40 Am. Crim. L. Rev. 1443 (2003).

FORENSIC USE OF ACTUARIAL RISK ASSESSMENT WITH SEX OFFENDERS: ACCURACY, ADMISSIBILITY AND ACCOUNTABILITY

Eric S. Janus, J.D.[*]
Robert A. Prentky, Ph.D.[**]

Introduction

The decade of the nineties ushered in an unprecedented number of state and federal laws intended to manage sexual offenders.[1] Among the most controversial are the so-called "sexually violent person" (hereinafter "SVP") laws[2]—schemes that use civil commitment to supplement criminal sentences in order to incapacitate the "most dangerous" sex offenders. Risk assessment—the prediction of sexual recidivism—is essential to this legislative agenda. As a result, the demand for specialized risk assessments has been rapidly growing, and has produced a "cottage industry of forensic psychologists"[3] and vigorous development of actuarial and other structured approaches to supplement the traditional clinical assessment of this risk.[4]

[*] Professor of Law, William Mitchell College of Law, St. Paul, MN. Professor Janus has served as co-counsel in the *Linehan* litigation, challenging the constitutionality of Minnesota's sex offender commitment laws. He wishes to thank his colleagues at William Mitchell College of Law—particularly Professors Eileen Scallen, Peter Knapp, and Wayne Logan—for their insightful assistance on this paper, and his research assistants Jamie Polovitz and Conor Tobin.

[**] Director of Research, Justice Resource Institute, Bridgewater, MA.

[1] *See, e.g.*, Jacob Wetterling Crimes Against Children and Sexually Violent Offender Registration Program, 42 U.S.C. § 14071 (2003) (conditioning receipt of federal funds on establishment of rigorous registration requirements for those sexual offenders deemed to be very dangerous); Megan's Law, 42 U.S.C. § 13701 (2001) (mandating that the designated agency for each state "shall release relevant information that is necessary to protect the public concerning a specific person required to register"); Child Online Protection Act, 47 U.S.C. § 231 (1998) (making it illegal for commercial Web site operators to post "material that is harmful to minors" without blocking access to the site through a credit card requirement or other adult verification); American Civil Liberties Union v. Reno, 217 F.3d 162, 181 (3d Cir. 2000) (finding that 47 U.S.C. § 231 violates the First Amendment). *See generally* Virginia B. Baldau, U.S. Dep't of Justice, Summary of State Sex Offender Registries: Automation and Operation, 1998, NCJ177621 (1999), *available at* http://www.ojp.usdoj.gov/bjs/pub/pdf/sssorao.pdf (last visited Nov. 3, 2003) (noting that as of April 1998, approximately 276,000 sex offenders were registered under the respective laws); Scott Matson & Roxanne Lieb, Wash. St. Inst. Pub. Pol'y, Sex Offender Registration: A Review of State Laws (1996), *available at* http://www.wsipp.wa.gov/crime/pdf/regsrtn.pdf (last visited Nov. 3, 2003) (describing the nature of state sexual offender registration laws and the number of offenders on the registries).

[2] *See infra* note 13 and accompanying text.

[3] Thomas Grisso, *The Economic and Scientific Future of Forensic Psychological Assessment,* 42 Am. Psychologist 831 (1987) (quoting United States v. Downing, 753 F.2d 1224, 1232 (3d Cir. 1985)).

[4] *See infra* Part III.

These modern legislative initiatives for the management of sexual offenders have generated heated controversy, and their reliance on the assessment of risk has, in important ways, been central to the controversy. An initial wave of litigation in the mid-1990's questioned the constitutional foundations of this new legislative agenda. By calling attention to the inherently problematic task of assessing future risk, these challenges questioned the constitutionality of locking up people for future crimes that they might—or might not—commit. The courts were unimpressed, and confirmed the basic constitutionality of risk-based deprivations of liberty, despite well-known shortcomings in the prediction of dangerousness.[5]

A second wave of litigation is now focused on a more detailed articulation of the standards for assessing risk in these legal settings. Courts are asking in more detail what the legal standards for dangerousness mean, and what sort of evidence is legally available, and ought to be required, to prove those standards. In large measure, the legal challenges in these more recent cases have addressed the use of actuarially-derived (as distinct from "clinical") risk assessments, and it is this subject that we examine in this article. Clinical judgments of dangerousness —judgments that ultimately rest on the *ipse dixit* of a mental health professional —are a routine part of the judicial landscape. In contrast, actuarial risk assessment (hereinafter "ARA")—which employs empirically derived "mechanical" rules for combining information to produce a quantitative estimate of risk—is novel in the legal arena and seems to be setting off a variety of alarm bells. Critics of ARA have focused their objections on the admissibility of ARA-derived expert testimony.[6] Pointing to a variety of shortcomings, they argue that the relatively new ARA techniques do not merit admissibility under prevailing *Frye* or *Daubert* standards.[7] Such challenges have met with mixed success.

In this paper we explore the forensic use of actuarial risk assessment testimony, particularly in the context of SVP laws. Our thesis is straightforward: actuarial methods have proven equal or superior to clinical judgments.[8] Given the

[5] *See* Kansas v. Hendricks, 521 U.S. 346, 358 (1997) (quoting Schall v. Martin, 467 U.S. 253, 278 (1984) for the proposition that "from a legal point of view there is nothing inherently unattainable about a prediction of future criminal conduct"); State v. Post, 541 N.W.2d 115, 132 (Wis. 1995) (rejecting challenge based on impossibility of prediction); *In re* Blodgett, 510 N.W.2d 910, 917 n.15 (Minn. 1994) (granting broad deference to the opinions of mental health experts regarding predictions of future dangerousness).

[6] *See*, *e.g.*, Randy K. Otto & John Petrila, *Admissibility of Testimony Based on Actuarial Scales in Sex Offender Commitments: A Reply to Doren*, 3 SEX OFFENDER LAW REPORT 1 (2002); Terence W. Campbell, *Sexual Predator Evaluations and Phrenology: Considering Issues of Evidentiary Reliability,* 18 BEH. SCI. & L. 111, 128 (2000) (arguing that two particular actuarial instruments are not sufficiently reliable to support expert testimony).

[7] *See* Frye v. United States, 293 F. 1013 (D.C. Cir. 1923) (designating general acceptance by the scientific community as the standard for admissibility of expert testimony); Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579 (1993) (holding Federal Rule of Evidence 702 supercedes the Frye standard and discussing a factor test for determining admissibility of expert evidence).

[8] *See*, *e.g.*, William M. Grove et al., *Clinical Versus Mechanical Prediction: A Meta-Analysis*, 12 PSYCHOL. ASSESSMENT 19, 19 (2000) (finding that actuarial prediction techniques were, on average, ten percent more accurate than clinical predictions)*;* Howard E. Barbaree et al.,

legislative mandate to assess risk, as well as the routine, widespread use of clinical assessments of risk in the judicial system, it is logically incoherent to exclude evidence that presumptively improves upon the reliability and accuracy of these judgments.[9]

But the issues raised are larger than those of admissibility. Like any tool of science, ARA can work for good or ill. Used well, ARA can ameliorate some of the concerns about justice, efficacy, and public policy that swirl around SVP laws and the other recent legislative initiatives. In fact, ARA is a state-of-the-art technique, and courts should insist that it be employed as a major instrument of risk assessment. Used poorly, however, ARA can exacerbate concerns about justice and utility. Further, the use of ARA might have unanticipated and undesired consequences for broader areas of public policy. Improved ability to identify persons at high risk for violence may make expanded preventive detention laws politically impossible to resist. New laws, in turn, may demand better risk assessment, which may beget more aggressive and expansive prevention laws, and so on. In view of these negatives, we urge caution and mindfulness in using ARA.

Our argument for the admissibility of ARA must be understood in the context of two important preliminary points. First, as we will more fully discuss later, the development of ARA, like all good science, is evolutionary. The sophistication of ARA evolves over time as more is learned about the task of assessing sex offender risk and the functioning of particular ARA scales. ARA scales differ in their reliability and accuracy. Hence, they should not be considered equivalent and thus interchangeable. Like all products of science, they are works in various stages of development.

Second, it is imperative that ARA be used properly. This second issue raises several legitimate concerns; among the most critical is the proper interpretation of ARA information. To the extent that courts seek to measure the long-term, presumptively stable risk posed by individuals, ARA provides the most accurate information. But courts ought to be concerned as well with how risk can be managed and modified in the short- and medium-terms, through intervention such as treatment and community supervision.[10] This domain, generally referred to as "dynamic" risk assessment, represents the most recent entrée to the scientific literature and will likely be the focus of attention among scientists for the foreseeable future. Given its focus on long-term risk, however, ARA is of less

Evaluating the Predictive Accuracy of Six Risk Assessment Instruments for Adult Sex Offenders, 28 CRIM. JUST. & BEHAV. 490, 492 (2000) (discussing the accuracy of the Violence Risk Appraisal Guide at predicting recidivism by sex offenders).

[9] *See generally* JOHN MONAHAN ET AL., RETHINKING RISK ASSESSMENT: THE MACARTHUR STUDY OF MENTAL DISORDER AND VIOLENCE 7 (2001) (discussing the well-known superiority of actuarial predictive methods to clinical predictive methods).

[10] *See*, *e.g.*, Gabrielle Sjöstedt & Martin Grann, *Risk Assessment: What is Being Predicted by Actuarial Prediction Instruments?*, 1 INT'L J. FORENSIC MENTAL HEALTH 179, 180 (2002) (distinguishing between prediction of violence and management of risk); ROBERT A. PRENTKY & ANN W. BURGESS, FORENSIC MANAGEMENT OF SEXUAL OFFENDERS 236, 241 (2000) (describing the "management model," which aims to reduce risk and accordingly requires instruments that are sensitive to changes in risk status).

direct relevance, at least given the current state of the art.[11] Therefore, on these important questions of risk management and modification, courts may, for the time being, need to rely more heavily on carefully done clinical assessments, though it is likely that dynamic ARA will eventually complement these assessments as well.

