Revisiting the Reliability of Diagnostic Decisions in Sex Offender Civil Commitment

Richard L. Packard, Jill S. Levenson
Lynn University, Boca Raton, Florida, USA

[Sexual Offender Treatment, Volume 1 (2006), Issue 3]

Abstract

Levenson (2004) investigated the inter-rater reliability of DSM-IV diagnoses commonly assessed by forensic evaluators in sex offender civil commitment proceedings and determined that the reliability of civil commitment selection (kappa = .54) and DSM-IV diagnostic categories (kappa = .23 - .70) were poor. The current study first reviews the limitations of using kappa in reliability studies and the reasons why the statistic may lead to paradoxical findings. Next, using Levenson s data as a demonstration, alternative statistical analyses measuring raw proportions of agreement, odds and risk ratios, and estimated conditional probabilities were utilized to examine reliability. Agreement on the existence of the majority of the diagnosed disorders was rather high despite low values of kappa. The proportions of total agreement in diagnostic decisions ranged from 68% to 97%, indicating that, overall, civil commitment evaluations were a reliable process. The strengths and limitations of alternative methods of measuring inter-rater reliability are illustrated, and implications for policy and practice are discussed.

Key words: sex offender, sexual predator, civil commitment, inter-rater reliability, DSM, diagnosis, kappa

Author s note: The authors wish to thank John Morin and Paul Stern for their reviews of an earlier draft. Their valuable suggestions helped strengthen the manuscript.

Forensic examiners are often called upon to render opinions as to whether a person has a mental disorder. These questions are encountered in various legal contexts, including competency, criminal culpability, workers compensation, torts, sentencing, and psychiatric commitment (Melton, Petrila, Poythress, & Slobogin, 1997). Seventeen states have implemented civil commitment laws for sex offenders, by which sexually violent predators (SVP) can be involuntarily treated in secure facilities beyond their criminal sentence. Although not necessarily a legal requisite, evaluators typically use the nosology described in the Diagnostic and Statistical Manual for Mental Disorders, Fourth Edition, Text Revision (*DSM-IV-TR*) (American Psychiatric Association, 2000) for such purposes. Use of the *DSM* in forensic matters has been subject to criticism, particularly for a lack of data regarding reliability (Campbell, 1999;2004; Grove, Andreasen, McDonald-Scott, Keller, & Shapiro, 1981; Kirk & Kutchins, 1994; Meyer, 2002).

The purpose of this study is to examine the strengths and weaknesses of different methods of assessing the reliability of diagnostic decisions within the forensic context of sex offender civil commitment. In order to illustrate these issues, we analyzed data from a recent study investigating the reliability of SVP civil commitment criteria (Levenson, 2004). Levenson s data has some unique advantages for this purpose: it was derived from a large sample that was carefully assembled, it used field clinicians involved in real-world decisions rather than simulated cases, and it involved a forensic issue where the presence of a mental condition was required as part of the commitment criteria, thus obligating each clinician to specifically assess for mental illness using the DSM.

Sex Offender Civil Commitment

Seventeen states have passed legislation that allows for the civil commitment of dangerous sex offenders. These sexually violent predator policies allow high risk sex offenders to be civilly detained in a mental health facility following their incarceration so that they can receive treatment in a secure setting. Generally these statutes require that there be a legal finding that the person has a mental abnormality or personality disorder that makes the person likely to engage in acts of sexual violence ("Kansas v. Hendricks," 1997). Although state laws vary in their definitions of the term mental abnormality, many of them use language similar to that of the Florida statute: a mental condition affecting a person s emotional or volitional capacity which predisposes the person to commit sexually violent offenses ("Jimmy Ryce Act," 1998). Evaluators customarily use the diagnostic categories described in the *DSM* when arriving at their conclusions, and the sections that discuss and define personality disorders and paraphilia disorders are particularly relevant (Doren, 2002).

Likelihood of reoffense is usually determined through the use of actuarial risk assessment instruments. These tools use empirically derived risk factors to estimate probability of reoffense by referring to the known recidivism rates of sex offenders with similar characteristics. The most commonly used instruments for this purpose are the Static-99 (Hanson & Thornton, 1999), the Rapid Risk Assessment for Sex Offense Recidivism (RRASOR) (Hanson, 1997), and the Minnesota Sex Offender Screening Tool-Revised (MnSOST-R) (Epperson, Kaul, Huot, Hesselton, Alexander, & Goldman, 1999). Although these instruments cannot specifically predict whether or not a given individual will reoffend, they have demonstrated moderate predictive validity and are valuable in informing expectations regarding the likelihood of recidivism (Barbaree, Seto, Langton, & Peacock, 2001).

Whether or not an individual meets criteria for civil commitment is a legal finding ultimately made by the trier of fact, but the diagnostic conclusions drawn by mental health professionals are central to the court process. States that have adopted sexual predator civil commitment laws typically rely upon at least one assessment conducted by a mental health professional, but often SVP candidates are evaluated by more than one clinician. Levenson (2004) investigated the inter-rater reliability of the Florida SVP civil commitment criteria. The study was intended to determine if two evaluators would independently reach the same conclusions about the same sex offender. In addition to the inter-rater reliability of a number of relevant *DSM-IV* diagnoses, Levenson (2004) also examined the inter-rater reliability of several actuarial risk assessment instruments often used in these evaluations, as well as the recommendations ultimately made by the evaluators.

