators consider them at some point in their deliberations. The principles of Bayesian analysis are therefore increasingly being applied in SVP evaluations. The set of articles at hand will hopefully accelerate this momentum and encourage similar exchanges in the future.

References

1. Adkins, G., Huff, D., & Stageberg, P. (December, 2000). The Iowa Sex Offender Registry and recidivism. Iowa City, IO: Iowa Department of Human Rights.
2. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). Standards for educational and psychological testing. Washington, D.C.: American Educational Research Association.
3. American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders (4th ed. Text Revision). Washington, D.C.: American Psychiatric Association.
4. American Psychological Association (2002). Ethical principles of psychologists and code of conduct. Retrieved on 12.7.2006 from http://www.apa.org/ethics/code2002.html (PDF format).

5. American Psychological Association (March 9, 1991). Specialty Guidelines for forensic psychologists. Washington D.C.: American Psychological Association.
6. Barnoski, R. (2005). Sex offender sentencing in Washington State. Document No. 05-08-1203. Washington State Institute for Public Policy at Olympia (www.wsipp.wa.gov).
7. 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.
8. Boccaccini, M. T., Turner, D., & Murrie, D. (2008). Do some evaluators report consistently higher or lower PCL-R scores than others? Psychology, Public Policy, and Law, 14, 262-283.
9. Campbell, T. (2007). Assessing sex offenders. Springfield, IL: Charles C. Thomas.
10. Donaldson, T., & Wollert, R. (2008). 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, 20, 206-217.
11. Doren, D. (2004). Stability of the interpretative risk percentages for the RRASOR and Static-99. Sexual Abuse, 16, 25 36.
12. Doren, D. M., & Epperson, D. (2001). Great analysis, but problematic assumptions: A critique of Janus & Meehl (1997). Sexual Abuse, 13, 45-51.
13. Doren, D. (2006). Battling with Bayes: When statistical analyses just won t do. Sex Offender Law Report, 7(4), 49-50, 60-61.
14. Doren, D., & Levenson, J. (2009). Diagnostic reliability and sex offender civil commitment evaluations. Sex Offender Treatment.
15. Epperson, D., Kaul, J., & Hesselton, D. (1999). Minnesota Sex Offender Screening Tool Revised (MnSOST-R): Development, performance, and recommended risk level cut scores. Unpublished manuscript (available from dle@iastate.edu).
16. Fienberg, S. E. (2006). When did Bayesian inference become Bayesian ? Bayesian Analysis, 1(1), 1-40.
17. First, M. B., & Halon, R. (2008). Use of DSM paraphilia diagnoses in sexually violent predator cases. The Journal of the American Academy of Psychiatry and the Law, 36, 443-454.
18. Frances, A., Sreenivasan, S., & Weinberger, L. (2008). Defining mental disorder when it really counts: DSM-IV-TR and SVP/SDP statues. The Journal of the American Academy of Psychiatry and the Law, 36, 375-384.
19. Fuller, A. K., Fuller, A. E., & Blashfield, R. K. (1990). Paraphilic coercive disorder. Journal of Sex Education & Therapy, 16, 164-171.
20. Groth, A. N. (1985). Men Who Rape. New York: Plenum Press.
21. Hanson, R. K. (2006). Does Static-99 predict recidivism among older sexual offenders? Sexual Abuse, 18, 343-355.
22. 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.
23. Hare, R. D. (1991). The Psychopathy Checklist Revised. Toronto: Multi-Health Systems.
24. Harris, A. J. R., Helmus, L., Hanson, R., & Thornton, D. (October 2008). Are new norms needed for Static-99? Retrieved on 12.28.2008 from http://www.static99.org/pdf.docs/atsa2008static-99.pdf.
25. Harris, G. T., & Rice, M. (2007). Characterizing the value of actuarial violence risk assessments. Criminal Justice and Behavior, 34, 1638 1658.
26. In re the Detention of Brooks (2001). 145 Wn 2d 275, 296; 36P. 3d 1034.
27. Jackson, R. L., & Richards, H. (2007). Evaluations for the civil commitment of sex offenders. In R. Jackson (Ed.), Learning forensic assessment (pp. 183-209). Mahweh, NJ: Lawrence Earlbaum Associates.
28. Jackson, R. L., Rogers, R., & Shuman, D. (2004). The adequacy and accuracy of sexually violent predator evaluations. International Journal of Forensic Mental Health, 3, 115-129.

