Inaccurate Arguments in Sex Offender Civil Commitment Proceedings[1]

Dennis M. Doren, Ph.D.

Few, if any forensic area in which psychologists and psychiatrists do assessments is more contested than sex offender civil commitments. Seventeen states currently have laws that require evaluations of certain sex offenders for possible commitment. Conducting a civil commitment evaluation can result in the clinician being required to give days of court testimony, with scores of arguments being tested during each occasion.

Despite, or maybe because of the highly contested nature of sex offender civil commitment proceedings, there are some commonly made arguments that are nevertheless completely inaccurate. This chapter is designed to describe many of those inaccurate arguments. As the reader will see, some of these inaccuracies are regularly made by attorneys on either side of the litigation, and many are made by the expert witnesses themselves. The purpose of this chapter is to expose these errors, and indicate instead what is accurate.

Most of the delineated arguments below can be found in various court proceedings, meaning no one expert witness or attorney seems ultimately responsible for the persistence of the error. In those situations, this chapter will avoid naming any single individual as accountable for the fallacious argument. To single out one or two people from the many who make a common error would not be fair to those couple people. The fact is that most of the following inaccurate arguments cannot be found anywhere in professional writings. Maybe that is a good thing, but it also means that the reader may

[1] The opinions expressed in this chapter do not necessarily reflect those of the State of Wisconsin Department of Health and Family Services.

not easily find a published reference for the error being discussed. On the other hand, where a publication is known to exist espousing the inaccurate argument, the publication will be noted. Either way, the philosophy behind this chapter is that we are each fully responsible for what we offer to the courts, and need to be as accurate as possible in our testimony. It is hoped that this chapter will facilitate that process.

As a caveat, this review should not be considered comprehensive of all the inaccuracies made in sex offender civil commitment arguments and testimony. Selected here are simply what are considered to be the most common and most fundamental errors of a conceptual, statistical, and/or factual nature within sex offender civil commitment proceedings around the country.

There are four rubrics covered, each with its own set of multiple inaccurate arguments:

(1) Diagnostic issues:

- (a) Paraphilia Not Otherwise Specified, Nonconsent does not exist,
- (b) Anyone who rapes has a paraphilia,
- (c) Paraphilia NOS, Nonconsent is not in the DSM-IV-TR,
- (d) Everyone who molests a child is pedophilic.

(2) Conceptual issues related to risk assessment:

- (a) Prediction is the same thing as risk assessment,
- (b) Flawed instrument developmental procedures equate to "meaningless" outcomes,
- (c) All sexual recidivism base rates are the same,
- (d) Revisions of instrumentation mean they are not yet ready to be used,

    (e) "Actuarial" means based on historical (unchangeable) data only.

(3) Statistical issues:

    (a) Correlations and their derivatives are equal to predictive accuracy,

    (b) The ROC equates to accuracy of a risk assessment within civil commitment arena.

(4) Actuarial instrument usage issues:

    (a) The instruments have not been shown to have inter-rater reliability,

    (b) The instruments lack validity,

    (c) The instruments represent a novel principle,

    (d) Actuarial risk percentages below statutory threshold mean the person does not meet commitment criteria,

    (e) Structured clinical judgments are better than actuarial data,

    (f) The instruments are only good for screening purposes,

    (g) His category has 50% likelihood, but which half is he in?

    (h) The instruments are not "good enough", and nothing else is either.

Diagnostic Issues

The sex offender civil commitment process involves two main assessment areas appearing to require expert testimony: diagnostic and risk assessment. This first section addresses some commonly made inaccurate arguments involving diagnostic issues.

Paraphilia Not Otherwise Specified, Nonconsent does not exist

The diagnosis of Paraphilia Not Otherwise Specified (NOS) Nonconsent is probably the only commonly made diagnosis within sex offender civil commitments that is attacked based on the argument that the condition does not exist, as opposed to the more obvious idea a given diagnosis does not apply to the subject of the commitment or reexamination petition. (Hereafter, throughout this chapter, the subject, detainee, patient, resident, etc., will be referred to as the respondent, though it is recognized that the subject could also be the person who filed for re-examination of his or her commitment. This simplification is purely to make the chapter more easily read, rather than mixing terms, or regularly listing multiple terms.) The argument also differs from that which asserts the diagnosis does not exist in the latest Diagnostic and Statistical Manual (APA, 2000), an issue that is addressed below.

The bottom-line assertion in this first argument is that there is no paraphiliac condition (i.e., disorder of sexual arousal) specifically involving sexual arousal to non-consensual interactions with others. To test the accuracy of this assertion, that a certain condition does not exist, all we need to find is a single example to the contrary.

There are, in fact, various places where one can document the existence of people who show sexual arousal specific to non-consenting interactions within a sexual context (cf., Doren, 2002; Lalumière, Quinsey, Harris, Rice, & Trautrimas, 2003). Treaters commonly can describe cases in which the person has acknowledged ongoing sexual fantasies, urges, and historical behaviors involving raping others. One such case example is briefly described below, for the reader who may not have met such a person.

The person was middle-aged at the time I became familiar with him. He had been incarcerated for some years already, most recently for a sexual assault of an adult. He was not psychopathic (Psychopathy Checklist - Revised = 18; PCL-R; Hare, 1991), and expressed a good deal of upset about his repetitive history of rape. Penile plethysmographic (PPG) data and his self-report concurred in showing him to become far more sexually aroused with depictions of rape compared to consensual sexual contact. Even so, there was no indication that the specific infliction of pain or injury to others was sexually arousing for him. He complained of ongoing sexual fantasies of raping, fantasies that bothered him but over which he felt a significant lack of control. Occasionally, he also expressed urges to rape, though he was able to deal with those without actual raping while in institutions. In treatment, he described having these thoughts and desires since he was an adolescent, a report substantiated by his recorded criminal history.

Given examples like this exist, demonstrating a paraphilia related to raping, it is inaccurate to say the condition does not exist.

Anyone who rapes has a paraphilia

This is the flip side to the argument above. Instead of no one having a paraphilia related to raping, everyone who rapes has such a condition. This argument is just as inaccurate as that described earlier.

Again, this issue raises an empirical question. The question above was can anyone document such a condition. With this argument, the empirical issue is whether or not anyone who rapes can be found who does not show signs of a paraphilia.

In fact, research results with such a finding can be located without trouble. In Marshall & Barbaree (unpublished), only about 30% of convicted rapists showed sexual arousal to depictions of rapes while still showing arousal to depictions of consensual adult sexual contact. Lalumière et al. (2003) found that about 65% of rapists showed clear arousal to depictions of raping, meaning 35% did not. The fact of having committed rapes does not properly translate to the automatic conclusion that the perpetrator is paraphilic.

Paraphilia NOS, Nonconsent is not in the DSM-IV-TR

If the first inaccuracy above represents a faulty defense expert's statement, and the second error above represents a faulty prosecution expert's statement, then the issue here is one from defense attorneys themselves. The argument typically comes from cross-examination of prosecution witnesses when the diagnosis has been made concerning a respondent. The relevant questioning is usually put in the form something like "Isn't it true that the diagnosis 'Paraphilia NOS nonconsent' does not exist in the DSM-IV-TR?"

The accurate answer is of course it does, on page 576 of the hard cover version. The inaccuracy made here confuses the diagnosis of Paraphilia NOS with a descriptor called Nonconsent. The diagnosis title is Paraphilia NOS, with this phrase standing alone as the diagnosis. When added, the inclusion of the descriptor "Nonconsent" is for the purpose of clarity in professional communication. (The general explanation of an NOS diagnostic determination is on page 4 of the same manual. In this case, the descriptor, of "Nonconsent", indicates the NOS diagnosis was made for the second reason enumerated on page 4: "The presentation conforms to a symptom pattern that has not been included in the DSM-IV Classification but that causes clinically significant distress or impairment", as opposed to three other possible reasons listed for an NOS diagnosis.) Other examples of possible descriptors listed for the same reason include hebephilia, pornography, and auto-asphyxiation. Saying that the phrase "Paraphilia NOS

Nonconsent" does not exist in the manual in its complete form obfuscates the distinction between the diagnosis and a descriptor added for clarity.

Arguments made about the manual's lack of inclusion of the diagnosis sometimes will acknowledge the Paraphilia NOS as the essential diagnosis, but then point out the lack of inclusion of "nonconsent" in the enumeration of possible types within the Paraphilia NOS description (on page 576). To be clear, the term nonconsent initially appears in the first sentence defining paraphilias overall: "The essential features of a Paraphilia are recurrent, intense sexually arousing fantasies, sexual urges, or behaviors generally involving 1) nonhuman objects, 2) the suffering or humiliation of oneself or one's partner, or 3) children or other *nonconsenting* persons that occur over a period of at least 6 months…" (p. 566, emphasis added). It is, in fact, a defining characteristic of various paraphilias.

Everyone who molests a child is pedophilic

This argument, most typically heard by implication in what prosecutors ask, represents the same faulty logic as the first two arguments above (involving "all" or "none" statements). Again, the only thing necessary to demonstrate the inaccuracy of an "everyone" statement is a single exceptional case. In fact, there are many such cases including some to be found in the professional literature. PPG studies with child molesters typically find many who show sexual arousal to children (the essential meaning of pedophilia), but a portion of offenders who do not (e.g., Barbaree & Marshall, 1988). An argument can be made that the sensitivity of PPGs is not perfect, so some pedophiles are missed through this method of testing. Missing some, however, cannot be meaningfully interpreted as saying that everyone who fails on a PPG to show sexual arousal to children is really a pedophile despite the lack of evidence.