Levenson found that the inter-rater reliability of actuarial risk assessment instruments used to measure likelihood of reoffense was good (ICC = .77 - .85). This result was promising, and was consistent with previous research investigating the psychometric properties of the actuarial instruments (Barbaree et al., 2001; Bartosh, Garby, Lewis, & Gray, 2003; Bradley & Epperson, 2001; Harris, Rice, Quinsey, Lalumiere, Boer, & Lang, 2003).

Regarding the nine *DSM-IV* diagnoses studied, however, Levenson (2004) reported that the magnitude of the kappa coefficients was in the poor to fair range (kappa = .23 - .70), as defined by Bloom, Fischer, and Orme (1999). A similar result was found for the overall recommendation decision made by the evaluators (kappa = .54), with this decision also characterized as showing poor reliability. Such findings were alarming, as they raised concerns about the credibility of civil commitment evaluations. After all, if two forensic evaluators disagreed on the diagnosis and/or the commitment decision, how would we know which one is right and which one is wrong? And what are the consequences to individual liberty and to public safety if the incorrect determination should prevail in the courtroom?

This apparently poor reliability of the *DSM-IV* criteria at first seemed consistent with the extant empirical, anecdotal, and theoretical literature which takes a critical approach to *DSM* nosology (Campbell, 1999;2004; Kirk & Kutchins, 1994; Marshall & Hucker, 2006; Marshall, Kennedy, &

Yates, 2002; Meyer, 2002; O'Donohue, Regev, & Hagstrom, 2000; Reid, Wise, & Sutton, 1992). Upon further reflection, however, we questioned whether kappa is really the best way to measure diagnostic reliability.

Conventional measures of reliability

Since *Daubert v. Merrell Dow Pharmaceuticals, Inc. and Kumho Tire Co. v. Patrick Carmichael*, many jurisdictions have adopted standards for admitting scientific testimony that require the court to decide whether proposed testimony has a reliable basis in the knowledge and experience of (the) discipline (Wierner, 1999, p. 47). Experts rely upon studies that examine inter-rater reliability, and, when testifying, they need to be aware of the strengths and weaknesses of the various statistical analyses. For psychiatric diagnoses, inter-rater reliability is typically operationalized as the agreement between two or more independent evaluators regarding the presence of a mental disorder.

For such studies, the kappa coefficient has become known as the statistic of choice for measuring agreement (Uebersax, 1987, p. 140). It is easy to calculate and was intended to correct for random agreement -- in other words, to detect whether evaluators agree more often than would be expected by chance. Kappa takes the observed percentage of times that pairs of evaluators agree, or p_o, as well as the level of agreement expected by chance, p_c, to provide a chance-corrected measure of agreement, as represented in the formula, $k = (p_o - p_c) / (1 - p_c)$ (Cohen, 1960). Authors have applied interpretations regarding the magnitude of the kappa coefficient to describe a study s results (e.g. Bloom et al., 1999), thus providing an easily imparted meaning to statistical terms that can be used with a non-technical audience. Given these characteristics, it is little surprise that kappa has become a conventional measure in agreement studies.

However, important concerns regarding kappa have been noted by a number of authors. First, kappa assumes statistical independence between raters. In a two-rater study, each evaluator should generate a rating without knowledge of the other s. Kappa is not appropriate for a situation in which one observer is in the position of confirming or disconfirming a known previous rating (Sim & Wright, 2005).

Second, kappa assumes that the values in each diagnostic category will be approximately equivalent. When investigating diagnostic decision-making, the data are typically nominal or dichotomous and are organized in a contingency table.

Figure 1: Two-by-two contingency table

The assumption that the values in the rows and columns of a 2 X 2 table will be similar is described as the expectation of *homogenous marginals*. Large disparities in either the vertical or horizontal marginals have been noted by Feinstein and Cicchetti (1990) to often produce paradoxical results, in that even when a large proportion of evaluators agree, kappa may be lower than expected. Feinstein and Cicchetti (1990) explained that the paradoxical effect of *disproportionate prevalence* is greater for larger values of kappa than for smaller ones; thus, ironically, large proportions of agreement can lead to surprisingly low kappa coefficients. This problem has also been noted by other authors as related to widely discrepant base rates within the diagnostic categories (Grove et al., 1981; Langerbucher, Labouvie, & Morgenstern, 1996; Uebersax, 1987). It has been suggested that kappa not be used when base rates are 5% or less (Grove et al., 1981). Sim and Wright (2005) advocated that researchers report a prevalence index to inform readers about the likelihood of a

disproportionate number of either positive or negative agreements, as these would result in a high number of chance agreements, thereby reducing the value of the kappa coefficient.[1] They further suggest that the computation of the Prevalence Adjusted Bias Adjusted Kappa (PABAK, Byrt, Bishop, and Carlin, 1993) be reported to indicate the likely effects of prevalence and bias in the study.