29. Janus, E. S. (2001). Sex offender commitments and the inability to control. In A. Schlank & F. Cohen (Eds.), The sexual predator: Vol. II (pp. 1/1 1/30). Kingston, NJ: Civic Research Institute.
30. 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.
31. Kansas v. Crane, 534 U.S. 407 (2002).
32. Langan, P. A., Schmitt, E., & Durose, M. (2003). Recidivism of sex offenders released from prison in 1994. Washington, DC: U.S. Department of Justice. Available at http://www.ojp.usdoj.gov/bjs/.
33. Levenson, J. S., Brannon, Y., Fortney, T., & Baker, J. (2007). Public perceptions about sex offenders and community protection policies. Analyses of Social Issues and Public Policy, 7, 1-25.
34. Levenson, J. S. (2004a). Reliability of sexually violent civil commitment criteria in Florida. Law and Human Behavior, 28, 357-369.
35. 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.
36. Levenson, J. S., & Morin, J. (2006). Factors predicting selection of sexually violent predators for civil commitment. International Journal of Offender Therapy and Comparative Criminology, 50, 609-629.
37. Lucken, K., & Bales, W. (2008). Florida s Sexually Violent Predator Program. Crime & Delinquency, 54, 95-127.
38. 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: John Wiley.
39. McLawsen, J. E., Jackson, R., Vannoy, S., Gagliardi, G., & Scalora, M. Professional perspectives on sexual sadism (2008). Sexual Abuse, 20, 272-304.
40. Meehl, P. E. (1996). Bootstrap taxometrics. American Psychologist, 50, 266-275.
41. Mercado, C. C., Schopp, R., & Bornstein, B. (2005). Evaluating sex offenders under sexually violent predator laws. Aggression and Violent Behavior, 10, 289-309.
42. Minnesota Department of Corrections (April 2007). Sex Offender Recidivism in Minnesota. Downloaded on May 30, 2008 from http://www.corr.state.mn.us
43. Montaldi, D. F. (2007). The logic of sexually violent predator status in the United States of America. Sexual Offender Treatment, 2(1), 1-28.
44. Mossman, D. (2006). Another look at interpreting risk categories. Sexual Abuse, 18, 41-64.
45. Mosteller, F. & Youtz, C. (1990). Quantifying probabilistic expressions. Statistical Science, 5, 2-34.
46. Murrie, D., Boccaccini, M., Turner, D., Meeks, M., Woods, C. & Tussey, C. Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? Psychology, Public Policy, and Law, 15, 19-53.
47. New York State Division of Probation and Correctional Alternatives (May 2007). Research Bulletin: Sex Offender Populations, Recidivism, and Actuarial Assessment. Available at http://www.dpcs.state.ny/statistics.htm
48. Nickerson, R. S. (1998). Confirmation bias. Review of General Psychology, 2, 175-220. Office of Program Policy Analysis and Government Accountability (February 2000). The sexually violent predator program s assessment process continues to evolve. Report No. 99-36. Tallahassee, FL: Author.
49. Ohio Department of Rehabilitation and Correction (April 2001). Ten-year Recidivism Follow-up of 1989 Sex Offender Releases. Retrieved May 30, 2008, from