In urging the use of ARA, we do so against the backdrop of existing SVP laws and serious concerns raised about their constitutionality and wisdom. Although our paper is premised on findings that ARA is superior to clinical assessment methods, we do not take the position that either is sufficient to justify the massive and long-term deprivation of liberty inherent in SVP laws. Our point is simply that *if* courts deprive people of liberty based on assessed risk, *then* ARA should be part of the assessment. Courts should use ARA in part because it will improve risk assessment. But more importantly, from our perspective, ARA brings a transparency that will allow for a clearer understanding of the true nature of risk assessment, including its significant limits and potential for misuse.

This paper proceeds as follows. We first set the context by briefly describing SVP laws, highlighting two salient features: their lack of clear standards for confinement, and their extraordinary cost. Second, we discuss notions of risk assessment and predictions of dangerousness, and outline the key legal concepts that control their use. Third, we introduce the distinctions between clinical and actuarial methods of risk assessment, and summarize the empirical basis for the claim that ARA is generally superior to clinical assessment of risk. Fourth, we discuss the treatment of ARA under prevailing standards for admissibility, the *Frye* and *Daubert* tests, arguing that SVP courts should admit ARA testimony. Finally, we propose a set of guidelines for courts to use to minimize the potential for misuse and prejudice, and maximize the beneficial consequences, in connection with ARA testimony. In closing, we argue that performing risk assessments without ARA is unethical for mental health professionals and improper for courts. But we warn that increasingly accurate methods of risk assessment may encourage the expansion of civil-commitment-style "violent person" laws, an approach to prevention that, in our view, is questionable both morally and practically.

I. SEXUALLY DANGEROUS PERSON COMMITMENT LAWS: MASSIVE DEPRIVATION OF LIBERTY, HUGE COSTS, AND FEW STANDARDS

Sexually violent person laws adopt the framework of mental illness civil commitment to address recidivist sexual violence.[12] Aimed primarily at convicted

[11] On the subject of dynamic risk assessment, see, e.g., R. Karl Hanson & Andrew J.R. Harris, *A Structured Approach to Evaluating Change Among Sexual Offenders*, 13 SEXUAL ABUSE: A JOURNAL OF RESEARCH AND TREATMENT 105 (2001); David Thornton, *Constructing and Testing a Framework for Dynamic Risk Assessment*, 14 SEXUAL ABUSE: A JOURNAL OF RESEARCH AND TREATMENT 139 (2002).

[12] *See*, *e.g.*, Eric S. Janus, *Sexual Predator Commitment Laws: Lessons for Law and the Behavioral Sciences*, 18 BEHAV. SCI. & L. 5, 5 (2000) (discussing Kansas v. Hendricks, 521 U.S. 346 (1997)).

sex offenders who are completing their prison sentences, SVP laws authorize long-term confinement in secure treatment centers for individuals who have a “mental disorder or abnormality” or “personality disorder” that produces a risk of future criminal sexual misconduct. Confinement continues until the individual can demonstrate that he no longer meets these criteria. Over the 13 years that these laws have been in place, only a small fraction of the committed individuals have met this release burden.[13] Thus, the population under commitment is growing. As of 2002, more than 1,600 individuals were committed and 840 were awaiting commitment proceedings.[14] As of September 2001, only 61 (roughly 5% of those committed at that time) had been cleared (found no longer “sexually dangerous”).[15] Sexually Violent Person programs are exceedingly expensive,[16] and growing populations of committed sex offenders will require an increasingly disproportionate share of treatment and prevention dollars.[17]

SVP laws are highly controversial for two reasons. First, they are morally and constitutionally suspect: by locking people up for long periods of time in order to prevent future crimes, SVP laws challenge deeply held notions of justice and ethics.[18] Second, SVP programs divert scarce funds not only from other mental health populations, but also from other, potentially more effective, sexual violence prevention strategies.[19]

[13] *See* W. Lawrence Fitch, *Sexual Offender Commitment in the United States: Legislative and Policy Concerns*, *in* Sexually Coercive Behavior: Understanding and Management 489, 492 (Robert Prentky, et al., eds., 2003) (noting that, out of a nation-wide total of 2,478 SVP’s in confinement, only 82 had been released).

[14] *Id.*

[15] *See* Robert A. Prentky, *A 15-year Retrospective on Sexual Coercion Research and Developments*, *in* Sexually Coercive Behavior: Understanding and Management 23 (Robert Prentky, et al., eds., 2003).

[16] *See* John Q. La Fond, *The Costs of Enacting a Sexual Predator Law*, 4 Psychol. Pub. Pol’y & L. 468, 478 (1998) (estimating that, in Washington in 1998, the cost per resident of one year of commitment was approximately $91,969); Eric S. Janus, *Minnesota’s Sex Offender Commitment Program: Would an Empirically-Based Prevention Policy Be More Effective?,* 29 Wm. Mitchell L. Rev. 1083, 1101 (2003) (noting that commitments in Minnesota cost about $20 million per annum, and will increase more than three-fold by 2010).

[17] *See* Janus, *supra* note 16, at 1090.

[18] *See, e.g*., Robert F. Schopp, *Sexual Aggression: Mad, Bad and Mad, in* Sexually Coercive Behavior: Understanding and Management 324, 335 (Robert Prentky, et al., eds., 2003) (arguing the blending of civil commitment and criminal sanctions distorts principle of retributive justice).

[19] *See, e.g.*, John Q. La Fond & Bruce J. Winick, *A Therapeutic Jurisprudence Approach to Managing Sex Offender Risk: A Proposal for Sex Offender Reentry Courts, in* Sexually Coercive Behavior: Understanding and Management 300, 309 (Robert Prentky, et al., eds., 2003) (arguing risk management approach to sexual offenders would be less costly and would serve a greater number of offenders); Janus, *supra* note 16, at 1109 (“[T]o the extent that the state spends *extraordinary* resources on the *ordinarily* risky, the resource misallocation . . . is exacerbated.”) (emphasis in original).

Risk assessment figures centrally in both of these concerns. Dangerousness is one of two constitutionally required components of civil commitment;[20] it is the unambiguous justification for the civil commitment of sex offenders (i.e., we are protecting society from the "most dangerous" perpetrators).[21] Although preventive detention would be legally and ethically problematic even *with* perfect knowledge about the future, the imperfection of risk assessment exacerbates constitutional and ethical concerns because it raises the likelihood that non-recidivists and low-risk individuals will be among the group suffering long-term loss of liberty. The same is true for the more utilitarian concerns about resource allocation and efficacy. The central justification for spending huge sums of money on SVP programs is that the "most dangerous" offenders are incapacitated. Public policy is not well served if, because of inaccurate assessment of risk, extraordinary resources are devoted to the ordinarily dangerous.[22]

II. RISK ASSESSMENT AND PREDICTIONS OF DANGEROUSNESS: BASIC CONCEPTS

The shortcomings in risk assessment can be traced to two sources. First, humans (experts or otherwise) have a limited ability to assess future risk of harmful behavior. These limits stem in part from the fact that the future is inherently unknowable, and in part from inherent shortcomings in human judgment.[23] Although the former limitation is inescapable, the limits of human judgment can, to some extent, be ameliorated through empirical research. Thus, as we show below, the quality of risk assessment is variable, and improvement is possible.

The second source of shortcomings in risk assessment resides in the legal system. The risk thresholds for invoking SVP commitments are vague, and courts have failed to set standards that are reviewable and enforceable, relying instead on unoperationalized[24] terms, such as "likely."[25] Frequently, the liberty-deprivation

[20] *See* Foucha v. Louisiana, 504 U.S. 71, 77-78 (1987) (noting civil commitment proceedings require a determination of current mental illness *and* dangerousness). *See also* Eric S. Janus & Paul E. Meehl, *Assessing the Legal Standard for Predictions of Dangerousness in Sex Offender Commitment Proceedings*, 3 PSYCHOL. PUB. POL'Y & L. 33, 37 (1997) (noting that for civil commitment, the state must have a compelling interest in preventing harm).

[21] *See* Kansas v. Hendricks, 521 U.S. 346, 364 (1997) (noting, with respect to the state's SVP law, that "the Kansas legislature has taken great care to confine only a narrow class of particularly dangerous individuals"); State v. Post, 541 N.W.2d 115, 124 (Wis. 1995), *cert. denied*, 521 U.S. 1118 (1997) (holding Wisconsin's SVP statute "is narrowly tailored to allow commitment only of the most dangerous of sexual offenders").

[22] Janus, *supra* note 16, at 1109.

[23] *See*, *e.g.*, David Faust, *Data Integration in Legal Evaluations: Can Clinicians Deliver on Their Premises?*, 7 BEHAV. SCI. & L. 469, 471 (1989) (stating research "raises doubt about the capacity of individuals or clinicians to manage and grasp complex information or configural relations").

[24] We use the word "unoperationalized" to indicate the relative absence of clear guidelines and criteria that would ensure that some particular term is applied in a consistent way, such that all who use the term have a common understanding of what it means and apply it in the same way.

decision boils down to a credibility judgment between the clinical assessments of two competing expert witnesses.[26] As a result, there is no assurance that risk thresholds are uniform or that risk assessments are performed at the highest standards. Thus, the legal system fails to take advantage of the scientific virtues of risk assessment, and exacerbates the failings of behavioral science by inviting arbitrariness to join the mix.

In this article, we argue that the use of ARA can address both sources of shortcomings. As we will explain, ARA represents the best that behavioral science has to offer in risk assessment. ARA also provides transparency to risk assessment, and thus allows courts to set and enforce clear standards in this area.