Likewise, just as was described above concerning paraphilia related to raping, sex offender treaters and evaluators quite regularly know of people with whom they have worked who have molested a child without qualifying for the diagnosis of pedophilia. Reasons for such offending, besides being driven by paraphilic desire, include loneliness coupled with a sense of inadequacy with adults, psychopathic lack of concern about who is the partner, desire to hurt the child's parent, psychotic delusion, and others.

Conceptual issues relate to risk assessment:

This is the first of three sections in this chapter related to the risk assessment portion of civil commitment proceedings. This first section covers general conceptual issues, while the latter two sections relate to more specific topics within the risk assessment overall rubric.

Prediction is the same thing as risk assessment

There is a very common, but fundamental confusion between the process of making predictions and the process of making a risk assessment. The two different concepts are frequently used interchangeably as if they mean the same thing. This error leads to various other arguments that are then inaccurate as well.

First, to be clear, we need to describe the difference between the two terms. A prediction is a forecasting of an event or outcome. A person states what is projected *will* occur the future (in some specified time period, location, etc.). A risk assessment is an estimate of the likelihood for an event or outcome. (Risk assessments, in general, typically also involve estimates of harm or severity, frequency, and imminence, though those considerations do not typically apply within the civil commitment evaluation process.) The risk assessor states what is the estimated likelihood for a future event or outcome (again within a specified time period, etc.).

The difference can be described metaphorically. If you have a standard, evenly balanced coin, and are going to flip it, then there are two possible outcomes (ignoring the coin's landing on its edge); its landing on side A or its landing on side B. If you make a prediction for the coin to land on side A upon (the first) flipping, then you cannot know if that prediction is correct or not until the coin is actually flipped. In other words, your accuracy in making a prediction can necessarily only be known in the future, after the predicted event did or did not happen. Then, and only then, is when you determine you were completely correct, or completely incorrect, with no in between possibility. If, however, before flipping the coin, you were to do a risk assessment, you would look at the fact that there are two equal outcomes upon any flipping of the coin. This would lead you to conclude that there is 50% likelihood that the coin will land on side A (for instance) upon flipping. In contrast to the predictive process, you do not need to flip the coin, or go into the future at all to find out if you are accurate in your risk assessment. All you have to do is verify that your assessment of two equal possible outcomes is correct, again something that could be done without ever flipping the coin (by studying the characteristics of the coin itself). You can conclude that you are perfectly correct without ever flipping the coin. In fact, flipping the coin after the risk assessment to see a single outcome would not tell you that your risk assessment was correct or incorrect. The coin's landing on side A would not tell you that you were correct and its landing on side B would not indicate you were incorrect. You would always have been correct, before any flipping of the coin, when the risk assessment concluded with 50% likelihood for landing on side A, no matter what the result from the next flip of the coin.

We deal with probability in our lives all the time, and know better than to view such circumstances as predictions. We do not (typically) predict a bad thing will happen to us when we buy insurance. Instead we buy insurance because we are aware of the probability (or at least possibility) of a bad thing happening to us and wish to lowering the degree to which that bad thing will hurt us if it happens. We take an umbrella with us when the meteorologist states there is a 75% chance of rain not because we heard "it will rain", but because it costs us little to carry an umbrella compared to an assessed high likelihood for rain. We require licensure for someone to practice medicine, not because we see this as guaranteeing fine medical practice, but because the requirements for a license are viewed as lowering the likelihood that bad medical practice will occur.

With all of this being said, how the described conceptual inaccuracy occurs within civil commitment proceedings can be explicated. Sometimes, the confusion is stated straightforwardly in a single comment, such as the statement by Janus and Prentky (2003), "Risk assessment – the prediction of sexual recidivism…" (p. 1443) or even within an article title by Berlin, Galbreath, Geary, and McGlone (2003): "The use of actuarials at civil commitment hearings to predict the likelihood of future sexual violence" (p.377). Most typically, the initial confusion causes a compound set of subsequent errors. Campbell (2000, 2004) is the main proponent for this degree of confusion between prediction and risk assessment. He states his argument in the following ways, first from Campbell (2000), and then from Campbell (2004):

> …risk assessments made for sexual predator hearings will lead to one of the following four outcomes:
>
> (i) The offender is correctly classified as an individual who would commit future sexually violent offenses if released into the community.
>
> (ii) The offender is correctly classified as an individual who would not commit future sexually violent offenses if released into the community.
>
> (iii) The offender is incorrectly classified as an individual who would commit future sexually violent offenses if released into the community, but in fact would not commit such offenses.
>
> (iv) The offender is incorrectly classified as an individual who would not commit future sexually violent offense if released into the community, but in fact would commit such offenses. (p. 114)

> Arguments regarding a "likelihood threshold for sexual reoffending" are also disingenuous. Civil commitment proceedings reach decisions that ultimately amount to one of four outcomes: (1) true positive, (2) false positive, (3) true negative, or (4) false positive [sic: clearly the phrase "false negative" was meant, but not written here]. In other words, the outcomes of civil commitment proceedings do not equate to a continuum of "more or less likely" to reoffend. Instead, these outcomes assume one of two dichotomous events occurring-The offender will, or will not, reoffend. (p. 122)

Dr. Campbell may find the argument against his statements as disingenuous, but the fact is that his statements are contrary to each and every one of the current 17 sex offender civil commitment statutes. In none of them do the commitment criteria include the assessment or determination of whether or not the offender will offend. Instead, all of them describe that a specified degree of likelihood is required for commitment (Doren, 2002).

To emphasize how the espoused “prediction model” ignores the relevant statutory language, we can look at how the various commitment laws (coupled with related case law) describe differing degrees of risk as their commitment thresholds. For example, (1) California and North Dakota’s commitment thresholds have been determined specifically to be lower than “more likely than not” (e.g., North Dakota’s definition for “likely” is “propensity towards sexual violence is of such a degree as to pose a threat to others”, a threshold specifically not as high as “more likely than not”), (2) Iowa and Washington’s thresholds are “more likely than not”, and (3) Arizona, Illinois, and Minnesota’s thresholds have been determined specifically to be beyond “more likely than not” (i.e., described as “highly probable”, “much more likely than not” and “high probability”). Is it even reasonable to say that all of these are really just saying the same thing: will or will not reoffend? In any other arena, these widely different risk levels would be seen for what they are: widely different.

The differentiation between predictions and risk assessments is far more than semantic or minor. Various errors stem from this initial confusion, with resultant inaccurate arguments being brought into the courtroom.

One of the main secondary errors pertains to the discussion of the sensitivity and specificity of the actuarial instruments (cf., Campbell, 2004). Sensitivity and specificity refer to the degrees to which discriminations are made without error. Within the civil commitment realm, the supposed discrimination of interest is between people who would actually reoffend and those who actually would not reoffend (sensitivity meaning the degree to which all future reoffenders are included in the selection process and specificity meaning the degree to which all future non-reoffenders are excluded in that same process). If one believes that evaluators are really making predictions about which offender "will, or will not, reoffend", then a discussion about sensitivity and specificity might make some sense (though it really still does not, as explicated below). Without making predictions of who will and will not reoffend, however, computations concerning sensitivity and specificity cannot even be made, no less be argued accurately as relevant.

To put the above another way, let's go back to the coin-flipping metaphor. If you devised a system for making predictions concerning the coin's landing on side A, and you implemented that system, you could test how accurate your system is by making the predictions and flipping the coins hundreds of times. You could then also compute how many times side A came up that you correctly and incorrectly predicted (to determine the sensitivity of your system), and how many times side B came up that you correctly and incorrectly predicted (thereby determining the specificity). On the other hand, if you conducted a risk assessment of the likelihood that side A would come up when the coin was flipped, you would have determined there was a 50% probability without ever flipping the coin. If you then flipped the coin, its outcome would not have made you right or wrong. You could not count the outcome as a "positive" or "negative", your original assessment as a "true" or "false" positive or negative, and you could not compute sensitivity and specificity figures based on your original assessment. All of these concepts do not apply, and the numbers cannot be computed because you did not make a prediction in the first place. (For the record, the proper statistical measure for accuracy of a risk assessment is the confidence interval, or probability interval.)

The reason why any of this is important is because discussions about sensitivity and specificity within the civil commitment context usually occur for the purpose of concluding there is too much error, that the evaluators are using faulty instrumentation. Using improper statistics to make a conclusion, any conclusion is the real error here.