Third, kappa is also known to give paradoxical results when rates of agreement differ substantially in the proportion of positive or negative cases (the *b* and *c* cells in a contingency table) (Sim and Wright, 2005). This phenomenon is termed bias and results in the second paradox described by Feinstein and Cicchetti (1990). The less that two coders agree, overall, the fewer chance agreements are estimated in kappa. Therefore, in situations in which there is *less* agreement in the *b* and *c* cells, the kappa will tend to be higher, which is contrary to the intent of the statistic. In order to advise readers of the possible presence of this phenomenon, Sim and Wright (2005) recommended that researchers routinely report a bias index in studies using kappa.[2]

Fourth, various authors have prescribed differing interpretations of the kappa coefficient. Bloom, Fischer and Orme (1999) recommended that kappas of less than .60 be described as indicating poor agreement, kappas from .60 to .74 as fair , and those of .75 and above as indicating good agreement. Landis and Koch (1977), however, recommended different descriptions and defined kappas of .20 or less as indicating slight agreement, those from .21 to .40 as fair , .41 to .60 as moderate , .61 to .80 as substantial , and those from .81 and above as almost perfect agreement (Landis & Koch, 1977). Muñoz and Bangdiwala (1997) recommended yet different ranges and descriptors for 3 X 3 and 4 X 4 contingency tables. Uebersax (1987) noted that the *chance correction* (p_c) element of the kappa equation represents the level of agreement expected under a null hypothesis that raters were making their conclusions on a random basis. In circumstances where there is evidence to reject the null hypothesis of random decision-making, however, it is not clear how the chance correction influences the coefficient or how it should be interpreted. Uebersax (2001) recommended using the statistical significance of kappa to determine if raters agreed more than would be randomly expected, but that the effect size interpretations should be disregarded. Others have agreed that guidelines for interpreting the magnitude of kappa are somewhat arbitrary (Sim and Wright, 2005) and may be deceiving (Hripcsak & Heitjan, 2002).

Alternative measures of reliability

There are a number of other ways of measuring agreement, and while no consensus exists as to the superiority of any one method, the selection of the method used should be made according to hypotheses under consideration and the nature of the data (Uebersax, 2001).

Simple descriptive statistics in the form of proportions of agreement have several advantages over more complex statistical measures. Proportions of agreement calculate the number of cases in which two independent evaluators agree that a diagnosis exists (positive agreement), or that the diagnosis does not exist (negative agreement). Disparity between raters can be calculated and the null hypothesis that the raters are independent can be tested using a Pearson chi-square. These frequencies are then combined in order to determine the overall proportion of cases in which forensic examiners agree on whether the subject met criteria for the particular diagnosis. These indices are important descriptive statistics, although they are often neglected in favor of seemingly more sophisticated statistical measures (Uebersax, 1987).

The odds ratio is another method often used in assessing occurrence of a diagnosis, or agreement, in two groups. In an agreement study, the odds ratio provides a measure of the likelihood that a second clinician will arrive at the same judgment as the first clinician. It is not influenced by the frequency of decisions in any particular direction, but rather is the odds of agreement divided by the odds of disagreement. The odds ratio results in values between zero and infinity, with 1 being the neutral value. Odds that approach zero or infinity mean increasingly large differences such that an odds ratio larger than one means that the first group has a larger proportion than the second group.

If the two groups are reversed, the odds ratio will take on the inverse (1/OR). The odds ratio is advantageous if the researcher is also using logistic regression (American College of Physicians, 2000).

Odds ratios are commonly used, though often difficult to intuitively understand. The risk ratio, also known as the relative risk, is easier to understand, since it provides probabilities rather than odds. It has many of the advantages of the odds ratio and uses the same scaling, but since the risk ratio uses proportions of agreement, it is closer to how most people seem to think when comparing the relative likelihood of events. For example, if group A has a 25% chance of mortality and group B has a 75% chance of mortality, the risk ratio of 3 is easier to interpret than the odds ratio of 9. For these reasons, reporting the risk ratio is often preferred, although both ratios include the equivalent information and are equally valid measures (American College of Physicians, 2000).

Two other statistics applicable in agreement studies are the positive predictive value (PPV) and negative predictive value (NPV). These statistics are often used in medical research to indicate the probability that a subject identified as having a disease actually has the disease, or in the case of the NPV, the probability that the subject does not have the disease when testing negative for its presence. The PPV is calculated as the number of true positive cases divided by the sum of the true positive and false positive cases (PPV = TP/(TP + FP)). The NPV is the number of true negative cases divided by the sum of the true negative and false negative cases (NPV = (TN + FN)). They are not appropriate for use in samples where the prevalence of the disorder was artificially controlled (Simon, 2005), so their use in a field study is desirable. Used in the context of an agreement study, the PPV indicates the likelihood that both evaluators will agree on the presence of a diagnosis, given that the first evaluator concludes that it is present. The NPV indicates the likelihood that both evaluators will agree on the absence of a diagnosis, given that the first evaluator concludes that it is absent. The PPV and NPV are also sensitive to the prevalence rates in the sample, such that samples with low prevalence will tend to have a lower PPV and higher NPV (Riddle & Stratford, 1999).

Purpose of the Study

The purpose of this study was twofold: (1) to re-examine Levenson s (2004) data using various statistical analyses to investigate the inter-rater reliability of the *DSM-IV* diagnostic categories used to determine whether a sex offender meets criteria for civil commitment; and (2) to illustrate the paradoxical interpretations of the kappa statistic using Levenson s (2004) data. Specifically, using alternative methods for establishing inter-rater reliability, this study set out to explore how often independent evaluators reached the same conclusions about sex offender clients who are assessed for civil commitment. Multiple analyses were utilized to examine the degree of agreement between forensic examiners.

Because unreliable outcomes in the civil commitment selection process can have grave consequences both for sex offenders and communities, this topic was considered to be an important area of inquiry. Few investigations have been conducted into the inter-rater reliability of Paraphilia diagnoses, and the literature on the reliability of *DSM* diagnoses in general is fairly limited. Often, reliability studies are simulated, with raters specially trained and factors controlled to maximize consistency. This study examines the inter-rater reliability of an important diagnostic application in a real-world setting.