In accordance with increasingly common practice among both clinicians and forensic examiners,[27] we use the concept of risk rather than dangerousness. Risk has greater utility and more flexibility than dangerousness for several reasons. First, risk addresses not only the presence of a potential hazard, but the probability of its occurrence.[28] It is thus dimensional and continuous, whereas dangerousness is too often thought of in dichotomous terms (i.e., either the offender *is* dangerous or *is not* dangerous).[29] Second, dangerousness connotes a narrow but not precisely defined swath of human behavior, typically consisting of acts of interpersonal violence;[30] risk, by contrast, captures a much broader range of behaviors (e.g., an offender's risk of eloping, violating parole, drinking, using drugs, developing depression, committing suicide, etc.). Third, while discussions of dangerousness often conflate several distinct concepts (e.g., the probability and

[25] *See* Commonwealth v. Boucher, 780 N.E.2d 47, 49-50 (Mass. 2002). The *Boucher* court rejected the trial court's holding that "likely" meant "more likely than not," and concluded instead that:

> In assessing the risk of reoffending, it is for the fact finder to determine what is 'likely.' Such a determination must be made on a case-by-case basis, by analyzing a number of factors, including the seriousness of the threatened harm, the relative certainty of the anticipated harm, and the possibility of successful intervention to prevent that harm.

Id. See also People v. Ghilotti, 44 P.3d 949, 954, 972 (Cal. 2002) (holding that "likely" means there is "a serious and well-founded risk, that he or she will commit such crimes if free in the community," and rejecting the argument that "likely" means "better than even chance of new criminal sexual violence"); *In re* R.S., 773 A.2d 72, 75 (N.J. Super. Ct. App. Div. 2001) (affirming trial court's application of New Jersey statutory provision requiring assessment of whether person is "likely to engage in acts of sexual violence").

[26] *See* Eric S. Janus, *The Use of Social Science and Medicine in Sex Offender Commitment*, 23 NEW ENG. J. ON CRIM & CIV. CONFINEMENT 347, 369 (1997) (noting the conflicting results reached by two courts assigning different weights to a "disorder" expert and a "traits" expert).

[27] *See* PRENTKY & BURGESS, *supra* note 10, at 100 (discussing the use of risk in place of dangerousness).

[28] *Id.*

[29] *Id.*

[30] *See*, *e.g.*, MONAHAN ET AL., *supra* note 9, at 3, 17 (suggesting interpersonal violence is the referent underlying statutory "dangerousness" assessment).

magnitude of harm),[31] discussions of risk demand clarity about the specific type of behavior in question. Finally, the use of risk brings criminology in line with numerous other disciplines, from health care (measuring health care outcomes, where the term "risk adjustment" is used) and environmental protection (environmental health risk management) to meteorology.[32]

In his 1974 book, Professor Brooks enumerated four components of dangerousness: "(1) the magnitude of harm, (2) the probability that the harm will occur, (3) the frequency with which the harm will occur, and (4) the imminence of the harm."[33] Sex offender commitment statutes often target offenders who have committed repeated acts of sexual violence; such statutes address the first and third of Professor Brooks' components, which have earned a fair amount of judicial attention.[34]

The second of Brooks' dangerousness components—the probability that harm will occur—is the most contentious, and often the major focus of concern in SVP litigation.[35] The assertion that only the most dangerous sex offenders are civilly committed is based on four assumptions concerning this component: "(a) the probability of dangerousness is susceptible of measure, (b) there is a way to discriminate between predictions of higher and lower probability, (c) there are standards that allow commitments based on the former while excluding confinement based on the latter, and (d) these standards are, in fact, enforced."[36] The first two of these assumptions address the scientific aspects of risk assessment, while the last two address the legal dimensions.

An additional distinction will be helpful. Over twenty years ago, Professors John Monahan and David Wexler sought to clarify the probability standards used in civil commitments.[37] Monahan and Wexler proposed a bifurcated standard: a standard of commitment and a standard of proof. The standard of commitment sets the substantive threshold for civil commitment. It refers to a characteristic of the defendant and is described in terms of the likelihood of dangerous behavior. At least at a superficial level, this standard of

[31] *See* Marie A. Bochnewich, Comment, *Prediction of Dangerousness and Washington's Sexually Violent Predator Statute*, 29 CAL. W.L. REV. 277, 298 (1992) (citing ALEXANDER D. BROOKS, LAW, PSYCHIATRY AND THE MENTAL HEALTH SYSTEM (1974)).

[32] *See* PRENTKY & BURGESS, *supra* note 10, at 101.

[33] Bochnewich, *supra* note 31, at 298.

[34] *See*, *e.g.*, *In re* Robb, 622 N.W.2d 564, 573 (Minn. Ct. App. 2001) (holding that statutory requirement of "harmful sexual conduct" does not necessitate violence); *In re* Hince, No. C9-94-1366, 1994 WL 637755, at *1 (Minn. Ct. App. Nov. 15, 1994) (finding two instances of sexual violence is not sufficient to show "habitual" pattern).

[35] *See* cases cited *supra* note 25; Eric S. Janus, *Legislative Responses to Sexual Violence: An Overview*, *in* SEXUALLY COERCIVE BEHAVIOR: UNDERSTANDING AND MANAGEMENT 247, 253 (Robert Prentky, et al., eds., 2003) (canvassing judicial interpretations of the probability element of proof).

[36] Janus & Meehl, *supra* note 20, at 38.

[37] *See generally* John Monahan & David Wexler, *A Definite Maybe: Proof and Probability in Civil Commitment*, 2 LAW & HUM. BEHAV. 37 (1978).

commitment refers to a "fact" in the real world, and is typically associated with such terms as "likely" and "highly likely."[38] The standard of proof, in contrast, measures the degree of certainty or conviction that the trier of fact must have in order to find that a particular fact is true.[39] The standard of proof is operationalized in the courtroom with such phrases as "clear and convincing evidence," and "beyond a reasonable doubt."[40]

As Monahan and Wexler submitted, and as Janus and Meehl discussed, the relationship between these two standards is "intricate and complex," and, at times, intertwined.[41] A full discussion of this complex relationship is beyond the scope of this paper. It suffices here to point out that each standard invokes a notion of probability, but there is no clear line between the two probabilities invoked. For example, when a mental health expert opines that the defendant is "likely" to commit a crime, is she talking about a "fact" in the real world (i.e., a characteristic or quality of the defendant), or about the level of certainty of her knowledge of the real world (e.g., that she has moderate confidence that the defendant will commit a crime)? In some ways, the use of ARA testimony demystifies these questions. As we shall describe, ARA testimony involves facts in the real world—facts concerning the measured frequency of sexual recidivism among individuals with described characteristics—that clearly address the standard of commitment. The standard of proof, in contrast, will address how certain the trier is that these facts are true, and, more to the point, that they support the legal elements for commitment.

This discussion of the standards of commitment and proof leads logically to a final preliminary matter. Risk assessment testimony is evaluated at three distinct stages of the litigation process. First, judges make a threshold decision of admissibility. Key considerations are relevance or "fit," prejudice, and reliability.[42] The second stage, which only comes into play if evidence is

[38] *See*, *e.g.*, Commonwealth v. Boucher, 780 N.E.2d 47, 50 (Mass. 2002) ("As used in the statute, however, the term 'likely' is not intended as a standard or burden of proof. Rather, it is descriptive of one characteristic ('likely to engage in sexual offenses') of a sexually dangerous person.").

[39] *See*, *e.g.*, *In re* W.Z., 773 A.2d 97, 114 (N.J. Super. Ct. App. Div. 2001) ("The burden of proof, which denotes a degree of certainty of conviction, is a consideration separate and apart from the likelihood of reoffense, which denotes a factual probability.").

[40] *See* Monahan & Wexler, *supra* note 37, at 41 (discussing complex relationship between standards of proof and standards of commitment and noting one court's adoption of the "reasonable doubt" standard for both); *see also Boucher*, 780 N.E.2d at 50 (distinguishing the "more likely than not" standard of proof from the descriptive term, "likely to engage in sexual offenses," and noting that the descriptive term does not require any particular mathematical quantum of proof).

[41] *See* Monahan & Wexler, *supra* note 37, at 41; *see also* Janus & Meehl, *supra* note 20, at 42-43 (arguing the two standards are intertwined, but characterizing the "standard of proof as a standard for measuring epistemological uncertainty, and the standard of commitment as a standard for measuring ontological uncertainty).

[42] Courts and legal scholars use various terms when referring to the "integrity" of testimony. Terms such as reliability, validity, dependability, and trustworthiness are frequently used. *See generally* David L. Faigman et al., *How Good is Good Enough?: Expert Evidence Under Daubert and Kumho*, 50 CASE W. RES. L. REV. 645 (2000) (discussing standards of reliability for

admitted, is legal sufficiency. This is a determination made by the judge as to whether the evidence, if fully credited by the jury or other fact finder, satisfies the legal standard for commitment.[43] The third stage is the assessment of the weight of the testimony. This is a function of the jury or other finder of fact. The questions here concern whether, and to what extent, the witness's testimony is credible, and how much influence this testimony ought to have in deciding the ultimate question (here, the legally defined risk posed by the defendant).[44] Put another way, the finder of fact applies the "standard of proof" in judging the weight of the evidence.

In the past, the first two stages have played a minimal role in the legal evaluation of risk-finding processes. As we discuss below, courts have generally given clinical risk assessment (CRA) testimony a green light at the admissibility stage, with little or no scrutiny.[45] As for the sufficiency judgment, courts have ruled that risk assessment testimony is constitutionally sufficient as a general matter, but have failed to characterize their decisions about individual risk assessments as judgments of legal sufficiency.[46] Rather, courts have uniformly relegated evaluation of the risk-finding process to the third stage, the assessment of "weight." [47]

This shifting of the location for the evaluation of the risk-finding process is significant, because it represents an abdication by the courts of their role as standards-setters. Weight-of-the-evidence decisions are generally viewed as unreviewable findings of fact.[48] Courts are thus relieved of the burden of

admissibility of expert evidence). In this article, we have tried to be consistent in our use of the word "reliability" when referring to the legal test for judging expert testimony.