As an aside, there is another commonly made inaccurate argument made by people who talk about the sensitivity and specificity of civil commitment "predictions". Those discussions regularly talk about these concepts as if society is necessarily interested in minimizing both of these types of error, and both are of equal importance. Within the civil commitment realm, however, this is demonstrably not accurate. Given how selections are made for commitment referrals across the country, the real concern is that the respondents *assessed as meeting criteria* for commitment are properly assessed (Doren, 1998, 2001). Translating that into (improper) predictive terms, this means the real issue is that those predicted to reoffend would have really reoffended if released into the community. This is not the same as sensitivity (which in this context means the degree to which all actual future recidivists are correctly predicted as such), but a statistic called positive predictive power (PPP): the degree to which predictions of future recidivism are accurate. A high PPP, within this context, means that those respondents assessed as meeting criteria would really have been future recidivists if released. It means high accuracy for those the evaluators say "commit" to, but does not speak to the degree to which all future recidivists are predicted as such. The data are available demonstrating that no State tries to commit each and every future recidivist (Doren, 1998, 2001), meaning that discussions about inadequate sensitivity (Campbell, 2004; Lloyd & Grove, 2002) are improper even within the view that evaluators make predictions; representing error upon error.

There are other errors that flow from the original failure to differentiate predictions from risk assessments. Rather than belabor this main point, however, only some of these will be discussed, with these discussions occurring later in this chapter.

Flawed instrument developmental procedures equate to "meaningless" outcomes

An argument typically made concerning the Minnesota Sex Offender Screening Tool – Revised (MnSOST-R; Epperson et al., 1999), but sometimes generalized to other risk assessment instruments is that the instrument's developmental process was flawed, and hence, the resultant product is, at best inadequate. The most poignant description of this argument has been written by Lloyd and Grove (2002). (Unpublished works are rarely cited in this chapter, but an exception was made here due to the fact the Lloyd and Grove paper has been cited elsewhere within a published work; Grisso, 2003, as discussed by Knight.) In their paper, Lloyd and Grove argue that the MnSOST-R had a flawed developmental process, and this caused the instrument's accuracy to be "meaningless" compared simply to predictions made using recidivism base rates.

Issues related to using a "prediction" accuracy model were described above, and while applicable here, do not need to be reiterated. What is new here is the idea that "flawed developmental" procedures automatically lead to a meaninglessness resultant product. (The issue of whether or not there were sub-optimal procedures used in the development of the MnSOST-R, or any other instrument, will not be addressed here, as that is only a tangential issue to the inaccurate argument being discussed in this section.)

Metaphorically, we can think of the utility of penicillin. The development of penicillin was not only sub-optimal, it was accidental. Yet, the outcome was phenomenally positive. Of course, improvements could, and have been made to the original product since the original development, but the fact the developmental process was sub-optimal did not result in a "meaningless" product.

The point is that something useful can be developed without optimal or even scientifically standard procedures. Flawed developmental procedures lower the *likelihood* a new instrument will be useful, but the real test for the meaningfulness of the instrument is whether or not it works in the way it should when tested with various new samples. If an instrument is empirically found to work consistently across various samples (i.e., consistently shows results supportive to validity), the fact that there may have been sub-optimal developmental procedures only means that some potential improvement can be made to the instrument as it currently exists. Despite such a shortcoming, the current instrument is anything but "meaningless".

All sexual recidivism base rates are the same

There are various statistical reasons why recidivism base rates are important within a risk assessment (Doren, 2002). None of these issues seems to be brought up regularly in sex offender civil commitment testimony except one. That one concerns a comparison of the supposedly "true" sexual recidivism base rate and the risk threshold for commitment. The purpose for testimony concerning this topic is in an attempt to show how unlikely it is that the specific respondent actually meets the commitment threshold criterion. The usual comparison of this type shows that "the" sexual recidivism base rate is much lower than the commitment threshold, such that the rarity of recidivism makes it very difficult to predict who will be a recidivist accurately. In this argument, the base rate that is used is also presented as the risk assessment of the individual respondent, at least by implication.

Ignoring descriptive parameters. There are three flaws with the argument as typically made. First of all, "the" sexual recidivism base rate presented is regularly not an accurate portrayal of the base rate of relevance to sex offender civil commitment proceedings. There are three parameters in defining a sexual recidivism base rate with accuracy: (a) the relevant time period, (b) the outcome measure employed, and (c) the type of sex offenders included in the computation. Inaccurate arguments involving recidivism base rates quite regularly ignore one, two, or all three of these parameters.

Concerning the time period of relevance, the current implementation of all 17 civil commitment statutes involves the working definition of risk as pertaining to (certain) sexual reoffending over the respondent's remaining lifetime. In contrast, Lloyd & Grove (2002) make their statistical arguments using a 5-year sexual recidivism rate (borrowed from Hanson and Bussière, 1998). In reality, the typical respondent across the country has far longer than an average five years remaining to his/her expected life span. Average respondent ages across the states do not tend to go beyond the offenders' mid-40's, with some states such as North Dakota and Pennsylvania having a far lower median respondent age. The portrayal of "the" relevant sexual recidivism rate using shorter-term estimates ignores the empirically demonstrated fact that different follow-up time periods regularly demonstrate different average sexual recidivism figures, with recidivism rates (using any one specific type of measure) continuing to rise as the time period for follow-up is extended (Doren, 1998). For instance, 5-year sexual recidivism rates are only about half of the 25-year rate using the same outcome measure (of reconviction or rearrest) (Doren, 2002). This means that, on average, using a 5-year rate to represent life time risk represents an underestimation by at least half of the true rate.

Concerning the outcome measure employed to derive the base rate figure, the Hanson and Bussière average sexual recidivism figure (like virtually all such research-derived recidivism rates) was derived almost completely from studies of reconviction and rearrest rates, not actual reoffending rates. No commitment statute requires the assessment that the respondent will be caught and legally processed, however; only the likelihood for certain sexual reoffending. We know reconviction figures are lower than rearrest rates (e.g., Doren, 1998; Langan, Schmitt, & Durose, 2003), and tend to believe that both are underestimates of the true reoffending rates (Doren, 1998; Hanson, Morton, & Harris, 2003). Inaccurate statements about recidivism base rates result from ignoring the effects of these differences in outcome measurements.

Concerning the inclusion of statutorily ineligible offenders, some studies are cited that include a large proportion of probationers, incest offenders, exhibitionists, etc. These subpopulations of the sex offender population are typically ineligible for commitment (though exceptions exist, such as for exhibitionists). Probationers and purely incest offenders typically show lower sexual recidivism rates than do child molesters and rapists (with time period and outcome measures controlled; Doren, 1998). Exhibitionists show the opposite (Doren, 2002). Citing a base rate from a study that very disproportionately includes probationers and incest offenders compared to populations eligible for commitment represents a seriously flawed argument. (For example, Adkins, Huff, &

Stageberg, 2000, has been cited various times even though the study involves only a 3-year follow-up period with a large proportion of the sample being probationers.)

Analogously, the errors described above in defining base rates is similar to describing the amount of interest one will earn based on a bank deposit. We can talk about the average interest earned across a lot of people, but what would the figure mean? If we said that the average interest earned was $250, would that be good, bad, or indifferent? If that interest were earned within six months, would that be good? If you knew nothing more than that interest was earned based on an average deposit of $5000, might that not be seen as good; that is, until you learn it took 20 years to earn that interest when you change your assessment? Without knowing the time period over which the interest was earned, the specific interest rate, and the initial amount of the deposits, meaningful statements about the average interest people earn can be seriously misleading. A simple figure, without qualifiers, has no inherent meaning, and can be very misleading.

<u>The base rate equals the accuracy of the risk assessment.</u> The second flaw involving testimony concerning recidivism base rates stems from the use of a base rate as the determinant of the degree of accuracy in a risk assessment. This flaw is a carryover from the inaccurate argument described above where the process of prediction is confused with the process of a risk assessment. As applied here, the inaccurate argument is of the following type. If one makes the prediction that no one will reoffend, one will be wrong equal to the base rate. (For example, if one says the relevant base rate is 15%, and you predict that no one will reoffend, then your predictions would be in error 15% of the time.) Improvements upon that error rate can be very difficult due to the relative rarity of the recidivism acknowledged. (People who use the argument involving the

accuracy of predictions almost invariably cite rather low sexual recidivism base rates for their computations; cf. Campbell, 2004; Lloyd & Grove). Besides the issue of using base rate figures that are too low, the discussion of predictive accuracy is flawed in its assumption. As explicated above, sex offender civil commitment laws do not require that predictions be made, and evaluators of respondents do not make predictions of recidivism. The use of base rates to compute potential prediction error rates represents an improper process.

Applying non-specific base rates to every respondent. Thirdly, the use of even a well defined and statutorily applicable base rate to describe a respondent can represent an inaccuracy. The issue here is that the most proper "well defined and statutorily applicable" base rate needs to be applied to the respondent. People with different characteristics can be members of subgroups with different base rates. Research has shown, for example, that first time molesters of girls have lower average sexual recidivism base rates (over various time periods) than do people convicted several separate times for sex offenses at least sometimes against boys (e.g., Hanson & Thornton, 2000). The most accurate base rate for the first subgroup would not include members of the latter subgroup, and vice versa. Ultimately, these differences in base rates based on respondent characteristics are the basis for actuarial instrument score interpretations. The recidivism percentage corresponding to each instrument score represents a separate base rate for subgroups of sexual offenders.

As described, there are various ways in which inaccurate statements about sexual recidivism base rates make their way into inaccurate arguments in the courtroom. Essentially, if a statement about a sexual recidivism base rate does not include some descriptors related to the amount of time post-incarceration, the type of outcome measure, and the type of sexual offender involved, the statement has set the stage for inaccurate arguments to follow.