Method

Sample

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 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 being hired to conduct SVP evaluations under Florida s Jimmy Ryce Act ranged from zero to eighteen years (mean = 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 all under age 18 comprised slightly less than half of the sample (45%). Mixed offenders (30%) had both adult and minor victims.

Data Collection Procedure

The original research was conducted in accordance with prevailing ethical guidelines and was approved by an Institutional Review Board. Data were collected through file review of SVP evaluation reports provided by the Florida Department of Children and Families. *DSM-IV* diagnoses were coded dichotomously (yes/no) and included those most commonly considered: Pedophilia, Sexual Sadism, Exhibitionism, Paraphilia NOS, Antisocial Personality Disorder, Personality Disorder NOS, Other Personality Disorder, Substance Use Disorder, and Other Major Mental Illness (major mood disorder or psychotic disorder). The evaluators recommendations for civil commitment (yes/no) were also recorded. Levenson s complete data set was accessed for the analyses utilized in the current study. Sampling and data collection procedures have been described in more detail elsewhere (Levenson, 2004).

Data Analysis Procedures

Data were analyzed using SPSS Version 12 (SPSS, 2004). Chi-square procedures were used to calculate the proportion of agreement for pairs of raters. Two-by-two tables were generated in order to determine the observed values in each cell (see Figure 1). The cells report the number of cases in which two independent evaluators agreed that the diagnosis existed (positive agreement), or that the diagnosis did not exist (negative agreement). In the remaining two cells, the values represent the frequency of disparity between raters. These frequencies were combined in order to calculate the overall proportion of cases in which evaluators disagreed on whether the subject met criteria for the particular diagnosis.

The data were also analyzed to calculate proportions of raw agreement and confidence intervals using SisPorto 2.1 (Ayres-de-Campos, Bernardes, Garrido, Marques-de-Sá, & Pereira-Leite, 2000). Marginal homogeneity was tested using McNemar tests. Odds ratios, relative risk and positive and negative predictive values were calculated using JavaStat Contingency Table Analysis (Pezzulo, 2005).

Results

Table 1 depicts the kappa coefficients, which were statistically significant ($p < .001$) for all diagnoses and for the overall commitment recommendations[3], thus rejecting the null hypothesis that the evaluators level of agreement was no better than chance. One diagnosis, Sexual Sadism, had marginals that significantly departed from homogeneity, as evidenced by the McNemar s test. Therefore, kappa results for this diagnosis are not likely to be credible due to the known distortions with imbalanced marginals. Using the non-significant McNemar tests as the criterion, the kappa results for the other diagnostic categories and the overall commitment recommendations do not appear to be compromised by significantly imbalanced marginals.

Using the Bloom et al (1999) rating schema, all but two diagnostic categories were judged to be in the poor range, with Pedophilia and Other Mental Illness in the fair range. Using the Landis and Koch (1977) ratings, four diagnostic categories (Sexual Sadism, Paraphilia NOS, Personality Disorder NOS, and Other Personality Disorder) were considered to have fair agreement. Four other categories (Exhibitionism, Antisocial Personality Disorder, Substance Use Disorder, and the overall civil commitment recommendation) were in the modest range of agreement and two categories were considered to have substantial agreement (Pedophilia and Other Mental Illness).

Prevalence and bias indices were computed in order to indicate whether the paradoxical effects described by Feinstein and Cicchetti (1990) were present in this study. These results, also in Table 1, indicate that the Prevalence Index (PI) was in excess of .80 for three diagnostic categories: Sexual Sadism (PI = .95), Exhibitionism (PI = .87) and Other Personality Disorder (PI = .93), thus indicating that the proportion of chance agreement used in the kappa coefficient was likely to be high, thereby reducing kappa, even in the presence of substantial agreement. The Bias Index (BI) results were uniformly low in this sample, with the BI less than .05 for all categories. This result indicates that the kappa coefficients were not paradoxically increased due to a disproportionate number of disagreements between the evaluators on positive or negative cases. Computing the Prevalence Adjusted Bias Adjusted Kappa statistic (PABAK) indicates that kappa is influenced by the presence of prevalence and/or bias on several diagnostic categories, particularly Sexual Sadism ($k = .30$, *PABAK* = .93), Exhibitionism ($k = .47$, *PABAK* = .87), and Other Personality Disorder ($k = .29$, *PABAK* = .91). The effects were less, but still notable in two other categories, Other Mental Illness ($k = .70$, *PABAK* = .87) and the overall Civil Commitment Recommendation ($k = .54$, *PABAK* = .65).

Table 1: Kappa and related statistics for diagnoses

Diagnoses	n	Kappa (95% CI)	Bloom et al. Rating	Landis & Koch Rating	Prevalence Index	Bias Index	PABAK	McNemar's
Pedophilia	277	.65*** (.55 - .75)	Fair	Substantial	.36	.00	.70	.00
Sexual Sadism	277	.30*** (-.03 - .63)	Poor	Fair	.95	.02	.93	5.44*

Exhibitionism	277	.47*** (.26 - .68)	Poor	Modest	.87	.01	.87	.89
Paraphilia NOS	277	.36*** (.25 - .47)	Poor	Fair	.12	.01	.36	.18
Personality Disorder NOS	277	.23*** (.11 - .35)	Poor	Fair	.44	.01	.38	.05
Antisocial Personality Disorder	277	.51*** (.41 - .61)	Poor	Modest	.20	.04	.53	1.86
Other Personality Disorder	276	.29*** (.02 - .56)	Poor	Fair	.93	.02	.91	1.92
Substance Use Disorder	277	.43*** (.32 - .54)	Poor	Modest	.04	.00	.43	.01
Other Mental Illness	276	.70*** (.57 - .83)	Fair	Substantial	.75	.00	.87	.00
Civil Commitment Recommendation	295	.54*** (.43 - .65)	Poor	Modest	.65	.01	.65	.31