We also use the terms "reliability" and "validity" in their technical sense when referring to the psychometric properties of ARA instruments. See *infra* Part V.B.1. In that sense, "reliability" refers to whether a test will yield the same results when applied repeatedly to the same individual. The more reliable the test, the less error it has arising from variations in the measurement process. But a test can be very reliable, in the technical sense, but not "accurate" if it lacks "validity," a complex term that generally means that the test measures what it claims to measure. *See generally* JOY PAUL GUILFORD, PSYCHOMETRIC METHODS (1954).

[43] *See*, *e.g.*, *In re* Coffel, No. ED 79989, 2003 WL 716682, at *13 (Mo. Ct. App. March 4, 2003) (reversing SVP commitment on the ground that the state's evidence on "likelihood" of reoffense was insufficient).

[44] *See*, *e.g.*, *In re* Joelson, 385 N.W.2d 810, 811 (Minn. 1986) (holding when findings rest almost entirely on expert opinion testimony, court's evaluation of credibility is of particular significance); *In re* Curiel, 597 N.W.2d 697, 711 (Wis. 1999) (noting the weight given to the testimony of an expert witness is always an issue properly before the trier of fact).

[45] *See infra* note 91.

[46] *See*, *e.g.*, State v. Post, 541 N.W.2d 115, 126 (Wis. 1995), *cert. denied*, 521 U.S. 1118 (1997) (explaining predictions of dangerousness are "difficult" but "attainable," while noting that "[t]he Supreme Court has refused to proscribe strict boundaries for legislative determinations of what degree of dangerousness is necessary for involuntary commitment.")

[47] *See*, *e.g.*, cases cited *supra* note 44.

[48] *See*, *e.g.*, Westerheide v. State, 767 So. 2d 637, 658-59 (Fla. Dist. Ct. App. 2000) (skipping directly from the admissibility of expert testimony to the weight of such testimony, and ignoring entirely the court's role in judging legal sufficiency).

articulating legal standards for dangerousness and explaining, based on those standards, their judgments about risk. Risk-finding decisions become opaque applications of vague catch phrases such as "likely" or "highly likely," and subjective applications of phrases like "beyond a reasonable doubt."[49]

A central thesis of this paper is that ARA can provide the necessary transparency to allow courts to evaluate the sufficiency of risk assessment testimony. A prime example is *Cooley v. Superior Court*,[50] a case concerning a probable cause hearing under California's SVP law. This is one of the few cases in which the trial court dismissed an SVP petition for insufficiency of risk.[51] The state's experts relied, in part, on the Static-99, an actuarial risk assessment instrument.[52] The defendant's witnesses attacked the adequacy of the Static-99, and the way in which the state's experts had used this actuarial information.[53] The trial court found that the state's experts had been effectively impeached, and held that the state had failed to establish probable cause on the issue of risk.[54] The California Supreme Court approved of the trial court's method, but ultimately remanded because the trial court had used an improper standard for measuring risk.[55] The key point is that the trial court did not exclude the ARA evidence, but rather used the transparency of ARA to evaluate the adequacy of the risk assessment testimony.[56] This type of evaluation would not have been possible if the only basis for the testimony had been clinical judgment.

III. CLINICAL VS. ACTUARIAL RISK ASSESSMENT

Central to our discussion is the distinction between clinical and actuarial methods of risk assessment. In this section, we describe the difference between the two methods, and summarize the scientific findings that compare the two methods.

The literature generally discusses two methods of risk assessment, clinical and actuarial. One study contrasts the two methods as follows:

> In the clinical method the decision-maker combines or processes information in his or her head. In the actuarial or statistical method the human judge is eliminated and conclusions rest solely on

[49] *See*, *e.g.*, *Commonwealth v. Boucher*, 780 N.E.2d 47, 49 (Mass. 2002) (adopting a "case-by-case" approach to judging likelihood).

[50] 57 P.3d 654 (Cal. Ct. App. 2002).

[51] *Id.* at 658.

[52] *Id.* at 659.

[53] *Id.* at 660-61.

[54] *Id.* at 662.

[55] Cooley v. Superior Court, 57 P.3d 654, 675 (Cal. Ct. App. 2002).

[56] *Id*. at 661-62.

> empirically established relations between data and the condition or event of interest.[57]

The clinical method is ubiquitous in judicial settings.[58] In a typical, well-done clinical evaluation, the expert examines the individual, gathers and reviews as much other information (e.g., medical and institutional records, court records and other documents pertaining to criminal history) as possible, and applies his expertise to produce an opinion. In many situations, the opinion is expressed in terms that are of direct relevance to the legal question before the court. Thus, in SVP cases, experts are asked, in essence, to characterize the individual's level of risk, using the legally relevant terms such as "highly likely" or "substantially probable."[59]

The actuarial method of risk assessment is relatively new in the judicial context, though it has been used rather extensively in other settings.[60] The term "actuarial" refers to the work done by actuaries or individuals trained to calculate risks using statistics, typically for insurance companies. As we explain in greater detail in part 0, actuarial scales are developed using statistical analyses of groups of individuals (in the present case, released sex offenders) with known outcomes during a "follow-up" period (either arrested for or convicted of a *new* sexual offense, or not identified as having committed a new sexual offense). These analyses tell us which items ("predictor variables") do the best job of differentiating between those who reoffended and those who did not reoffend within a specified time period. Since some of these variables inevitably do a better job than others, these analyses also help us to determine how much weight should be assigned to each item. The variables are then combined to form a scale, which is tested on many other groups of offenders (cross-validation). When the scale has been used on many samples with a sufficiently large number of offenders, the scores derived from the scale may be expressed as estimates of the probability that individuals with that score will reoffend within a specified time frame. At this

[57] Robyn M. Dawes et al., *Clinical Versus Actuarial Judgment*, SCIENCE, March 31, 1989, at 1668, 1668. The authors continue:

> Clinical judgment should not be equated with a clinical setting or a clinical practitioner. A clinician . . . may use the clinical or actuarial method . . . . To be truly actuarial, interpretations must be both automatic (that is, prespecified or routinized) and based on empirically established relations.

Id.

[58] *See*, *e.g.*, Kirk Heilbrun et al., *Risk Communication: Clinicians' Reported Approaches and Perceived Values,* 27 J. AM. ACAD. PSYCHIATRY L. 397, 398 (1999) (listing numerous types of cases in which risk assessment issues are important).

[59] *See Cooley*, 57 P.3d at 674 (reversing in part because experts were not asked to "consider the definition of 'likely' that we conclude applies in the context of a probable cause hearing—whether there is a *substantial danger*, that is, a *serious and well-founded risk*, that Marentez is likely to reoffend") (emphasis in original).

[60] *See* John Monahan, *Violence Risk Assessment: Scientific Validity and Evidentiary Admissibility*, 57 WASH. & LEE L. REV. 901, 906 (2000) (relating the "long tradition in criminology of using actuarial techniques in predicting recidivism by released prisoners").

point, it is possible to develop a "life" or "experience" table that provides probabilistic estimates of reoffense for each score, or range of scores, for different time frames (e.g., within 12, 36, 60 or 120 months).[61] These estimates are usually expressed in terms of the percentage of individuals in the development and validation samples with a particular score who reoffended sexually within the specified time period. These ARA scales are sometimes referred to as "mechanistic," because a statistical formula is used to derive an individual's overall score.

The "experience table" then becomes the focus in risk assessments. The individual to be assessed is scored on the factors, which are combined according to the formula, and the resultant risk score is compared to the table, which yields a probability representing the proportion of the reference group that reoffended. Speaking precisely, we can say that an individual with a particular score has characteristics that place him in a group of persons with the same score who were observed (over the follow-up period) to have a given probability, or frequency, of sexual recidivism.

About a decade ago, in response to increasing requests from the courts to offer opinions on matters of reoffense risk, a number of highly sophisticated researchers began working in earnest to develop reliable, valid actuarial risk assessment instruments for use with sex offenders. This "second generation" of empirical research on risk assessment is a response, at least in part, to the widespread doubts, expressed most vocally in the 1970s, about the ability of mental health professionals to "predict dangerousness."[62] The pace of developments has been rapid, with empirically-driven revisions to and support for existing "static" scales emerging roughly every 12 months, as well as a new wave of research on "dynamic" scales within the past several years,[63] and adaptations of existing scales for special populations, such as juveniles.[64]

In 1954, Paul Meehl wrote a seminal paper, arguing that actuarial methods provide more accuracy than do clinical methods.[65] The empirical support for Meehl's thesis has been demonstrated repeatedly over the ensuing decades, with

[61] *See* Robert A. Prentky & Sue Righthand, *Juvenile Sex Offender Assessment Protocol-II Manual*, at 6, *available at* http://www.csom.org/pubs/JSOAP.pdf (last visited Nov. 18, 2003).

[62] *See infra* text accompanying notes 115 through 118.

[63] *See* sources cited *supra* note 11.

[64] *See, e.g.*, Prentky & Righthand, *supra* note 61, at *i*-7; J.R. Worling, & T. Curwen, *The "ERASOR": Estimate of Risk of Adolescent Sexual Offence Recidivism* (2000) (unpublished manuscript, on file with the American Criminal Law Review); Thomas Grisso, Ethical Issues in Evaluations for Sex Offender Re-offending, Address at the Symposium on Sex Offender Re-Offence Risk Prediction (March 6, 2000) (on file with the American Criminal Law Review).