Revisions of instrumentation mean they are not yet ready to be used

The Static-99 (Hanson & Thornton, 2000) was developed by borrowing from the earlier instrument called the Rapid Risk Assessment for Sex Offender Recidivism (RRASOR; Hanson, 1997). The MnSOST-R was developed borrowing from the MnSOST.

The inaccurate argument made that uses the above facts states that because the instruments have only "recently" been revised, this means that they cannot yet be good enough to be applied to real life decision-making. This argument makes two assumptions that are faulty: (a) that once science finds something that works, all work on improving that thing stops, and (b) anything that can be improved is not sufficient for practical use. Of course these assumptions are quite regularly false, including in the area of risk assessment. The fact that improvements can be found does not imply earlier forms were not appropriate to be applied real-life, or that current forms are insufficient if work to improve those forms is ongoing.

Again, metaphorically, one can think of any number of medical procedures that were useful and important in their time, even though improvements on those procedures have since been made and implemented in real-life medical practice. The fact is we all hope improvements will always continue to be made, both in medical practice and in sex offender risk assessments.

"Actuarial" means based on historical (unchangeable) data only

There is a common misconception concerning actuarial instrumentation: that actuarial data are synonymous with historical and essentially unchangeable information. To be fair, this misunderstanding may stem from the fact that most current sexual reoffense risk actuarial instruments commonly used in sex offender civil commitment assessments do largely, sometimes solely use historical data.

Actuarial scales, however, do not necessarily have to use historical information, in total, or even at all. The Level of Service Inventory – Revised (LSI-R; Andrews & Bonta, 2001), for instance, involves numerous changeable characteristics. The Sex Offender Needs Assessment Rating (SONAR; Hanson & Harris, 2001) is comprised almost solely of characteristics that can show change over time. Actuarial assessment simply implies the use of data that are specifically delineated, involve specific coding rules, and involve specific interpretative schemes. Whether or not an actuarial instrument contains solely historical information was simply a reflection of how it was developed.

As a further comment, actuarial assessment procedures actually represent the basis for all empirically developed psychological testing. This is true for intelligence tests, personality tests, attitudes tests, etc. A review of existing psychological tests will show that a vast majority of what is tested within any of those measures is not historical in nature.

Statistical Issues

Testimony concerning statistical issues may be the most difficult for judges and juries to follow. The terms used are strange, technical, and do not easily translate naturally into life experience. It is not surprising, then, that some inaccurate arguments creep in to sex offender civil commitment proceedings when it serves someone's purpose to do so, or out of ignorance.

There are probably a wide variety of inaccurate arguments made involving statistical concepts. Because of the technical nature of statistical analysis, this chapter will concentrate on just two of the most common errors, beyond what has already been described above.

Correlations and their derivatives are equal to predictive accuracy

A vast majority of empirical studies trying to delineate risk and protective factors related to sexual reoffending have offered a correlational statistic in their summary of findings. Maybe that is why there is a common misconception about the meaning of that statistic in relation to the predictive accuracy of those same risk and protective factors. The fact is, as Quinsey stated (as quoted in Grisso, 2002) stated "correlations and percent variance accounted for are not measures of predictive accuracy; they are measure of association" (p.245). Given the frequency at which inaccurate arguments are made regarding this same concept, however, one can only surmise that this concept is not well understood.

The inaccurate arguments seem to come in one of two forms: (a) in terms of the correlation figure itself not being high enough, or (b) in terms of the derived "variance accounted for", again this derived figure being deemed not high enough. You can sometimes find both of these forms within the same testimony and written work (cf., Campbell, 2004; Wollert, 2002), as the thinking behind one is the same as the other.

The first argument states that (a) the correlation between supposed risk factor X and sexual recidivism is "just" a certain figure, (b) since correlational figures essentially go from 0.0 through 1.0 and the correlation described is far closer to 0.0, (c) this shows that the predictive accuracy of risk factor X still leaves far too much to be desired; i.e., is not showing enough predictive accuracy. This argument wrongly assumes that a correlational statistic is a measure of predictive accuracy, as will be explained below.

The second form of the argument takes the correlational statistic, multiplies it by itself (i.e., squares it), describes the resultant figure as a percentage (of variance accounted for) and compares that percentage to the full range from 0% – 100%. For example, through this process, a correlation of .30 would be multiplied by itself (to equal .09), put into percentage form (changing .09 to 9%), and compared to the range of 0%-100%. Using this type of process, the user is allegedly computing the degree to which the variability found in sexual recidivism outcomes (i.e., whether someone was found to be an actual recidivist or not) can be accounted for (i.e., explained by) the variability in risk factor X. This argument (cf., Berlin, Galbreath, Geary, & McGlone, 2003) again wrongly assumes that a correlational statistic is a measure of predictive accuracy. An additional fault is the assumption that the best measure of "variance accounted for" is the square of the correlation, an assumption that is statistically questionable (Ozer, 1985).

Two not-so-technical methods will be used here to demonstrate the impropriety of viewing a correlation as equal to predictive accuracy. More statistical methods for this demonstration exist, but will not be offered here. The explication here will be of a type thought more useful in courtroom testimony, where sophisticated statistical analysis might not otherwise be well understood.

The more analytical method of the two for demonstrating that a correlation does not equate to predictive accuracy uses Figure 1. (In a courtroom, like elsewhere, a picture can be worth a thousand words.) In this figure, the risk measure is divided into 6

Insert Figure 1 here.

---

categories, numbered 1 through 6. The different columns represent the proportion of actual recidivists and non-recidivists that multiple large pieces of research have consistently found to have scored in each category. If you run a correlational analysis on these findings, you would discover only a small correlational relationship between the risk measure scores and the differentiation between actual recidivism and non-recidivism (if $n = 200$, $r = .16$).

The (inaccurate) argument often made would involve looking at this correlational figure and drawing the conclusion that this risk measure shows poor predictive accuracy. Likewise, if squaring the correlation was believed to be telling you something meaningful, the conclusion would again be drawn that this risk measure is virtually useless (that squared figure being described as representing less than 3% of the variance).

The actual predictive accuracy can be quite the opposite from those conclusions, however. In explanation, predictive accuracy, like the predictive process upon which it is determined, needs to have a division drawn between what will lead to a prediction in the affirmative (recidivism) versus a prediction in the negative (lack of recidivism). The determination of where the “line” should be drawn is crucial. Importantly, the process of deciding on a “cut-score” is not in keeping with what a correlational statistic typically does. A correlation describes the relationship between the whole measure and the outcome of interest (i.e., a correlation measures the degree of association). A cut-score treats a whole scale as if it has only two levels, above and below the cut-score.

Additionally, predictive accuracy is often viewed as if its determination must include all types of predictions made; that is, it must be computed including predictions made for both recidivism and non-recidivism. In practical circumstances, however, this is not always correct, in particular in the sex offender civil commitment realm.

The current implementation of every sex offender civil commitment statute involves some type of screening process by which not every offender thought to represent some recidivism risk is pursued for commitment. In fact, most of the actual future recidivists are knowingly not pursued (cf., Doren, 1998; 2001). Virtually all states screen out from any individual evaluator's consideration the majority of the people who will go on to recidivate sexually. Counting such "errors" in prediction (what some would improperly term "false negatives") as relevant to evaluators' judgments highly distorts what evaluators are in a position to do. Given the real life screening that occurs, the (improperly described) predictive issue for evaluators is not whether the risk measure is fully accurate in differentiating all future non-recidivists from all future recidivists. The real issue is specifically and solely the degree to which "predictions" of recidivism are accurate. In other words, when respondents are recommended for commitment because they are "predicted" to recidivate, how accurate are those predictions? (This issue was described above, in reference to the concept of positive predictive power.) This is the (predictive) question posed to courts in all sex offender commitment hearings, and is the real predictive accuracy of importance.

Within that context, and using the data from Figure 1, if we draw our predictive differentiation line (cut-score) at a score of 6 versus anything lower, we can see that the predictive accuracy is *perfect*. Every positive prediction for recidivism would have been accurate. An evaluator using this risk measure and this cut-score would be correct 100% of the time the evaluator said "will reoffend" (or more accurately, does meet criteria in terms of the risk assessment). Of course, in this illustration, most of the actual recidivists would not have been predicted to be such (i.e., there is low sensitivity), but this is simply analogous to how respondents are selected from among the complete sex offender population in the real world. (In this example, 5% of the actual future recidivists would have been assessed as such while 95% of the recidivists would not have been differentiated from the complete set of non-recidivists. This 5%, or what represents 2.5% of the total set of recidivists and non-recidivists, may seem like a small percentage, but the fact is that most states do not refer a much larger proportion of their convicted sex offenders for commitment; Doren, 1998, 2001.) The issue for the court in sex offender civil commitment proceedings is whether or not the referred individual (the respondent) meets criteria, not what portion of actual future recidivists were never referred for commitment. Referred respondents find themselves in that position at least partially because someone assessed the person as meeting criteria, so the question for the court is whether or not that assessment, and that specific assessment, is correct.

In summary, within the real life context, a measure showing a low correlation can still be a fine measure, even a perfect measure in terms of the predictive accuracy of relevance. The correlational statistic is simply the wrong one with which to make the determination of the predictive accuracy of a measure.