Bloom Rating (Bloom et al); Landis & Koch Rating (Landis & Koch); PABAK = Prevalence Adjusted Bias Adjusted Kappa; * $p < .05$, ** $p < .01$, *** $p < .001$

Table 2 illustrates the odds and relative risk ratios which indicate that, regardless of the direction of the decision, evaluators were likely to agree on the diagnosis. For the odds ratio (OR), if one evaluator diagnosed a subject with Pedophilia (or not), another evaluator is 26 times more likely to agree than disagree. When diagnosing Sexual Sadism, the two raters were 67 times more likely to agree with each other than not. The poorest agreement is with Personality Disorder NOS, but even in this category evaluators were 3 times more likely to agree than disagree. For the overall Civil Commitment Recommendation, evaluators were 14 times more likely to agree than not. All of the 95% confidence intervals are above 1, meaning that receiving a diagnosis from one evaluator indicates that the odds were also greater that a second evaluator would agree.

The relative risk ratios (RR) are also provided. Although mathematically similar to the odds ratio, because they measure a proportion instead of odds, these data indicate that if a person was diagnosed (or not) with Pedophilia by one evaluator, there is a 7 fold increased likelihood that the second evaluator would do the same. The categories with the strongest relative risk ratios also included Sexual Sadism (RR = 53), Exhibitionism (RR = 13), Antisocial Personality Disorder (RR = 4), Other Personality Disorder (RR = 13), and Other Mental Illness (RR = 20). The relative risk ratio was not lower than 2 for any of the categories and none of the 95% confidence intervals included 1. The lowest relative risk ratios were found with Paraphilia NOS (RR = 2), Personality Disorder NOS

(2), Substance Use Disorder (RR = 2), and the overall Civil Commitment Recommendation (RR = 2). Even these, however, indicate a greater than two-fold likelihood of the evaluators arriving at the same conclusion.

Table 2: Odds and Risk Ratios for diagnoses

Diagnoses	n	Odds Ratio	95% CI	Relative Risk	95%CI
Pedophilia	277	25.75	(13.25 - 50.03)	6.84	(4.76 - 9.74)
Sexual Sadism	277	66.50	(7.73 - 555.86)	53.40	(7.29 - 399.82)
Exhibitionism	277	29.22	(9.46 - 90.77)	13.35	(6.46 - 23.98)
Paraphilia NOS	277	4.48	(2.70 - 7.44)	2.22	(1.71 - 2.87)
Personality Disorder NOS	276	2.87	(1.64 - 5.02)	2.03	(1.41 - 2.87)
Antisocial Personality Disorder	277	10.19	(5.79 - 17.92)	4.01	(2.87 - 5.63)
Other Personality Disorder	276	21.67	(4.72 - 101.96)	12.81	(4.10 - 28.40)
Substance Use Disorder	277	6.27	(3.72 - 10.55)	2.46	(1.90 - 3.19)
Other Mental Illness	276	71.91	(26.47 - 195.61)	19.78	(11.00 - 34.38)
Civil Commitment Recommendation	295	14.15	(7.56 - 26.45)	2.45	(1.92 - 3.15)

The proportions of agreement between raters and the total agreement are displayed in Table 3. The overall proportions of agreement (Pa) ranged from .68 for Paraphilia NOS to .97 for Sexual Sadism, with all proportions of agreement being statistically significant. None of the 95% confidence intervals included zero. Evaluators agreed in more than 80% of their decisions for five of the nine diagnoses, as well as the overall Civil Commitment Recommendation. The diagnoses with less agreement were Paraphilia NOS (*Pa* = .68), Personality Disorder NOS (*Pa* = .69), Antisocial Personality Disorder (*Pa* = .76), and Substance Use Disorder (*Pa* = .71). The proportions of agreement on the absence of a diagnosis ranged from .54 for Substance Use Disorder to .97 for Sexual Sadism, again with none of

the 95% confidence intervals including zero, therefore indicating a high degree of agreement between the evaluators when a subject did not meet the criteria for a diagnosis. The proportions of agreement on the presence of a diagnosis were lower, ranging from .18 for Sexual Sadism to .79 for the overall Civil Commitment Recommendation. Two categories had 95% confidence intervals that included zero: Other Personality Disorder and Sexual Sadism.

The proportions of negative agreement were greater than the proportions of positive agreement for eight of the nine diagnostic categories, with Substance Use Disorder being the only exception; this was also true for the overall civil commitment recommendation. These indices have been described by Cichetti and Feinstein (1990) as analogous to specificity and sensitivity. Such a pattern implies that the evaluators were applying stringent criteria for inclusion in a diagnosis, with a preference given for eliminating false positives in favor of potentially allowing a greater proportion of false negatives.