[65] *See* PAUL E. MEEHL, CLINICAL VERSUS STATISTICAL PREDICTION: A THEORETICAL ANALYSIS AND A REVIEW OF THE EVIDENCE (1954) (discussing the varying strengths and weaknesses of both the actuarial and clinical predictive methods and concluding that the actuarial method is more accurate and is the soundest way to ensure the accuracy of clinical predictive methods).

recent contributions noteworthy for their clarity and persuasiveness.[66] A recent paper reported on a meta-analysis of 136 studies in which predictions by both human judges and "mechanical-prediction schemes" had been compared.[67] In all instances, the predictions fell in the realm of psychology or medicine (i.e., all predictions involved human behavior or medical diagnoses), and in all instances the clinician and the actuarial expert had access to the same predictor variables and made their predictions on the basis of the same criterion.[68]

In only eight out of the 136 studies was clinical prediction superior to actuarial prediction.[69] In 128 studies, either the results were comparable or actuarial prediction was superior. Actuarial prediction was found to be superior in 33% to 47% of the studies, depending on the type of analysis used.[70] Across all of the studies, whether the clinician had access to *more* data did *not* significantly alter the superiority of actuarial prediction.[71] Moreover, in those instances in which the clinician had access to a clinical interview, the superiority of actuarial prediction was even greater.[72] The authors concluded:

> Even though outlier studies can be found, we identified no systematic exceptions to the general superiority (or at least material equivalence) of mechanical prediction. It holds in general medicine, in mental health, in personality, and in education and training settings. It holds for medically trained judges and for psychologists. It holds for inexperienced and seasoned judges.[73]

[66] *E.g.*, ROBYN M. DAWES, HOUSE OF CARDS: PSYCHOLOGY AND PSYCHOTHERAPY BUILT ON MYTH 7-37 (1994); Dawes et al., *supra* note 57, at 1673; Robyn M. Dawes et. al., *Statistical Prediction Versus Clinical Prediction: Improving What Works*, *in* A HANDBOOK FOR DATA ANALYSIS IN THE BEHAVIORAL SCIENCES: METHODOLOGICAL ISSUES 351-367 (1993); David Faust & Jay Ziskin, *The Expert Witness in Psychology and Psychiatry*, SCIENCE, July 1, 1988, at 31, 33-35 (stating actuarial predictions are more accurate than those of professionals and laypersons); William M. Grove & Paul E. Meehl, *Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy*, 2 PSYCHOL. PUB. POL'Y & L. 293, 296-99 (1996)(presenting meta-analysis of studies on risk prediction and concluding the majority of studies favor actuarial methods); Grove et al., *supra* note 8; Janus & Meehl, *supra* note 20; John A. Swets et al., *Psychological Science Can Improve Diagnostic Decisions*, 1 PSYCHOL. SCI. PUB. INTER. 1, 10-11 (2000) (arguing for use of statistical prediction rules in the prediction of future violence because of the weakness of clinical judgment alone); JAY ZISKIN & DAVID FAUST, COPING WITH PSYCHIATRIC AND PSYCHOLOGICAL TESTIMONY (4th ed. 1988).

[67] Grove et al., *Clinical versus Mechanical*, *supra* note 8.

[68] *Id.*

[69] *Id.* at Table 1.

[70] *Id.* at 19.

[71] *Id.*

[72] *Id.*

[73] Grove et al., *Clinical versus Mechanical*, *supra* note 8, at 25.

Two recent meta-analyses further support the conclusion that actuarial assessments of risk are generally superior to clinical assessments.[74] Both meta-analyses reported very small (non-significant) correlations between clinical judgments and recidivism, and stronger correlations between actuarial assessments and recidivism. First, in an aggregation of 61 sexual offender recidivism studies (in which the number of subjects studied was an impressive 23,393), the correlation (r) between clinical assessments and sexual recidivism was .10. The correlation with violent recidivism was .06, and the correlation with recidivism in general was .14. Actuarial methods, in contrast, were much more strongly associated with recidivism, (r of .46 for sexual recidivism, .46 for violent recidivism, and .42 for general recidivism).[75] The second study involved the aggregation of 64 independent samples (derived from 58 studies) in order to examine predictors of recidivism for mentally disordered and non-disordered offenders.[76] In this meta-analysis, the relevant or governing offense was *non*-sexual in 97% of the cases. The correlation between general recidivism and clinical judgment ranged from .06 (lower bound) to .16 (upper bound). By contrast, the correlation between general recidivism and "objective risk" (actuarial) assessment ranged from .34 (lower bound) to .44 (upper bound).[77]

Based on this, and similar, empirical evidence, many scholars have concluded that the predictive efficacy of actuarial methods of risk assessment is superior to clinically derived assessments of risk.[78] Monahan and his colleagues, for example, stated: “The general superiority of statistical over clinical risk assessment in the behavioral sciences has been known for almost half a century.”[79] A similar conclusion was expressed in a review paper published by the Solicitor General of Canada, where “[o]ne of the most consistent findings is that evidence-based, actuarial measures are more accurate in the prediction of offender re-offending or recidivism than professional, clinical judgment.”[80] Another recent article notes, "In literally hundreds of comparisons over many domains including

[74] *See* R. Karl Hanson & Monique T. Bussiere, *Predicting Relapse: A Meta-analysis of Sexual Offender Recidivism Studies*, 66 J. CONSULTING & CLNICAL PSYCHOL. 348 (1998) (examining recidivism studies and identifying the risk factors most positively associated with reoffense); James Bonta et al., *The Prediction of Criminal and Violent Recidivism Among Mentally Disordered Offenders: A Meta-analysis*, 123 PSYCHOL. BULL. 123, 123-142 (1998).

[75] *See* Hanson & Bussiere, *supra* note 74, at 356 tbl. 5.

[76] *See* James Bonta et al., *supra* note 74.

[77] *Id*.

[78] *See* MONAHAN ET AL., *supra* note 9, at 4-8 (comparing weaknesses in accuracy findings of clinical risk assessments with relatively greater accuracy of actuarial methods); Swets et al., *supra* note 66, at 10-11. *See generally* Dawes et. al., *Statistical Prediction Versus Clinical Prediction*, *supra* note 66, at 351-67; Grove et al., *supra* note 8 (reporting first completed meta-analysis of studies comparing clinical and mechanical predictions);

[79] MONAHAN ET AL., *supra* note 9, at 7.

[80] *See*, *e.g.*, Dep’t Solicitor Gen. of Can, *Research Summary: Guidelines for Offender Risk Assessment*, 7 RES. SUMMARY: CORRECTIONS RES. & DEV.6 (2002), *available at* http://www.sgc.gc.ca/publications/corrections/pdf/200211_e.pdf (last visited Oct. 19, 2003).

the prediction of recidivism, clinical judgment has essentially never been found to be superior to actuarial methods, whereas the converse has most often been demonstrated."[81] In commenting on the demonstrated superiority of actuarial over clinical judgment, Meehl remarked, "I do not know of any controversy in the social sciences in which the evidence is so massive, diverse, and consistent."[82] In sum, actuarial methods should be considered, at this point, to represent an "upper bound"[83] in our ability to predict the risk of sexual recidivism.

Clinical risk assessment is, by definition, an exercise in human judgment. The susceptibility of human judgment to error has been the subject of considerable empirical scrutiny. Although by no means exhaustive, the following sources of error in clinical judgments have been noted: (1) ignoring or using incorrect base rates, (2) assigning suboptimal or incorrect weights to information (e.g., over-weighting "high profile" but relatively non-predictive information), (3) failing to take into account regression toward the mean, (4) failing to properly take into account covariation, (5) relying on illusory correlations between predictor variables and the criterion (i.e., basing decisions upon the presence or absence of information that is unrelated or only weakly related to the criterion), [84] (6) failing to acknowledge the natural bias among forensic examiners toward "conservative" judgments, defined as an increased potential for incorrect judgments of dangerousness associated with a reluctance to find someone *not* dangerous,[85] and (7) failing to receive, and thus benefit from, feedback on judgment errors.[86]

In large measure, the superiority of actuarial risk assessment arises from the elimination or reduction of these and other sources of error. As Professors Will Grove and Paul Meehl observe, "[T]he clinician's brain is functioning as merely a poor substitute for an explicit regression equation or actuarial table.

[81] Grant T. Harris et al., *Appraisal and Management of Risk in Sexual Aggressors: Implications for Criminal Justice Policy*, 4 PSYCHOL. PUB. POL'Y & LAW 73, 88 (1998).

[82] Paul E. Meehl, The Power of Quantitative Thinking, Speech upon Receipt of the James McKeen Cattell Fellow Award at the Meeting of the American Psychological Society 3 (May 23, 1998), transcript *available at* http://www.tc.umn.edu/~pemeehl/PowerQuantThinking.pdf (last visited Oct. 19, 2003).

[83] Dawes et al., *Clinical versus Actuarial Judgement*, supra note 57, at 1673.

[84] *Cf.* JOHN MONAHAN, PREDICTING VIOLENT BEHAVIOR: AN ASSESSMENT OF CLINICAL TECHNIQUES 57-67 (Sage Publications, 1981) (discussing common errors in clinical prediction, including vagueness in specifying "dangerousness," disregard of statistical base rates, reliance on weak or nonexistent correlations, and failure to incorporate environmental factors into the analysis).

[85] Edwin I. Megargee, *Methodological Problems in the Prediction of Violence*, *in* VIOLENCE AND THE VIOLENT INDIVIDUAL 179, 188 (J. Ray Hays et al. eds., 1981) ("[M]ental health personnel are much more inclined to over-predict dangerous behavior; that is, we are more likely to be conservative and classify doubtful cases as dangerous.").