A second way to demonstrate the "degree of association" nature of correlations, versus their supposed relatedness to "accuracy of prediction" comes from real life examples. Meyer, Finn, Eyde, Kay, Moreland, Dies, Eisman, Kubiszyn, & Reed (2001) conducted a review of research in which a large number of correlational findings were summarized from a wide variety of contexts. As illustration, here are some of the findings that were summarized (quoted from portions of pages 130-137), with the correlations listed after each set of variables:

(1) Aspirin and reduced risk of death by heart attack: .02

(2) Chemotherapy and surviving breast cancer: .03

(3) General batting skill as a Major League baseball player and hit success on a given instance at bat: .06

(4) Coronary artery bypass surgery for stable heart disease and survival at 5 years: .08

(5) Effect of nonsteroidal anti-inflammatory drugs (e.g., ibuprofen) on pain reduction: .14

(6) Graduate Record Exam Verbal or Quantitative scores and subsequent graduate GPA in psychology: .15

(7) Scholastic Aptitude Test scores and subsequent college GPA: .20

(8) Sleeping pills (benzodiazepines or zolpidem) and short-term improvement in chronic insomnia: .30

One can see that even the highest of the above listed correlations (i.e., .30) would still represent less than 10% of the variance accounted for if computed by squaring that figure, despite the fact that this figure stems from a relationship that most people probably consider obvious: taking sleeping pills helps you get to sleep.

To put the above correlations into a real-life context to see how inappropriate this summary statistic is concerning predictive accuracy, we can take the smallest of the above figures rather than the largest. In a study by the Steering Committee of the Physicians' Health Study Research Group (1988), the researchers found part way through the planned study that the number of people dying from heart attacks from among the people taking aspirin was about half the rate from among people not taking aspirin; about a 50% relative improvement rate! The researchers were so impressed by these early findings that they felt it unethical to continue denying aspirin to people in the control group (i.e., the group not taking aspirin). The original study was halted for this reason. Such a dramatic finding, a 50% lower mortality rate, would seem to suggest that aspirin can be of high importance in lowering one's risk for dying from a heart attack, the conclusion drawn by the researchers. Still, the correlation between taking aspirin and the reduction in heart attacks was only .02, tiny by anyone's measure. How can the correlation between taking aspirin and the risk of death from a heart attack be so low given the dramatic results?

The answer lies in the fact that a vast, vast majority of the approximately total group of 22,000 people did not die from a heart attack during the length of the study. The "great" improvement in reducing risk by taking aspirin stemmed from a comparison of a the very small proportion of people who died from heart attacks in both groups, with the aspirin-taking group showing only half the other group's number. Statistically, however, when one takes into consideration all of the people in both groups who did not die from a heart attack, the *degree of association* between aspirin-taking and death from heart attack was very small. After all, it was only a very small subset of people who died from heart attacks in either group, and the vast majority of the "variance" (i.e., varying outcomes) in the correlational analysis was contributed by the people who did not die from heart attacks. This results in a very small correlation, despite an effect the researchers considered so dramatic that it was thought unethical to continue to withhold potentially life-saving treatment.

The point is that a correlational statistic looks at the degree of association between two complete set of variables. Under various circumstances, the degree of association is a poor surrogate for the true degree of predictive accuracy or utility of a measure or intervention.

The ROC equates to accuracy of a risk assessment within civil commitment arena

The Receiver Operating Characteristic (ROC) is another statistic often mentioned in sex offender civil commitment testimony. ROC figures are typically mentioned when an expert witness is discussing the relative utility of actuarial risk assessment instruments. Flawed arguments begin when an instrument's ROC figure is described as synonymous with the accuracy of a sex offender civil commitment risk assessment derived from the use of that same instrument. That argument is not correct, for two reasons.

The ROC statistic is computed by comparing the sensitivity and specificity of predictions made using cut-scores at each successive level of a measure, and then summing those comparisons across the complete scale. In other words, the ROC represents a statistic that can only be computed by presuming that it is proper to make predictions using cut-scores as thresholds. As described above in some detail, predictions of any kind are not the same thing as a risk assessment. Since ROCs are completely based on predictions, this statistic is improper for evaluating the accuracy of a risk assessment.

Additionally, the sex offender civil commitment evaluation does not involve specifically the determination of some absolute degree of risk. In other words, these evaluations do not necessarily involve determining if the respondent's relevant risk is at 22%, or 47%, or 52%. The relevant statutory question across the country is whether or not the respondent's risk is above or below a specified threshold; that is, above or below "more likely than not", for instance. A sex offender civil commitment risk assessment is therefore accurate to the degree that determinations are on target concerning whether respondents are above or below the specified legal threshold. Differentiations among "low" risk percentages (say of 12%, 22%, or 33%) do not matter relative to the conclusion of "below threshold" in a state where the legal threshold for commitment is above those figures (e.g., where the threshold is "more likely than not"). Likewise, discriminations among 55%, 67%, and 82% (again, for example) do not matter in that same state. Specific absolute levels of risk are not of primary relevance in a civil commitment risk assessment, but only the discrimination of above or below the legal threshold.

The ROC statistic does not tell us the accuracy of risk assessments of this type. This statistic adds up the degrees of error found across all possible cut-scores, or risk levels of a measure. Because a lot of this summation includes error that is not relevant to the specific risk assessment question at hand (i.e., above or below one threshold only), the ROC statistic necessarily gives a statistical summation of error that is not in keeping with the evaluator's task, or what the court needs to know.

For the record, the proper computation of the accuracy of a risk assessment within this context needs to consider the dichotomous nature of the determination using a threshold for risk as the differentiation point, something that is conceptually and statistically different from both the process of making predictions and the process of making determination of respondents' absolute degree of risk.

Actuarial Instrument Usage Issues

Probably the most contested aspect of current risk assessment procedures within sex offender civil commitment proceedings is the use of, and findings from actuarial risk assessment instruments. Maybe this is because courts are not used to methodical risk assessments beyond what is typically labeled "clinical judgment". Perhaps, the degree of contention concerning actuarial instruments simply reflects that making arguments against them (i.e., the more methodical process) is a lot easier than making arguments against a professional's general clinical judgment. Actuarial figures can be proven to be incorrectly computed, misinterpreted, or otherwise in error, whereas clinical judgments are less open to scrutiny concerning how the opinions are derived, and how much error is included in the conclusion drawn.

No matter the reason, the fact is that numerous arguments have been raised about the use of actuarial risk assessment instruments in sex offender civil commitment proceedings. This section describes many such arguments that are regularly made, but are nevertheless inaccurate. Some are technical arguments, while others reflect more conceptual matters of relevance to courts.

In he following, the reader can assume that the phrase “commonly used” in regards to actuarial instruments within the sex offender civil commitment evaluation setting refers to at least these three specific instruments: (a) the Rapid Risk Assessment for Sex Offender Recidivism (RRASOR; Hanson, 1997), (b) the Static-99 (Hanson & Thornton, 2000), and (c) the Minnesota Sex Offender Screening Tool – Revised (MnSOST-R; Epperson, Kaul, Huot, Hesselton, Alexander, & Goldman, 1999). Each topic being discussed will also typically be found to apply equally as well to other actuarial instruments not used as commonly as the above in commitment evaluations, such as the Violence Risk Appraisal Guide (VRAG; Webster, Harris, Rice, Cormier, & Quinsey, 1994) and Sex Offender Risk Appraisal Guide (SORAG; Quinsey, Harris, Rice, & Cormier, 1998).

The instruments have not been shown to have inter-rater reliability

Two sets of authors have made negative statements about the inter-rater of commonly employed actuarial risk assessment instruments. Campbell (2004) reviews a select set of research, and makes a distinction between “field reliability” (i.e., among people working in real life settings) versus “research reliability” (i.e., within research studies) to draw the conclusion that adequate research on inter-rater reliability is lacking for all actuarial instruments. Otto and Petrila (2002) make the statement that “…interrater reliability and measurement error are unknown for these instruments…” Both of these views are flawed.

By definition, inter-rater reliability is a *characteristic of the device* being tested, not of the raters employed. To test a device meaningfully, one must use raters well trained in the scoring system. The issue concerning inter-rater reliability is whether or not an instrument's coding rules are sufficiently precise, given real life data input, for trained raters largely to agree on instrument scores. The use of poorly trained or untrained raters cannot represent a meaningful empirical test of an instrument's inter-rater reliability. Such raters' errors all too easily reflect the raters' lack of knowledge of scoring rules, and not the sufficiency of the scoring rules themselves.

With this understanding in mind, the distinction between "research reliability" and "field reliability" is artificial and of no meaning. If the distinction were simply to point out that there are inadequately trained people who will use instruments anyway, then that point is granted. The fact improperly trained people use an instrument does not reflect on the instrument, though. It only reflects on the people who use an instrument when they are not trained to do so. If the distinction between "field" and "research" reliability is to reflect something inherent about real life versus research settings, then the distinction is fictitious. Virtually every piece of inter-rater reliability research concerning the commonly used actuarial instruments has involved real life cases with real life file materials.