Table 3: Proportions of agreement

Diagnoses	n	Pa	95% CI	Pa-	95% CI	Pa+	95% CI	X^2
Pedophilia	277	.85	(.81 - .89)	.80	(.74 - .85)	.62	(.53 - .71)	117.88***
Sexual Sadism	277	.97	(.95 - .99)	.97	(.95 - .99)	.18	(-.05 - .41)	4.65***
Exhibitionism	277	.93	(.91 - .96)	.93	(.90 - .96)	.33	(.15 - .51)	60.93***
Paraphilia NOS	277	.68	(.63 - .74)	.56	(.49 - .63)	.47	(.39 - .55)	35.06***
Personality Disorder NOS	276	.69	(.63 - .74)	.64	(.58 - .70)	.23	(.20 - .36)	14.04***
Antisocial Personality Disorder	277	.76	(.71 - .81)	.67	(.61 - .74)	.54	(.46 - .63)	72.97***
Other Personality Disorder	276	.95	(.93 - .98)	.95	(.93 - .98)	.19	(.00 - .38)	25.60***
Substance Use Disorder	277	.71	(.66 - .77)	.54	(.49 - .62)	.57	(.50 - .64)	50.95***
Other Mental Illness	276	.93	(.91 - .96)	.93	(.90 - .96)	.58	(.43 - .73)	134.51***
Civil Commitment Recommendation	295	.82	(.78 - .87)	.48	(.39 - .58)	.79	(.74 - .84)	84.63***

Pa = Proportion of agreement, overall; Pa- = Proportion of agreement, negative; Pa+ = Proportion of agreement, positive; $p \leq .05$; ** $p < .01$; *** $p < .001$.

Positive and Negative Predictive Values are outlined in Table 4. The Positive Prediction Value (*PPV*) provides the probability that both evaluators will agree on the presence of a diagnosis, given that the first evaluator concludes that it is present. The Negative Prediction Value (*NPV*) gives the probability that both evaluators will agree on the absence of a diagnosis, given that the first evaluator concludes that it is absent. When the first evaluator concluded that the diagnostic criteria were met for a disorder, the probability that the second evaluator would agree was greater than .70 for three of the disorders: Pedophilia (*PPV* = .76), Substance Use Disorder (*PPV* = .72) and Other Mental Illness (*PPV* = .73). These probabilities were greater than .50 for three other disorders, Exhibitionism (*PPV* = .56), Paraphilia NOS (*PPV* = .65) and Antisocial Personality Disorder (*PPV* = .67). The three remaining disorders had a lower probability of agreement on the presence of the disorder; Sexual Sadism (*PPV* = .20), Personality Disorder NOS (*PPV* = .45), and Other Personality Disorder (*PPV* = .43). When one evaluator concluded that a subject did not meet the criteria for the diagnosis, the probability that the second evaluator reached the same conclusion was greater than .70 for all of the diagnoses. When considering the overall Civil Commitment Recommendation, there was a high degree of agreement. When the first evaluator concluded that a subject met the global criteria, the probability that the second evaluator would concur was .89.

Table 4: Positive and Negative Predictive Values

Diagnoses	n	NPV	95% CI	PPV	95% CI
Pedophilia	277	.89	(.85 - .92)	.76	(.69 - .82)
Sexual Sadism	277	.99	(.99 - .99)	.20	(.06 - .28)
Exhibitionism	277	.96	(.94 - .97)	.56	(.35 - .74)
Paraphilia NOS	277	.71	(.66 - .75)	.65	(.58 - .71)
Personality Disorder NOS	276	.78	(.75 - .81)	.45	(.36 - .54)
Antisocial Personality Disorder	277	.83	(.79 - .87)	.67	(.61 - .73)
Other Personality Disorder	276	.97	(.96 - .97)	.43	(.16 - .73)
Substance Use Disorder	277	.71	(.65 - .76)	.72	(.67 - .77)
Other Mental Illness	276	.96	(.94 - .98)	.73	(.61 - .83)
Civil Commitment Recommendation	295	.64	(.55 - .71)	.89	(.86 - .92)

Discussion

These results indicate that kappa does have limitations in measuring and interpreting interrater reliability. Single coefficients such as kappa do not fully capture the complexity of rater agreement. Many of its statistical assumptions may not be easily met in field studies. As demonstrated here, interrater agreement in the majority of cases can result in paradoxically low kappa coefficients.

Small values of kappa do not necessarily represent poor reliability, and indeed, differing interpretations of kappa coefficients have been proposed throughout the literature.

Unlike physiological tests or pathology reports for medical conditions, there are no independent, concrete, gold-standard procedures that can definitively rule in or rule out the existence of most psychiatric disorders. Confirmation that a diagnosis is correct is generally measured by the degree of agreement between independent evaluators, leading to a determination of the reproducibility or precision of the reference standard (Hripcsak & Heitjan, 2002). Despite considerable criticism and even warnings to the contrary (Lantz & Nebenzahl, 1996; Uebersax, 1987), the kappa coefficient has often been used for this purpose. As illustrated here, however, kappa may be low in spite of high levels of agreement, and in such cases it may be erroneous to conclude that ratings are unreliable. Although kappa s limitations have been well-documented by scholars, in practice, few nosologists pay attention to these issues, using their preferred measure of diagnostic agreement on the basis of familiarity and custom (Langerbucher et al., 1996, p. 1285).

In this investigation of SVP civil commitment decision-making, raw agreement on the existence of the majority of the diagnosed disorders was rather high despite low values of kappa. (It also bears repeating here that although significance levels were erroneously omitted in Levenson s original report, kappa values were all statistically significant.) Alternative methods of assessing reliability consistently demonstrated that the agreement among the evaluators was better than implied by the kappa coefficients. For most of the disorders, odds ratios ranged from 10 to above 70. Furthermore, evaluators had a generally high degree of agreement on the absence of a disorder. Finally, the evaluators had a high level of agreement in their final recommendations.