[86] *See also* William R. Freudenburg, *Perceived Risk, Real Risk: Social Science and the Art of Probabilistic Risk Assessment*, SCIENCE, Oct. 7, 1988, at 44, 44 (spelling out, in addition to well known sources of human error, the impact of external social forces, monetary and political pressures, and overconfidence, all of which are relevant to present consideration).

Humans simply cannot assign optimal weights to variables, and they are not consistent in applying their own weights."[87] To be sure, ARA has faults, and some ARA tools are better than others. Yet, even the weakest of the actuarial assessment methods appears to be systematically better than clinical judgments. As has been pointed out, any problems present in a poorly designed actuarial method are likely to be equaled or exceeded in clinical assessments.[88]

Given the courts' routine reliance on clinical risk assessment to support long-term liberty-deprivation, it is illogical to exclude demonstrably more reliable ARA tools. In making determinations with serious implications for individual liberty, courts must adopt state-of-the-art methods. As the above discussion indicates, a corpus of empirical evidence demonstrates the predictive superiority of ARA over clinical judgments.

IV. EVALUATING THE ADMISSIBILITY OF ACTUARIALLY-BASED RISK ASSESSMENT IN SVP CASES

We turn now to the forensic arena in which actuarial methods have been most directly addressed: the question of admissibility. We begin by canvassing the basic frameworks that courts employ to judge admissibility. We argue that, at bottom, these methods seek to judge three things about expert or scientific testimony: whether it is sufficiently reliable to be used in the legal context before the court; whether it is sufficiently relevant to the kind of risk that needs to be assessed ("fit"); and whether its use will unduly prejudice or distort the proceedings.

The answer we propose to the first question is, under the circumstances, easy. The baseline for reliability is manifested in the routine admission and reliance on clinical risk assessment in SVP cases. As we have discussed, ARA is at least as reliable, and probably more reliable, than clinical assessments. Thus, ARA meets or exceeds the actual reliability standard in use by SVP courts. Having established this, we wish to make clear that we are not addressing what is a more abstract question: whether the reliability threshold is set too low, or, to put it the other way around, whether, in the abstract, ARA tools have sufficient reliability to support liberty-deprivation. Were the question posed in the abstract, we might agree with Litwack's conclusion that risk assessment tools are not sufficiently validated "for use in determining when individuals should be confined on the grounds of their dangerousness."[89] But if we affirm Litwack's conclusion, we must conclude, *a fortiori*, that clinical judgments of dangerousness are not sufficiently validated either, a conclusion that ought to lead to the abolition of SVP laws.

[87] Grove & Meehl, *supra* note 66, at 315.

[88] *Id.* at 301-02 (arguing in favor of using actuarial instruments in clinical settings because of greater accuracy of combined techniques over clinical judgment alone).

[89] Thomas R. Litwack, *Actuarial Versus Clinical Assessments of Dangerousness*, 7 PSYCHOL. PUB. POL'Y & L. 409, 409 (2001).

The question, however, is not posed in the abstract. Legislatures have mandated that courts perform risk assessments in SVP cases, and courts will undoubtedly continue to oblige by admitting clinical judgments of risk, even if ARA is excluded.[90] The question is not *whether* courts should assess risk, but rather, *how* the risk assessments that are mandated by law should be undertaken.

The questions of fit and prejudice are more difficult. Regarding "fit," we will dismiss many of the broader critiques of ARA, while noting that ARA is not yet very sensitive to the changes in risk status that might be accomplished through effective treatment or well-designed community supervision. We will argue that these are characteristics of ARA that require interpretation and heightened care in its application, but that ought not render ARA inadmissible as evidence.

The question of prejudice raises the concern that the complexity of ARA and its adoption of the mantle of science will combine to render its shortcomings invisible at trial. Invoking Professor Laurence Tribe's denunciation of "trial by mathematics,"[91] we might anticipate that the presumption of innocence will be compromised more by the seemingly inflexible results of a mechanistic formula than by the opinions proffered by clinicians. But we think that there is a potential for prejudice with clinical risk assessment as well, and that ARA offers the best hope of bringing some transparency, accountability and consistency to the judicial risk-finding process. We do acknowledge the dangers of ARA (as well as other forms of risk assessment), and therefore recommend a series of protections in the risk-finding process to increase accountability while reducing the possibility of arbitrary and otherwise improper assessments of risk.

A. The Frye *and* Daubert *Tests*

Courts generally obtain risk assessment from expert testimony.[92] Although several distinct tests are employed to determine the admissibility of expert

[90] *See*, *e.g.*, Commonwealth v. Reese, No. CIV.A. 00-0181-B, 2001 WL 359953, at *9 (Mass. Supp. April 5, 2001) ("Clinical judgment is the predictive means anticipated by the Legislature when it enacted the 1999 legislation regarding sexually dangerous persons . . . ."), *vacated by* 781 N.E.2d 1225 (Mass. 2003). Trial judges in Arizona and Missouri cases explicitly held that the exclusion of ARA testimony did not bar introduction of clinical assessments of risk. *See In re* Woods, No. OP200000005 (Ariz. Sup. Ct. March 20, 2001) (order granting motion to exclude expert testimony based on ARA); *In re* James Francis, No. CV-299-108MH, Docket Memorandum and Judgment 2, ¶ 4 (Circuit Ct. 2000) (finding that clinical assessments "are sufficiently based on empirical findings and approved methodologies such as would withstand *Frye* and *Daubert* scrutiny"). *But see In re* Coffel, No. ED 79989, 2003 WL 716682, at *12 (Mo. Ct. App. 2003) (excluding clinical risk assessment testimony on the grounds that it was not sufficiently grounded in science).

[91] Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329, 1355 (1971).

[92] Some courts have held that fact-finders are not limited to the expert testimony on risk assessment, and some have intimated that courts need no expert testimony to make risk assessment findings. *Compare* Barefoot v. Estelle, 463 U.S. 880, 897-98 (1983) (suggesting that expert testimony is not required to establish dangerousness in death penalty case), *with In re* Pocan, No. 98-1039, 1998 WL 687244 at *2 (Wis. Ct. App. Oct. 6, 1998) ("The trier of fact may accept

testimony, courts appear to be particularly interested in three considerations: the connection between the testimony and the legal issue (relevance or "fit"), the reliability of the evidence, and the potential prejudicial impact of the testimony.

Courts address the admissibility question in two distinct ways. First, the *Frye* test admits "novel" scientific testimony only if it is based on principles generally accepted in the relevant scientific community.[93] Second, under the guidance of *Daubert* and its progeny, federal courts and some states have replaced the *Frye* approach, and require the trial judge to assess the reliability of the testimony.[94] In a sense, while *Daubert* places the reliability assessment on trial judges, *Frye* delegates the quality check to the community of scientists. Finally, some states have developed their own standards.[95]

In *Daubert et al. v. Merrell-Dow Pharmaceuticals*,[96] interpreting the Federal Rules of Evidence, the Supreme Court rejected the argument that *Frye's* general acceptance test remained the central inquiry in the admissibility of scientific testimony.[97] In its place, the *Daubert* Court emphasized the trial judge's independent "gatekeeping" function, determining at the threshold whether the "reasoning or methodology underlying the testimony is scientifically valid and whether that reasoning or methodology can be applied to the facts in issue."[98] The former inquiry relates to the "reliability"[99] of the testimony. The latter inquiry is dubbed "relevance" or "fit," and is further defined by the Court as "whether expert testimony proffered in the case is sufficiently tied to the facts of the case

portions of an expert's testimony, reject other portions and draw conclusions that differ from the expert's.").

[93] *See* Frye v. United States, 293 F. 1013, 1014 (D.C. Cir. 1923); *see also* K.R. FOSTER & P.W. HUBER, JUDGING SCIENCE: SCIENTIFIC KNOWLEDGE AND THE FEDERAL COURTS 225 (1999) (Noting "[t]he '*Frye* rule' was applied by federal courts for more than 50 years and is still enforced by many state courts").

[94] *See generally* Faigman et al., *supra* note 42, at 646, 656 (noting some courts have gone further than the *Frye* test in assessing the soundness of expert testimony).

[95] A recent annotation in the American Law Reports classifies states' approach to expert and scientific testimony. *See* Alice B. Lustre, Annotation, *Post-*Daubert *Standards For Admissibility Of Scientific And Other Expert Evidence In State Courts*, 90 A.L.R.5th 453 (2001). According to this classification, 10 SVP states retain the *Frye* test (Arizona, California, District of Columbia, Florida, Illinois, Kansas, Missouri, Minnesota, North Dakota, and Washington), five have adopted *Daubert* (South Carolina, Texas, Iowa) or the *Daubert* factors (Massachusetts, New Jersey), and two (Wisconsin and Virginia) have developed their own tests.

[96] 509 U.S. 579 (1993).

[97] *Id.* at 592-93 (holding Fed. R. Evid. 702 supercedes the *Frye* general acceptance test).

[98] *Id.*

[99] Courts and legal scholars use various terms when referring to the "integrity" of testimony. Terms such as reliability, validity, dependability, and trustworthiness are frequently used. *See, e.g.*, *Id.*; Faigman et al., *supra* note 42, at 657 (using the term "reliability"). In this essay, we have tried to be consistent in our use of the word "reliability" when referring to the legal concept of admissible testimony. The terms reliability and validity are also used in their technical sense when referring to the psychometric properties of ARA instruments.

that it will aid the jury in resolving a factual dispute."[100] The two inquiries are connected: the Rule "requires a valid scientific connection to the pertinent inquiry as a precondition to admissibility."[101]

The Court set forth four "nonexclusive" factors to be considered in judging scientific reliability: (1) whether the procedure being employed or the theory underlying the procedure is testable; (2) whether there is evidence reported in the literature from peer-reviewed studies supporting the expert's testimony; (3) the known or potential error rate; and (4) whether the procedure or theory was generally accepted by the scientific community.[102] Although these four factors were considered to be "nonexclusive" (i.e., none of four would alone be determinative), the first factor (whether the procedure or theory had been properly tested) was considered "a key question."[103] In a subsequent case, *Kumho Tire Co. v. Carmichael*,[104] the Supreme Court clarified that the "gatekeeping" role for the trial court applies to all expert testimony, not just "scientific" testimony: "The objective of [the] requirement is to ensure the reliability and relevancy of expert testimony."[105] *Daubert* and *Kumho* are interpretations of the Federal Rules of Evidence, and thus do not apply of their own force to state courts.