This brings us to the second inaccurate argument: that the inter-rater reliability for the instruments is unknown. In fairness to Otto and Petrila, they wrote their article before some of the more recent research was conducted and published. Their article, however, is cited in more current literature (cf., Janus & Prentky, 2003) as if that same assessment is still current as well.

An enumeration of empirical tests of inter-rater reliability for three of the most commonly used risk assessment instruments has already been published. (See Doren, 2004, concerning the RRASOR and Static-99; and Doren and Dow, 2003, concerning the

MnSOST-R.) The first two instruments show inter-rater reliability figures ranging between .88 - .94 stemming from about 8 studies each. The latter instrument shows figures ranging from .80 - .86 across 5 studies. While someone may argue that the number of empirical tests completed to date is still insufficient (a general criticism discussed below), the various studies documenting inter-rater reliability figures clearly demonstrate the inaccuracy of any statement suggesting that the inter-rater reliability of these instruments is unknown. Additionally, any argument that there are insufficient numbers of studies of inter-rater reliability must also account for the fact that the dozens of validity studies of these instruments (described in the next section) demonstrate inter-rater reliability each time validity is supported.

One other comment should be made here. Janus and Prentky (2003, 2004) point out that no matter how one looks at the current actuarial instruments, their inter-rater reliability must be better than the other type of expert assessments regularly accepted by the courts: clinicians' judgments unstructured by empirical findings. To say that the inter-rater reliability of actuarial instruments is insufficient is also to say that *all* types of evaluators' testimony concerning risk are insufficient in this same regard. The inter-rater reliability of actuarial instruments is the highest we have within the context of a risk assessment. If the instruments are not good enough in this regard, nothing else is either. The argument that nothing is good enough is discussed below.

The instruments lack validity

This argument comes in various forms. The most straightforward says that the commonly used instruments, or even all actuarial risk assessment instruments, lack sufficient demonstrations of their validity to be used (e.g., Campbell, 2004). "Sufficient" is a judgment call, not a scientific standard. The judgment is based on the degree to which the instrument has been empirically demonstrated to work as it should, especially within the context to which it is to be applied.

To make a judgment about sufficiency, one needs to be aware of the empirical findings that exist. Over two dozen validity studies have been conducted with the RRASOR and Static-99 (each), with the results nearly uniformly supportive in both cases. Those studies stem from at least eight different countries, including various U.S. states (again, for each instrument). While someone may fault single pieces of research, the consistency of the findings across samples and jurisdictions seems clear. A recent meta-analysis including most of these studies' results found very supportive results as well (Hanson & Morton-Bourgon, 2004). The MnSOST-R has been investigated fewer times, using about eight different samples from three different countries and four different states within the U.S. Although people will sometimes cite three different published works as showing non-supportive results for the instrument's validity, only one accurately represents such a finding (Bartosh, Garby, Lewis, & Gray, 2003). The conclusion of non-support by Barbaree, Seto, Langton, and Peacock (2001) was altered by follow-up work by Langton (2003) using the same subject sample. Wollert's (2002) conclusion was found to be incorrect by Doren and Dow (2003), though Wollert (2003) argues otherwise.

So, is this sufficient to demonstrate validity? A vast majority of the country's state and court appointed sex offender civil commitment evaluators think so, as determined by a series of informal surveys concerning which instruments are used in these assessments. (Contact this author for details).

A second form of "the instruments lack validity" argument stems states that only a portion of the above cited research has been published, and unpublished works should not be considered (i.e., if the study has not gone through a peer-review process, then we cannot be so certain of its scientific merit). There are, of course, smaller numbers of published studies for the instruments testing their validity compared to the number of studies both published and unpublished. (To see an enumeration of published works for each instrument, see the reference list that can be found at www.atsa.com.) Again, the determination of "sufficient" is a judgment call, this time being made within the context of another judgment call that only published works matter in determining the validity of an instrument. The issue for people making this judgment is whether or not they make the same kind of judgment in other professional areas (such as evaluations related to competency, criminal responsibility, and child custody). A discrepancy in the standard used (in defining sufficient validity for using an instrument within a forensic context) indicates a bias that needs to be explained.

A third form of the validity argument concerns the degree to which the individual risk percentages associated with each scale risk category remains stable across different samples of sex offenders. Wollert (2002) made an argument for insufficient stability specific to one of the commonly used instruments (the MnSOST-R). As mentioned earlier, Doren and Dow (2003) found otherwise by re-analyzing the data. The stability of the risk percentages for the RRASOR and Static-99 were both found to be well supported with large samples by Doren (2004).

Of course, someone can argue that these findings are not enough, that more is needed. As described above, no matter how much research there may be, the argument there is not yet enough research can literally always be made. The issue then becomes whether or not the person's threshold for "enough" is consistent across various types of evaluations and applications.

In contrast to the above problematic arguments, there is one accurate argument that very rarely makes it into the courtroom during sex offender civil commitment proceedings. This argument was stated most eloquently by Quinsey, as written in Grisso (2003): "The accuracy of a particular instrument is underestimated in follow-up research by the unreliability of the outcome measure" (p. 245). We know that reconviction and rearrest rates (the usual instrument outcome measure) are only surrogates for what we are trying to measure, the true reoffending process. To the extent that reconviction and rearrest rates are in error in measuring true reoffense rates, the actuarial risk assessment instruments are constrained in their potential accuracy; not because of the design of the instrument, but because the instrument was developed with a flawed outcome measure. Error in measurement necessarily interferes with the demonstration of accuracy. From these points, we can conclude that the demonstrations of accuracy of our current risk assessment instruments are underestimations of what would be found if our outcome measure were more accurate.

The instruments represent a novel principle

This argument is typically offered when issues of evidentiary admissibility are raised. A consideration in some states for the admission of scientific information as evidence is that it does not represent a novel scientific principle. Unfortunately, there are some people who argue that the instruments represent just such a principle.

The fact is that the psychological assessment of individuals by comparison to systematically obtained group data goes back nearly a century. The first such process may have been the intelligence test called the Stanford-Binet. (Developed in 1916, this test actually was a revision by Lewis M. Terman of the 1908 Binet-Simon Scale.) Data from a group of people were obtained against which individual test scores were compared, and interpreted. Modern day psychological testing of virtually all types (intelligence, personality, attitude, etc.) maintains this same exact process. There is no basis in reality for saying it is a novel principle to assess an individual by comparing that person's scores to actuarial group data.

Of course, the application of this principle to risk assessment procedures is newer than one hundred years old. Even so, the application of actuarial procedures to risk assessments is older than what various people describe in testimony. Of actuarial risk assessment instruments that are still regularly used (at least in revised forms), probably the earliest were two instruments developed about the same time: (1) the Statistical Information on Recidivism scale (SIR; Nuffield, 1982); and (2) the Level of Service Inventory (LSI; Andrews, 1982). Both were developed about 22 years ago at the time this chapter was written. Of actuarial risk assessment instruments still very popular in their original form, the Violence Risk Appraisal Guide (VRAG, Webster, Harris, Rice, Cormier, & Quinsey, 1994) may be the oldest, being developed over a decade ago. Although some people may argue that "only" 10 years, or even 22 years of application is still novel (an argument that seems questionable on its face), the typical interpretation of the legal issue is actually not how novel the specific application of the scientific principle is, but how novel the principle is. The principle of comparing individuals to group data from which conclusions about the individual are drawn has a century-long history in psychology, with even the application to risk assessment procedures approaching a quarter century. It would seem that arguments about the instruments representing a novel scientific principle either ignore the above facts or push the time factor inherent in the concept of "novel" to the point of incredulity.

Actuarial risk percentages below statutory threshold mean the person does not meet commitment criteria

A common argument heard in civil commitment proceedings, by attorneys and expert witnesses alike, takes an actuarial risk percentage, compares it to the legal threshold for commitment, and finds it insufficient. At times, this conclusion can be perfectly accurate. There is an inaccuracy in the comparison, however, that can make the resultant conclusion also inaccurate.

The percentages attached to actuarial risk instrument scores describe empirical findings using specific outcome measures. Those outcome measures virtually always involved the subjects being reconvicted and/or rearrested. In contrast, none of the relevant statutes describes the issue of risk in terms of the person being reconvicted or rearrested for relevant crimes, but solely the likelihood for recommitting such a crime. There is already documentation that rearrest rates differ from reconviction rates within the same samples of sex offenders (Doren, 1998), and professional literature describing the accepted idea that true reoffense rates (over the same time periods) have to be larger than either of these surrogate measures. (For example, Hanson, Morton, & Harris, 2003, state "The observed rates underestimate the actual rates because not all offenses are detected…".) Likewise, the current actuarial instruments go out no further than 15 years in recidivism percentages, but research indicates that new first time sexual recidivism can still occur beyond that time period (cf., Doren, 1998; Hanson, Morton, & Harris), meaning that a person's risk even for being reconvicted or rearrested can also be underestimated by the current set of actuarial figures.

These differences between what actuarial figures represent and what the laws indicate is of relevance are ignored by some people. Wollert (2002), for example, uses the phrase "recidivism risk" both in relation to an actuarial instrument's risk percentages and the statutory risk threshold for commitment, as if there is no differentiation between the two. Campbell (2004) does the same thing.