The current results suggest that diagnoses with more specific, behaviorally anchored criteria have better agreement. The diagnoses that occur more prevalently (and with which evaluators are therefore ostensibly more familiar) also have better agreement. Less frequently occurring diagnoses and those that have confusing, vague criteria (e.g., sadism) have poorer agreement. The disorders with relatively poorer positive predictive values (Sexual Sadism, Personality Disorder NOS, Other Personality Disorder, and Exhibitionism) also had low base rates within the sample (less than 15%). The inconsistency of these diagnoses could be the result of vagueness in the diagnostic criteria, or a true reflection of the lower prevalence of these disorders in the SVP population, and/or insufficient experience of the evaluators. Categories that are not otherwise specified also provide less specific diagnostic criteria, leading to a certain degree of subjectivity in decision making. Doren (2002) identified multiple problems with diagnosing Paraphila NOS, and made suggestions for improvement of the diagnostic criteria, and these were discussed as well by Levenson (2004).

These data were not collected as part of a hypothetical, simulated, or retrospective study of diagnostic reliability. Rather, they were collected in a real-world context in which SVP evaluators were well aware that the decisions made would likely have profound effects on others. How the perceived importance of their decisions may have affected the evaluation process is really an empirical question to which we have no answer. But it seems reasonable to speculate that evaluators were exceedingly cautious. This caution may have led to evaluators concluding that a person met criteria only when they could point to specific evidence that would meet a requisite legal standard. Such a high threshold may have led to diagnostic disagreement because an evaluator could not meet that rigorous, possibly self-imposed, standard.

One of the most controversial areas of forensic debate has been the accuracy and reliability of *DSM* criteria (Campbell, 1999; Grove et al., 1981; Kirk & Kutchins, 1994; Meyer, 2002), particularly when it comes to diagnosing paraphilias (Campbell, 2004; Doren, 2002; Marshall, 1997; Marshall & Hucker, 2006; Marshall et al., 2002; Marshall , Kennedy, Yates, & Serran, 2002; O'Donohue et al., 2000). In SVP commitment decisions, evaluators are faced with the task of attempting to understand, describe, and communicate symptoms of a statutorily-defined mental abnormality using the *DSM-IV* as their standard of reference. Especially because the stakes are so high for both offenders and communities, such decisions should be able to withstand various tests of reliability and rise to a high threshold of diagnostic agreement.

This study has limitations, several of which are the result of it being an uncontrolled field study. It was impossible to control for the data considered by each evaluator. It is feasible that one evaluator had access to and considered information not available to another. Furthermore, the second evaluator may have been aware of the opinions of the first, thus calling into question the independence of the assessments. As well, it was not possible to set an *a priori* base rate for the disorders being considered. When the researcher cannot control for the prevalence of the disorders, these data indicate that attempting to compensate for chance agreement becomes a complex matter. In such circumstances the effects of uncontrolled prevalence or bias upon unitary coefficients like kappa, as well as other agreement measures like the Positive and Negative Predictive Values, can influence the results.

In sum, interrater reliability is a complicated matter that is best determined using multiple methods of analysis. Kappa, the conventional statistic used for this purpose, can be misleading when utilized exclusively. Descriptive ratings applied to kappa coefficients are varied and somewhat arbitrary, making them difficult to interpret. As illustrated in this study, SVP civil commitment evaluation appears to be a highly reliable process, despite the conclusions previously drawn by Levenson (2004). Forensic evaluators testifying in these matter should be aware of the concerns outlined here, and capable of fully informing the court about the complexities of interpreting diagnostic reliability. A broader comprehension of the complexities of reliability indices will allow experts to more precisely construe and convey studies of diagnostic agreement when testifying to the reliability of forensic procedures.

References

1. American College of Physicians. (2000). Effective clinical practice: Primer on probability, odds and interpreting their ration. Retrieved 1/11/06, from http://www.acponline.org/journals/ecp/mayjun00/primer.htm
2. American Psychiatric Association. (2000). Diagnostic and Statistical manual of Mental Disorders, Fourth Edition, Text Revision. Washington, D.C.: Author.
3. Ayres-de-Campos, D., Bernardes, J., Garrido, A., Marques-de-Sá, J., & Pereira-Leite, L. (2000). SisPorto 2.0: a program for automated analysis of cardiotocograms. Journal of Maternal Fetal Medicine, 9(5), 311-318.
4. Barbaree, H. E., Seto, M. C., Langton, C. M., & Peacock, E. J. (2001). Evaluating the predictive accuracy of six risk assessment instruments for adult sex offenders. Criminal Justice and Behavior, 28(4), 490-521.
5. Bartosh, D., 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.
6. Bloom, M., Fischer, J., & Orme, J. G. (1999). Evaluating practice: Guidelines for the accountable professional (3rd ed.). Boston: Allyn and Bacon.
7. Bradley, M. A., & Epperson, D. L. (2001). Inter-rater reliability study of the MnSOST-R. Unpublished raw data.
8. Campbell, T. W. (1999). Challenging the evidentiary reliability of DSM-IV. American Journal of Forensic Psychology, 17(1), 47-68.
9. Campbell, T. W. (2004). Assessing Sex Offenders: Problems and Pitfalls. New York, NY: New York Academy of Sciences.
10. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psycyological measurement, 20, 37-46.
11. Doren, D. M. (2002). Evaluating sex offenders: A manual for civil commitments and beyond. Thousand Oaks, CA: Sage Publications.
12. Epperson, D. L., Kaul, J. D., Huot, S. J., Hesselton, D., Alexander, W., & Goldman, R. (1999). Minnesota sex offender screening tool - revised (MnSost-R): Development