The manner in which the two standards relate to each other is somewhat complex. As a general matter, however, *Daubert* may be somewhat more permissive when applied to newly developing scientific methods, in the sense that there may be new methods that demonstrate sufficient validity, but are nonetheless too new to have gained "general acceptance." But there may be some ways in which *Daubert* is more restrictive. The extension of *Daubert* to all expert testimony may lead judges to exercise the increased gatekeeping function with respect to clinical judgments, which, as we discuss below, have traditionally been immune from the admissibility scrutiny demanded of "scientific" testimony.[106]

Though there is contrary authority,[107] there is a substantial body of commentary that suggests that the two approaches converge, or are converging,

[100] *Daubert*, 509 U.S. at 591 (quoting United States v. Downing, 753 F.2d 1224, 1242 (3d Cir. 1985)).

[101] *Id*. at 592.

[102] Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 593-94 (1993).

[103] *Id*. at 593.

[104] 526 U.S. 137 (1999).

[105] *Id*. at 152.

[106] *See* cases cited *supra* note 118; Christopher Slobogin, *Doubts About* Daubert*: Psychiatric Anecdata as a Case Study*, 57 WASH & LEE L. REV. 919, 941 (2000) (stressing the reliability requirement applies to all expert testimony). *But see In re* Coffel, No. ED 79989, 2003 WL 716682, at *12 (Mo. Ct. App. March 4, 2003) (rejecting clinical risk assessment testimony on the ground that it was unsupported by science).

[107] *See, e.g.*, *In re* Seibert, No. 01-3361, 2003 WL 722871, at *1 (Wis. Ct. App. March 4, 2003) (per curiam) (holding admissibility of scientific evidence in Wisconsin "is not conditioned upon its reliability," and noting that, with some exceptions for evidence that is relevant, "the witness is qualified as an expert, and the evidence will assist the trier of fact in determining an issue of fact").

on the issue of reliability. Hyongsoon Kim, for example, argues that *Daubert*'s factors are merely "a stricter version of the 'general acceptance' test, a more detailed examination of the manner in which the scientific community has treated the specific theory or technique under review."[108] Peter Knapp demonstrates that Minnesota's application of the *Frye* "general acceptance" test really amounts to an assessment of "the same type of factors used in *Daubert* . . . to assess reliability."[109] K.R. Foster and P.W. Huber argue that "both [tests] refer to the criteria that scientists use to grade the quality, reliability, and overall validity of claims purporting to reflect scientific knowledge. In *Frye* the reference was indirect, by means of the surrogate label 'general acceptance.' In *Daubert,* the reference was direct and explicit."[110]

If, as we have argued, the admissibility standards focus centrally on reliability, we must determine how much reliability SVP courts should demand from risk assessment testimony. No testimony – expert or otherwise – is perfectly reliable. Thus, the reliability determination is not dichotomous, but a question of degree.[111] Neither the Rules of Evidence, nor the Supreme Court's *Daubert* progeny, specify what degree of reliability is required.[112]

In large measure, the reliability standard is a product of the legal purpose for which it is used. As explained by one study: "underlying the admissibility determination lies the policy judgment of how demanding courts should be regarding the level of experience or the amount of research that is necessary before testimony will be allowed."[113] Thus, while evidence law dictates that courts must make a determination of "reliability" at the threshold, it will be the context—here, civil commitment of dangerous individuals—that dictates the degree of reliability required.

A strong argument could be made for requiring a rather high level of reliability for risk assessment testimony. After all, the consequences resting on the assessments are momentous—long-term loss of liberty, on the one hand, and prevention of potential sexual violence on the other. Under such a rigorous standard, it is likely that no risk assessment testimony—clinical or actuarial—would pass muster.[114] Indeed, prevailing wisdom among mental health

[108] Hyongsoon Kim, *Adversarialism Defended:* Daubert *and the Judge's Role in Evaluating Expert Evidence*, 34 COLUM. J.L. & SOC. PROBS. 223, 241 (2001).

[109] Peter B. Knapp, *The Other Shoe Drops: Minnesota Rejects* Daubert, 27 WM. MITCHELL L. REV. 997, 1019 (2000).

[110] FOSTER & HUBER, *supra* note 93, at 228.

[111] *See* Faigman et al., *supra* note 42, at 664-65 (noting in many situations, it would not be appropriate to draw a sharp line between evidence that is admissible and evidence that is sufficient).

[112] *See* Kim, *supra* note 108, at 240 (noting the *Daubert* court "failed to determine what error rate is acceptable or even how to make that determination").

[113] Faigman et al., *supra* note 42, at 650.

[114] *See* Litwack, *supra* note 91, at 409 (noting even though some argue for replacement of clinical assessments of dangerousness with actuarial methods, even the best actuarial tool has not been validated such that it would be appropriate for use); Janus & Meehl, *supra* note 20, at 37 ("At

professionals has, over the years, asserted that the ability to predict violence using clinical methods falls well below a threshold of accuracy that justifies the use of such predictions in legal proceedings.[115] Professor Charles Ewing's unambiguous and often cited conclusion, first asserted in 1983, is that "[t]he psychiatrist or psychologist who makes a prediction of dangerousness violates his or her ethical obligation to render judgments that rest on a scientific basis."[116] Again, in 1991, Ewing asserted that there "is good reason to conclude that psychologists and psychiatrists act unethically when they render predictions of dangerousness that provide a legal basis for restricting another person's interests in life or liberty."[117]

Although we might prefer such a high standard, it is clearly not the reality. Rather, SVP courts routinely and uniformly admit clinical risk assessment testimony, thus establishing a reliability standard that is rather low.[118] The fact that some courts, in other contexts, have excluded unaided clinical judgment as

some point the validity of prediction testimony becomes so attenuated that it is ineffective to establish the requisite certainty of harm.").

[115] *See*, *e.g.*, Megargee, *supra* note 87, at 189-91 (discussing the high false positive rate of clinical predictions and expressing pessimism that clinical predictions will become more accurate); MONAHAN ET AL., *supra* note 84, at 67, 168-69; Joseph J. Cocozza & Henry J. Steadman, *The Failure of Psychiatric Predictions of Dangerousness: Clear and Convincing Evidence*, 29 RUTGERS L. REV. 1084, 1096-99 (1976) (arguing their study shows that psychiatric predictions of dangerousness were as likely to produce false positives as to correctly predict future dangerousness); Bruce J. Ennis & Thomas R. Litwack, *Psychiatry and the Presumption of Expertise: Flipping Coins in the Courtroom*, 62 CAL. L. REV. 693, 751-52 (1974) (stating predictions of dangerousness by psychologists and psychiatrists are not accurate enough to justify use in criminal proceedings); Charles P. Ewing, *"Dr. Death" and the Case for an Ethical Ban on Psychiatric and Psychological Predictions of Dangerousness in Capital Sentencing Proceedings*, 8 AM. J.L. & MED. 407, 418 (1983) ("[C]linical intuition . . . is rarely, if ever, an acceptable substitute for scientific knowledge" and clinical predictions of future violence are too inaccurate to be considered scientifically sufficient."); Charles P. Ewing, Schall v. Martin: *Preventive Detention and Dangerousness Through the Looking Glass*, 34 BUFF. L. REV. 173, 225 (1985) (arguing that inaccuracy in clinical predictions of future violence are so acute that many juveniles detained based on these predictions will be the victims of erroneous prediction); Charles P. Ewing, *Preventive Detention and Execution: The Constitutionality of Punishing Future Crimes*, 15 LAW & HUM. BEHAV. 139, 162 (1991) (stating there is "good reason to conclude that psychologists and psychiatrists act unethically when they render predictions of dangerousness that provide a legal basis for restricting another person's interests in life or liberty" because of the inaccuracy of such predictions).

[116] Ewing, "*Dr. Death*," *supra* note 115, at 418.

[117] Ewing, *Preventive Detention and Execution*, *supra* note 115, at 162.