http://www.drc.state.oh.us/web/Reports/compendium2002.pdf
50. Packard, R. L., & Levenson, J. (2006). Revisiting the reliability of diagnostic decisions in sex offender civil commitment. Retrieved on 12.19.2006 from http://www.sexual-offender-treatment.org/50.0html.
51. 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.
52. Roberts, C. F., Doren, D., & Thornton, D. (2002). Dimensions associated with assessments of sex offender recidivism risk. Criminal Justice and Behavior, 29, 569-589.
53. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.
54. Waggoner, J., Wollert, R., & Cramer, E. (2008). A respecification of Hanson s updated Static-99 experience table that controls for the effects of age on sexual recidivism among young offenders. Law, Probability and Risk, 7(4), 305-312.
55. Wollert, R. (August, 2007). Validation of a Bayesian Method for Assessing Sexual Recidivism Risk. Presented in San Franciso at the 2007 conference of the American Psychological Association.
56. Wollert, R. (2002). The importance of cross-validation in actuarial test construction: Shrinkage in the risk estimates for the Minnesota Sex Offender Screening Tool Revised. Journal of Threat Assessment, 2(1), 87 102.
57. Wollert, R. (2003). Additional flaws in the Minnesota Sex Offender Screening Tool-Revised: A response to Doren and Dow. Journal of Threat Assessment, 2, 65-78.
58. Wollert, R., Lytton, D., Waggoner, J., & Goulet, M. (November, 2005). Competent use of actuarial tests requires understanding sample-wise variations in both recidivism and test accuracy. Presented at the annual convention of the Association for the Treatment of Sexual Abusers. Salt Lake City, UT.
59. Wollert, R. (2006). Low base rates limit expert certainty when current actuarial tests are used to identify sexually violent predators. Psychology, Public Policy, and Law, 12, 56-85.
60. Wollert, R. (2007). Poor diagnostic reliability, the Null-Bayes Logic Model, and their implications for sexually violent predator evaluations. Psychology, Public Policy, and Law, 13, 167-203.
61. 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.
62. Zander, T. (2008). Commentary: Inventing diagnosis for civil commitment of rapists. The Journal of the American Academy of Psychiatry and the Law, 36, 459-469.

Authors Notes

Requests for reprints of this article and related correspondence should be sent to Dr. Wollert at 1220 SW Morrison St., Suite 930, Portland, OR 97205. He may also be contacted by phone at 360-737-7712 or by e-mail at rwwollert@aol.com. The address of his website is www.richardwollert.com. Correspondence to Dr. Waggoner should be sent to 5000 N. Willamette Blvd., Portland, OR 97203. She may also be contacted by phone at 503.943.8012 or by e-mail at waggoner@up.edu.

Footnotes

[1] For example, both the last sentence before the Conclusions section of Doren and Levenson s article and the fourth sentence of their Conclusions section assert that the likelihood ratio is determined by the base rate. Published formulas for the calculation of likelihood ratios do not,

however, include a term that represents the base rate (Donaldson & Wollert, 2008; Mossman, 2006; Wollert, 2007).

[2] Wollert used a worksheet format to explain his analyses. Formulas from his article were used in the present article for the sake of brevity.

[3] A content analysis was performed on Doren and Levenson s manuscript to clarify their arguments. Each sentence was assigned an alphanumeric identifier that referenced its section, paragraph, and paragraph placement. Sentences were then sorted into blocks of thematic content, the argument that ran through each block was summarized, and we composed a response to each argument. We did not respond to several isolated statements that were incorrect or polemical because this would have been digressive.

[4] A cognitive heuristic reduces the complicated task of assessing probabilities to a simple judgment (Tversky & Kahneman, 1974). The similarity heuristic, also called the representativeness heuristic, evaluates the probability that person P has characteristic C based on the perceived degree of similarity between P and C. Observers who justifiably perceive a person as bad because he is a sex offender, from this perspective, may precipitously jump to the conclusion that he also has the characteristics of an SVP when he actually does not differ from a typical criminal recidivist (Kansas v. Crane, 2002, p. 5).

[5] Some evaluators may be hesitant, in spite of such advantages, to present a clinical formulation based on non-SVP or non-DSM constructs because they believe that their rival theory must be shown to be true to a reasonable degree of certainty and that the SVP theory must be regarded as true if their alternative does not greatly exceed a clinical degree of certainty. These assumptions are wrong, however, in that the SVP theory must be proved to be true on its own merits. Rejecting a non-SVP theory, in other words, cannot confirm the SVP theory.

[6] The authors are indebted to Professor Karole Lucken, psychologist Dr. Natalie Novick Brown, and a second psychologist for discussing the Florida system at length with them.

[7] Doren s 2004 article reported risk percentages for seven data sets. We obtained frequency data for five data sets from Doren and for one from its compilers. We were able to estimate frequency data for the missing data set, which included only 172 offenders, because we knew the psychometric properties of all the other data sets.

Author address

Richard Wollert, Ph.D.
Independent Practice
1220 S. W. Morrison St., Suite 930
Portland, OR 97205
360.737.7712
rwwollert@aol.com

Jacqueline Waggoner, Ed.D.
University of Portland
5000 N. Willamette Blvd.
Portland, OR 97203
503-943-8012
waggoner@up.edu