performance, and recommended risk level cut scores., from http://psych-server.iastate.edu/faculty/epperson/MnSOST-R.htm
13. Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I: The problems of two paradoxes. Journal of Clinical Epidemiology, 43(6), 543-549.
14. Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis. Theory and practice. Archives of General Psychiatry., 38(4), 408-413.
15. Hanson, R. K. (1997). The development of a brief actuarial scale for sexual offense recidivism. Ottawa: Department of the Solicitor General of Canada.
16. Hanson, R. K., & Thornton, D. (1999). Static 99: Improving actuarial risk assessments for sex offenders. (No. User report 1999-02). Ottawa: Department of the Solicitor General of Canada.
17. Harris, G. T., Rice, M. E., Quinsey, V. L., Lalumiere, M. L., Boer, D. P., & Lang, C. (2003). A Multi-site Comparison of Actuarial Risk Instruments for Sex Offenders. Psychological Assessment, 15(3), 413-426.
18. Hripcsak, G., & Heitjan, D. F. (2002). Measuring agreement in medical informatics reliability studies. Journal of Biomedical Informatics, 35, 99-110.
19. Jimmy Ryce Involuntary Civil Commitment for Sexually Violent Predators' Treatment and Care Act, Florida Statute 394.912 (1998).
20. Kansas v. Hendricks, 117 S. Ct. 2072 (U.S. Supreme Court 1997).
21. Kirk, S. A., & Kutchins, H. (1994). The Myth of the Reliability of DSM. The Journal of Mind and Behavior, 15(1), 71-86.
22. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.
23. Langerbucher, J., Labouvie, E., & Morgenstern, J. (1996). Measuring diagnostic agreement. Journal of Consulting and Clinical Psychology, 64, 1285-1289.
24. Lantz, C. A., & Nebenzahl, E. (1996). The behavior and interpretation of the Kappa statistic: resolution of two paradoxes. Journal of Clinical Epidemiology, 49, 431-436.
25. Levenson, J. S. (2004). Reliability of Sexually Violent Predator Civil Commitment Criteria. Law & Human Behavior, 28(4), 357-369.
26. Marshall, W. L. (1997). Pedophilia: Psychopathology and theory. In D. R. Laws & W. O'Donohue (Eds.), Sexual Deviance. New York: Guilford Press.
27. Marshall , W. L., & Hucker, S. J. (2006). Issues in the Diagnosis of Sexual Sadism. Sex Offender Treatment, 1(2). http://www.sexual-offender-treatment.org/40.0.html
28. Marshall, W. L., Kennedy, P., & Yates, P. (2002). Issues concerning the reliability and validity of the diagnosis of sexual sadism applied in prison settings. Sexual Abuse: A Journal of Research & Treatment, 14(4), 301-311.
29. Marshall , W. L., Kennedy, P., Yates, P., & Serran, G. A. (2002). Diagnosing sexual sadism in sexual offenders: Reliability across diagnosticians. International Journal of Offender Therapy and Comparative Criminology, 46(6), 668-677.
30. Melton, G. B., Petrila, J., Poythress, N. G., & Slobogin, C. (1997). Psychological evaluations for the courts: A handbook for mental health professionals and lawyers. New York: Guilford.
31. Meyer, G. J. (2002). Implications of information-gathering methods for a refined taxonomy of psychopathology. In L. E. Beutler & M. L. Malik (Eds.), Rethinking the DSM: A psychological perspective (pp. 69-106). Washington, D.C.: American Psychological Association.
32. O'Donohue, W., Regev, L. G., & Hagstrom, A. (2000). Problems with the DSM-IV diagnosis of Pedophilia. Sexual Abuse: A Journal of Research & Treatment, 12(2), 95-105.
33. Pezzulo, J. C. (2005). 2-way contingency table analysis., from http://members.aol.com/johnp71/ctab2x2.html
34. Reid, W. H., Wise, M., & Sutton, B. (1992). The use and reliability of psychiatric diagnosis in forensic settings. Clinical Forensic Psychiatry, 15(3), 529-537.

35. Riddle, D. L., & Stratford, P. W. (1999). Interpreting validity indexes for diagnostic tests: an illustration using the Berg balance test. Physical Therapy, 79(10), 939-948.
36. Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Physical Therapy, 85(3), 257-268.
37. Simon, S. (2005). STATS, at Children's Mercy Hospital., from http://www.childrens-mercy.org/stats
38. SPSS. (2004). Statistical Package for the Social Sciences. Chicago, IL.: Author.
39. Uebersax, J. (2001). Statistical methods for rater agreement., from http://ourworld.compuserve.com/homepages/
40. Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. Psychological Bulletin, 101(1), 140-146.
41. Wierner, R. L. (1999). Extending Daubert beyond scientific expert testimony. APA Monitor, 30(6), 47.

Endnotes

1) Prevalence index = $\frac{|a-d|}{n}$

2) Bias index = $\frac{|b-c|}{n}$

3) It should be noted that the Kappa coefficients have been reported elsewhere (Levenson, 2004), but for ease of interpretation and comparison, they have been included here in Table 1. Importantly, statistical significance for Kappa coefficients was mistakenly omitted by the author in Levenson (2004) and is reported here in Table 1.

Author address

Jill S. Levenson, Ph.D., LCSW
Assistant Professor of Human Services
3601 N. Military Trail
Boca Raton , FL 33431
561-237-7925
E-mail: jlevenson@lynn.edu