The process of ignoring these definitional differences between what is measured and what the laws require for commitment allows for the potentially inaccurate argument to be made that any actuarial figure below the legal threshold means the respondent does not meet commitment criteria. That argument is flawed whenever proper consideration of factors beyond what any single actuarial scale measures move the assessed risk to above the commitment threshold despite a risk percentage that falls below.

Structured clinical judgments are better than actuarial data

There are some people who argue that the current set of actuarial instruments are simply not yet good enough to be used in forensic evaluations, an argument that is discussed below. Within that set of people, many then go on to say a different method of risk assessment, using structured lists of risk factors (otherwise called structured clinical judgment), is better and can be used.

The flaw in this argument stems from a mischaracterization of the relative utility and support for the two risk assessment methodologies. If the assertion is made that actuarial procedures are not yet good enough, but structured clinical judgments are, then there should be research showing the latter to perform better (either in risk assessments or in predictions of recidivism, both over the long-term). While a couple research results of that type can be found (e.g., deVogel, deRuiter, van Beek, & Meed, 2004), most empirical finding indicate that actuarial assessment is more accurate (Hanson & Morton-Bourgon, 2004). This is not to say that structured clinical procedures should be avoided, as they clearly have an important role in various types of risk assessments outside of the sex offender civil commitment realm (Doren, in press). It is only that asserting that structured clinical judgments are better than actuarial procedures in conducting sex offender civil commitment risk assessments is an empirically very questionable.

The instruments are only good for screening purposes

Berlin, Galbreath, Geary and McGlone (2003) appear to be the main proponents of this view. The argument stems from the initial conclusion they draw that the instruments are not sufficiently accurate in determining who will and who will not reoffend (i.e., predictions of sexual reoffending). If the actuarial instruments are not good enough in their predictive abilities, then the instruments can only be used within a screening process in assessing who should be committed and who should not.

The first flaw in their argument stems from the view that predictions need to be made. This issue was discussed in detail above, so it will not be reiterated here except to point out that without this flawed assumption, the inaccurate argument concerning the sole use for actuarials is for screening goes away. In other words, the rationale for saying the instruments can only be used for screening purposes stems from the flawed view that risk assessments and predictions are the same thing.

There is a corollary within this "only for screening purposes" argument that is also in error. Proponents of this argument state that because most of the people referred for commitment have a repetitive sexual offending history, have extrafamilial victims, have other (nonsexual) criminal histories, and have other such characteristics that are on current actuarial scales, then the scales cannot differentiate among these referred people between who will and will not reoffend. From this argument comes the conclusion the instruments can only be used to screen potential commitment candidates, and not beyond that point in the assessment.

This corollary argument lacks statistical meaning once placed back into the realm of risk assessment. To explain by analogy, it is similar to saying that once you screen a group of people for all of the high risk signs for cancer, you cannot differentiate between those with high risk for cancer and those with lesser risk for cancer. Of course, you cannot differentiate any further, because you already selected all of the high risk people in your "screening" process. If your task was to determine the complete group of high risk people, you already accomplished the task. No further work is needed. Your original differentiation process was all that was needed, whether we call it a "screening" or a complete assessment.

His category has 50% likelihood, but which half is he in?

I give credit to Harris (2003) for the title to this section, and quickly point out that the inaccurate argument in this regard certainly does not represent his view. The flawed argument states something like "the respondent may be in the 70% category for risk, but you cannot tell if he is in the 70% group who will reoffend or the 30% group who will not".

Again, this argument confuses prediction and risk assessment. If it is a fact that that respondent is accurately assessed within a certain risk category (let's say, in keeping with the example, 70%), then the risk assessment in that regard is complete. Any argument from that point on concerning whether or not the respondent will reoffend falls into the set of inaccurate arguments that stem from the confusion between risk assessment and prediction, what Harris (2003) describes as "not knowing the technical meaning of the word 'risk'" (p. 391). Unfortunately, this permutation on the inaccurate arguments can be found even in recent professional literature (cf., Berlin et al., 2003; Hart, 2003).

The instruments are not "good enough", and nothing else is either

This argument has been mentioned in various sections above due to its representing the bottom line to all arguments of "insufficiency" in risk assessment methodology. Nothing is good enough. As stated above as well, this argument ultimately represents a value judgment about what is "good enough", not a statement with an empirical basis. If a person sets the proverbial bar high enough, no risk assessment methodology, in fact no aspect of psychological science will be able to jump high enough to qualify as "good enough".

The reason why this value-laden argument is included here, among others that are more clearly simply inaccurate, is because there appears to be confusion even within the value-laden argument. The confusion is in defining "good enough" for what.

Campbell (2004) argues that "good enough" is what is defined by admissibility standards for scientific evidence (as often defined by *Frye v. United States* and *Daubert* *v. Merrell Dow Pharmaceuticals*). In his opinion, nothing science has to offer concerning risk assessments yet meets either of these evidentiary standards.

The view is certainly debatable, and in fact is not in keeping with the very large proportion of the country's courts that have already adjudicated relative to the admissibility of actuarial instrument testimony. Of importance here, however, is a completely different standard of relevance in determining "good enough for what". To this writer's knowledge, there are no or at least virtually no evaluators who solely rely on actuarial information in making conclusions about respondents' risk within the sex offender civil commitment context. Actuarial instruments typically serve as an anchor, or foundation upon which final clinical opinions are determined, but the instruments do not stand alone. The question, then, is the degree to which the instruments are "good enough" for this purpose.

The fact is that the courts are regularly willing, even determined to hear opinions about risk from expert witnesses. This is true apparently despite concerns about the accuracy of those opinions (cf., *Barefoot v. Estelle*). Therefore, the question "are the instruments good enough" does not pertain to legal standards for admissibility in court, but to professional standards in forming the basis for a forensic opinion.

Again, if the bar in making this determination is set high enough, the answer will always be no. The placement of "the bar" is a judgment call, not an empirically-based determination. If we place the bar's level by (a) considering the fact the courts need to elicit such testimony from someone (in order to conduct these matters of law), and (b) the fact that actuarial procedures represent the best we have to offer the courts in such elicited testimony, then the answer of yes to the question of "are they good enough" would seem quite reasonable professionally.

Summary Comments

The sex offender civil commitment process represents a very serious public policy. Individuals' life freedoms over potentially long periods of time are threatened, while society attempts to protect itself from the most dangerous sex offenders. Within this context, it would seem that professionals owe it to the courts, and to society in general, only to give the most accurate information to those people who are making the decisions in these cases. Inaccurate arguments should be avoided, and if made, corrected.

This chapter was designed to describe a number of the commonly made inaccurate arguments, ones that are repeated all too many times. Maybe Yogi Berra had it right when he said: "We made too many wrong mistakes". My hope is that this chapter will help diminish the degree to which this remains true.

References

American Psychiatric Association (2000). Diagnostic and Statistical Manual – Volume IV – Text Revision. Washington, D.C.: American Psychiatric Association.

Andrews, D.A. (1982). The Level of Supervision Inventory (LSI): The first follow-up. Toronto: Ontario Ministry of Correctional Services.

Andrews, D.A., & Bonta, J.L. (2001). The Level of Service Inventory – Revised user's manual. North Tonawanda, N.Y.: Multi-Health Systems.

Atkins, D., Huff, G., & Stageberg, P. (2000). The Iowa Sex Offender Registry and Recidivism. Iowa Department of Human Rights, Division of Criminal and Juvenile Justice Planning and Statistical Analysis Center.

Barbaree, H.E., & Marshal, W.L. (1989). Erectile responses among heterosexual child molesters, father-daughter incest offenders, and matched non-offenders: Five distinct age preference profiles. Canadian Journal of Behavioural Science, 21 , 70-82.

Barbaree, H.E., Seto, M.C., Langton, C., & Peacock, E. (2001). Evaluating the predictive accuracy of six risk assessment instruments for adult sex offenders. Criminal Justice & Behavior, 28(4), 490-521.

Barefoot v.Estelle (1983). 463 U.S. 880.

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

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

Campbell, T. W. (2000). Sexual predator evaluations and phrenology: Considering issues of evidentiary reliability. Behavioral Sciences and the Law, 18, 111-130.

Campbell, T.W. (2004). Assessing sex offenders: Problems and pitfalls. Springfield, Illinois: Charles C Thomas.

Daubert v. Merrell Dow Pharmaceuticals, Inc. (1993). 113 S. Ct. 2786.

Doren, D.M. (1998). Recidivism base rates, predictions of sex offender recidivism, and the "sexual predator" commitment laws. Behavioral Sciences and the Law, 16, 97-114.

Doren, D.M. (2001). Analyzing the analysis: A response to Wollert (2000). Behavioral Sciences and the Law, 19, 185-196.

Doren, D.M. (2002). Evaluating sex offenders: A manual for civil commitments and beyond. Thousand Oaks, CA: Sage Publications.

Doren, D.M. (2004). Stability of the interpretative risk percentages for the RRASOR and Static-99. Sexual Abuse: A Journal of Research and Treatment, 16(1), 25-36.

Doren, D.M. (in press). Recidivism risk assessments: Making sense of controversies. In W. Marshall, Y. Fernandez, & L. Marshall (Eds.). Sexual offender treatment: Issues and controversies. West Sussex, UK: John Wiley & Sons.

Doren, D.M., & Dow, E.A. (2003). What "shrinkage" of the MnSOST-R? A Response to Wollert (2002b). Journal of Threat Assessment, 2(4), 49-64.

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. Available at http://www.psychology.iastate.edu/faculty/epperson/mnsost_download.htm .

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

Grisso, T. (2003). Risk assessment: Discussion of the section. In R. A. Prentky, E.S. Janus, & M. C. Seto (Eds.) Understanding and managing sexually coercive behavior. (pp. 236-245). New York: Annals of the New York Academy of Sciences, Vol. 989.

Hanson, R.K. (1997). The development of a brief actuarial risk scale for sexual offense recidivism. Department of the Solicitor General of Canada, Ottawa, Ontario. Available at http://www.psepc-sppcc.gc.ca/publications/corrections/199704_e.pdf .

Hanson, R.K. & Bussière, M.T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. Journal of Consulting and Clinical Psychology, 66(2), 348-362.

Hanson, R. K., & Harris, A. J. R. (2001). A structured approach to evaluating change among sexual offenders. Sexual Abuse: A Journal of Research and Treatment, 13 (2), 105-122.

Hanson, R.K., Morton, K.E., & Harris, A.J.R. (2003). Sexual offender recidivism risk: What we know and what we need to know. In R. A. Prentky, E.S. Janus, & M. C. Seto (Eds.) Understanding and managing sexually coercive behavior. (pp. 154-166). New York: Annals of the New York Academy of Sciences, Vol. 989.

Hanson, R.K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. Law and Human Behavior, 24(1), 119-136.

Hare, R.D. (1991). The Hare Psychopathy Checklist - Revised. Toronto: Multi-Health Systems.

Harris, G. (2003). Men in his category have a 50% likelihood, but which half is he in? Comments on Berlin, Galbreath, Geary, and McClone. Sexual Abuse: A journal of Research and Treatment, 15(4), 389-392.

Hart, S.D. (2003). Actuarial risk assessment: Commentary on Berlin et al. Sexual Abuse: A Journal of Research and Treatment, 15 (4), 383-388.

Janus, E.S., & Prentky, R.A. (2003). Forensic use of actuarial risk assessment with sex offenders: Accuracy, admissibility and accountability. American Criminal Law Review, 40(4), 1443-1499.

Janus, E.S., & Prentky, R.A. (2004). Forensic use of actuarial risk assessment: How a developing science can enhance accuracy and accountability. Sex Offender Law Report, 5(5), 55-56 & 62-63.

Lalumière, M.L., Quinsey, V.L., Harris, G.T., Rice, M.E., & Trautrimas, C. (2003). Are rapists differentially aroused by coerecive sex in phallometric assessments? In R. A. Prentky, E.S. Janus, & M. C. Seto (Eds.) Understanding and managing sexually coercive behavior. (pp. 211-224). New York: Annals of the New York Academy of Sciences, Vol. 989.

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

Langton, C.M. (2003). Contrasting approaches to risk assessment with adult male sexual offenders: An evaluation of recidivism prediction schemes and the utility of supplementary clinical information for enhancing predictive accuracy. Unpublished doctoral thesis, University of Toronto, Toronto, Canada.

Lloyd, M.D., & Grove, W.M. (2002). The uselessness of the Minnesota Sex Offender Screening Tool-Revised (MnSOST-R) in commitment decisions. Submitted for publication.

Marshall, W.L., & Barbaree, H.E. (1995). Heterogeneity in the erectile response patterns of rapists and nonoffenders. Unpublished manuscript, Queen's University, Kingston, Ontario.

Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Moreland, K.L., Dies, R.R., Eisman, E.J., Kubiszyn, T.W., & Reed, G.M. (2001). Psychological testing and psychological assessment. American Psychologist, 56(2), 128-165.

Nuffield, J. (1982). Parole decision making in Canada: Research towards decision guidelines. Ottawa, Ontario: Solicitor General of Canada.

Otto, R.K., & Petrila, J. (2002). Admissibility of testimony based on actuarial scales in sex offender commitments: A reply to Doren. Sex Offender Law Report, 3(1), 1 & 14-16.

Ozer, D.J. (1985). Correlation and the coefficient of determination. Psychological Bulletin, 97, 307-315.

Quinsey, V.L., Harris, G.T., Rice, M.E., & Cormier, C.A. (1998). Violent offenders: Appraising and managing risk. Washington, D.C.: American Psychological Association.

Steering Committee of the Physicians' Heath Study Research Group (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. New England Journal of Medicine, 318, 262-264.

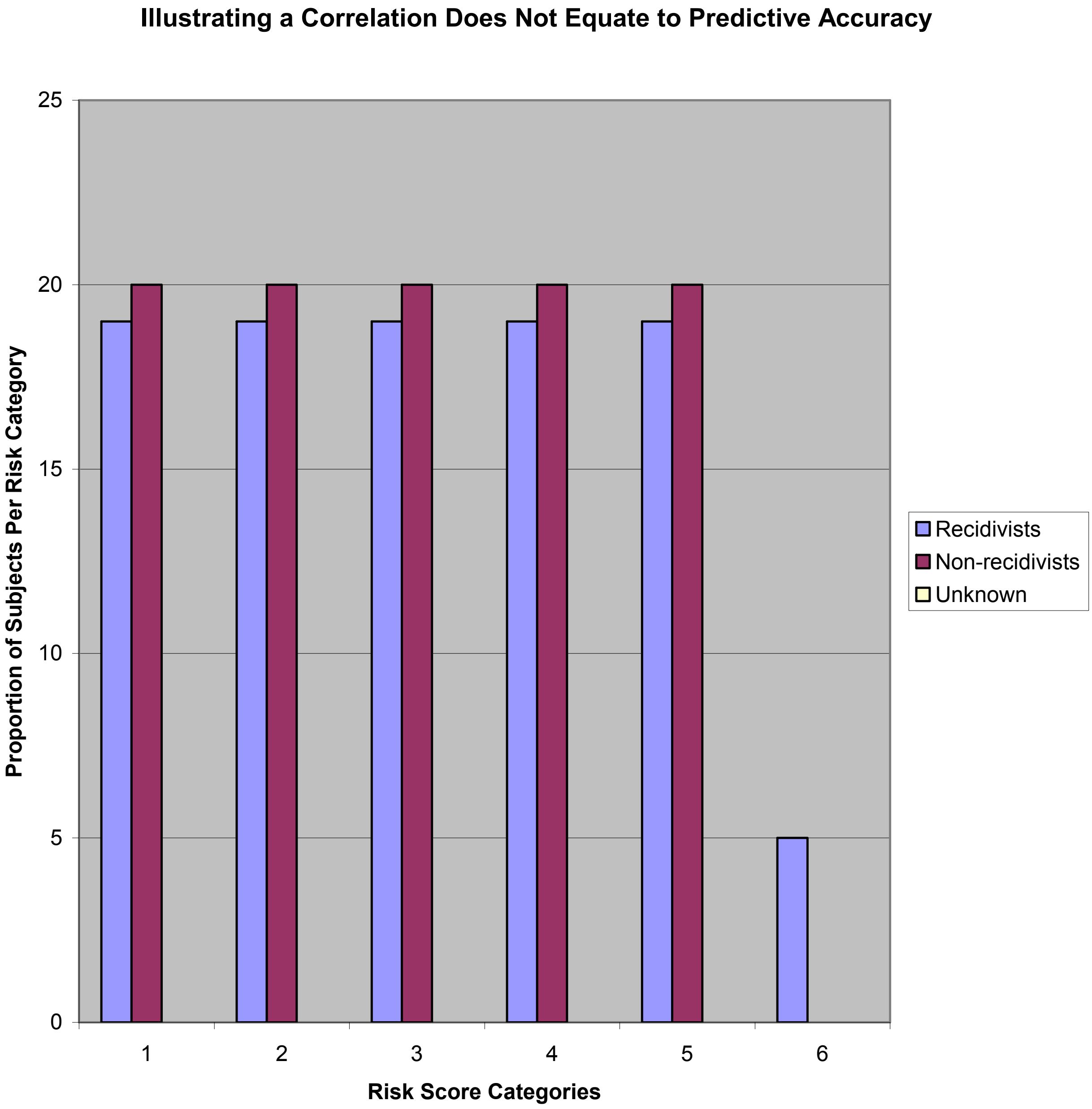
Vogel, V. de, Ruiter, C. de, Beek, D. van, & Mead, G.(2004). Predictive validity of the SVR-20 and the Static-99 in a Dutch sample of treated sex offenders. Law and Human Behavior, 28(3), 235-251.

Webster, C.D., Harris, G.T., Rice, M.E., Cormier, C., & Quinsey, V.L. (1994). The violence prediction scheme: Assessing dangerousness in high risk men. Toronto: University of Toronto, Centre of Criminology.

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

Wollert, R.W. (2003). Additional Flaws in the Minnesota Sex Offender Screening Tool - Revised: A Response To Doren and Dow (2002). Journal of Threat Assessment, 2(4), 65-78.

Figure 1


Illustrating a Correlation Does Not Equate to Predictive Accuracy
25
20
15
10
5
0
Proportion of Subjects Per Risk Category
1
2
3
4
5
6
Risk Score Categories
Recidivists
Non-recidivists
Unknown