[118] *See* People v. Ward, 83 Cal.Rptr.2d 828, 831 (Cal. Dist. Ct. App. 1999) ("*Kelly-Frye* applies to cases involving novel devices or processes, not to expert medical testimony, such as a psychiatrist's prediction of future dangerousness or a diagnosis of mental illness."); *see also* Westerheide v. State, 767 So. 2d. 637, 657 (Fla. Dist. Ct. App. 2000) ("The sciences of psychiatry and psychology have been an integral part of American jurisprudence since its inception and although this type of expert testimony is not amenable to mathematical precision, we find that predictions of future dangerousness are sufficiently accurate and reliable to be admissible."). *But see In re* Coffel, No. ED 79989, 2003 WL 716682, at *12 (Mo. Ct. App. March 4, 2003) (rejecting clinical risk assessment testimony on the ground that it was unsupported by science).

falling below the required level of reliability further underlines the laxness of the risk assessment threshold actually in use by courts in SVP cases.[119]

B. The Reliability of Actuarial Assessment

Our argument that ARA is sufficiently reliable to pass muster in SVP cases is bolstered by an examination of the underlying science. As we have indicated, actuarial methods are empirically based. Their development is based on empirically observed relationships between measurable characteristics of the individual and the outcome variable of interest (here, sexual recidivism). As our discussion in this section shows, there are several salutary consequences of this empiricism, accounting in large measure for the advantages of ARA compared to clinical methods. First, and most obviously, the empirical basis means that actuarial tools are likely to capture real, as opposed to illusory, relationships between predictors and outcomes. Second, the efficacy of the tools can be measured and reported with a high degree of precision (interrater reliability of judgments,[120] measurement error,[121] and predictive validity[122]). We can make

[119] *See, e.g.*, Wynacht v. Beckman Instruments, Inc., 113 F. Supp. 2d 1205, 1210-11 (E.D. Tenn. 2000) (excluding clinical medical testimony on causation of toxic injuries); Case of Canavan, 733 N.E.2d 1042, 1050 (Mass. 2000) (holding, with regard to the state's test for admissibility, "[t]here is no logical reason why conclusions based on personal observations or clinical experience should not be subject to the *Lanigan* analysis. Observation informed by experience is but one scientific technique that is no less susceptible to *Lanigan* analysis than other types of scientific methodology"); *cf.* Slobogin, *supra* note 108, at 922 (noting that clinical opinion testimony is "frequently of questionable validity," but arguing that where "such testimony concerns past mental state and is proffered by a criminal defendant, it should be admissible even under the *Daubert-Kumho* regime that exists in the federal courts and many state jurisdictions.").

[120] Reliability refers to how well the test measures what it was designed to measure (i.e., is the test consistent and accurate). There are a variety of ways of examining reliability, the most important of which for present purposes is "inter-rater reliability." Inter-rater reliability refers to the agreement between two (or more) raters. Presented with the identical information, how often will two raters independently agree on how they score or rate an individual. The most common reason for unreliability in scales such as ARA is poorly worded or ambiguously worded descriptions of the items in the scale. *See* GUILFORD, *supra* note 42, at 395-397.

[121] Whenever we set out to measure anything, that measurement will contain some amount of chance error. We are very concerned about "measurement error," because it undermines reliability. The amount and quality of information that is used for rating an ARA scale is undoubtedly the most common and the most obvious potential source of measurement error. In addition, however, there is the "human" factor. Ultimately, it is still humans that read and process information and make the ratings. We assume that raters are capable of some degree of precision and objectivity. We know full well, however, that there are many sources of personal bias that enter when it comes to making judgments. These sources of bias also contribute to measurement error. *See* GUILFORD, *supra* note 42, at 398-400.

[122] Undoubtedly, the most critical feature of a test is its validity. Validity refers, quite simply, to whether the test measures what it purports to measure. A test may be highly reliable but not valid. Although there are different ways of examining validity, the most important for an ARA scale is predictive validity. Predictive validity refers to how well the test, or in this case the ARA scale, predicts the outcome under scrutiny (in this case, sexual recidivism). In order to establish predictive validity, there must be empirical evidence that the scores on the ARA scale are highly correlated with sexual recidivism. This is a complex question, since it may be revealed that the

judgments about the adequacy of the tools based on these measurements, and we can compare the tools among themselves. Third, the science can build on itself incrementally. As shortcomings are identified in existing tools, developers can work to remedy the problems and improve accuracy.[123] Fourth, the actuarial method replaces the opacity of CRA with a transparency that makes visible the key strengths, and limitations, of risk assessment.

These are key characteristics of ARA that distinguish it from clinical risk assessment. They account, at least in part, for the advantages of ARA over clinical methods. In the following section, we show how the development and evolution of several of the more widely used ARA instruments give rise to these important characteristics of ARA.

1. *The Science*

We begin our discussion of the science of ARA by discussing the Violence Risk Appraisal Guide ("VRAG"), undoubtedly the most frequently reported actuarial risk assessment scale in the empirical literature.[124] The following summary of the development of the VRAG demonstrates the way in which ARA is empirically based, and illustrates the use of statistics to evaluate the accuracy of ARA tools in identifying recidivists. We follow that discussion with summaries of several other instruments, emphasizing the evaluative and evolutionary courses researchers have pursued in the development of ARA.

The VRAG was developed to assess violent recidivism. The initial development was based on a sample of 618 men (about 15% of whom were sex offenders) who had been committed – and later released -- as mentally disordered offenders to the maximum security psychiatric hospital in Penetanguishene, Ontario, for assessment or treatment. The men were followed after release to determine which engaged in any "violent" recidivism, an outcome variable that included, *inter alia,* all "hands-on" sexual offenses.[125] The average time "at risk" in the community was about seven years. Almost one-third of the sample committed a new violent offense during the follow-up period. A large number of

scale has greater predictive validity for certain types of sex offenders (e.g., it works better for rapists than for child molesters), or for certain types of criminal offenses (e.g., it works better for general violence than for sexual offenses), or for certain lengths of follow-up (e.g., better for short-term than long-term), etc. In addition, it is often the case that the predictive validity that is demonstrated on the development sample "shrinks" (is less impressive) when the scale is used on a new, different sample (i.e., on cross-validation). *See* GUILFORD, *supra* note 42, at 278-285.

[123] See *infra* text accompanying notes 138-162.

[124] *See generally* G. T. Harris et al., *Violent Recidivism of Mentally Disordered Offenders*, 20 CRIM. JUST. & BEHAV. 315 (1993); V.L. QUINSEY ET AL., VIOLENT OFFENDERS: APPRAISING AND MANAGING RISK (1998); M .E. Rice & G. T. Harris, *Violent Recidivism: Assessing Predictive Validity*, 63 J. CONSULTING & CLINICAL PSYCHOL. 737 (1995); M. E. Rice & G. T. Harris, *Cross-validation and Extension of the Violence Risk Appraisal Guide for Child Molesters and Rapists*, 21 L. & HUM. BEHAV. 231 (1997); C.D. WEBSTER ET AL., THE VIOLENCE PREDICTION SCHEME: ASSESSING DANGEROUSNESS IN HIGH RISK MEN (1994).

[125] V.L. QUINSEY ET AL., *supra* note 124, at 142 (explaining the decision to include all sexual assaults involving physical contact in the definition of violent offenses).

potential predictors of violence was examined, and twelve variables were selected as particularly related to subsequent violence:[126] (1) separation from parents before age 16; (2) elementary school maladjustment; (3) alcohol abuse history; (4) never married; (5) history of nonviolent offenses; (6) failure on prior conditional release; (7) age at index offense; (8) victim injury in index offense; (9) male victim in index offense; (10) diagnosis of any personality disorder according to the Diagnostic and Statistical Manual of the American Psychiatric Association (DSM[127]); (11) diagnosis of schizophrenia according to the DSM; and (12) Hare's Psychopathy Checklist (PCL-R) Score.[128] These variables were numerically combined to constitute the VRAG.

In an early study, the correlation between the VRAG (i.e., the combination of the twelve predictor variables) and violent recidivism was .46.[129] The two predictors with the highest correlations with violent recidivism were the Psychopathy Checklist (.34) and elementary school maladjustment (.31).[130]

As is common practice, the developers tested the VRAG using an independent sample of 159 sex offenders that were not included in the original construction sample. This cross-validation study yielded similar results.[131] That is, the correlation of the VRAG with violent recidivism was quite comparable (.47) to the correlation of .46 observed in the original study.[132]

As indicated, the VRAG was initially developed to predict violent recidivism (not limited to sexual recidivism). Given the level of interest in the prediction of sexual recidivism, the developers assessed the ability of the VRAG to predict outcomes that were limited to sexual recidivism. That inquiry suggested that the VRAG does a better job at predicting violent recidivism (nonsexual as well as sexual) than at predicting general sexual recidivism, which inevitably includes many crimes that are on the low end of a violence continuum. In the cross-validation study, the VRAG's correlation with sexual recidivism (i.e., only sexual crimes) was .20.[133]

Although there is no uniformly accepted index of accuracy for predictive models such as the VRAG, the "AUC value" is generally regarded as an index

[126] *Id.* at 147.

[127] AMERICAN PSYCHIATRIC ASSOCIATION, DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS (3d ed. 1980)

[128] ROBERT D. HARE, THE HARE PCL-R RATING BOOKLET (1991) (noting that psychopath is a personality construct that includes two core sets of traits, one of which focuses on the callous indifference to the welfare of others, and the other, which focuses on a chronically impulsive, antisocial lifestyle.)

[129] Harris, et al., *Violent Recidivism*, *supra* note 124.

[130] V.L. QUINSEY ET AL., *supra* note 124, at 147.

[131] Rice & Harris, *Violent Recidivism*, *supra* note 124, at 737.

[132] *Id.*

[133] *Id.*

that should be reported.[134] The AUC value corresponds to the probability of accurately predicting that a randomly selected, truly dangerous individual is more likely to be dangerous than a randomly selected, truly non-dangerous individual.[135] Near-perfect accuracy in discriminating between dangerous and non-dangerous individuals would yield a AUC value that approached 1.00, while chance prediction would yield a AUC value of .50.

Again, studies examining the AUC value suggest that the VRAG may have better predictive capabilities in terms of violent recidivism as compared to sexual recidivism. Mossman examined 58 studies of violence prediction, finding that the median AUC value for all 58 studies was .73 and the weighted average was .78.[136] Rice and Harris reported that the VRAG's AUC value associated with violent recidivism was .77, a result comparing favorably with the group of studies Mossman reported on. However, the VRAG's AUC value associated with sexual recidivism (.60) clearly was suboptimal. As Rice and Harris clearly stated, the mission for VRAG is interpersonal violence.[137] Thus, it is not surprising that the VRAG falls short when it comes to differentiating among samples exclusively comprised of sexual offenders, many of whom have minimal (or no) history of physical violence. The VRAG variable with the greatest weighting is the PCL-R score[138] and none of the twelve items capture the sexual pathology (e.g., sexually deviant thoughts/fantasies, intensity of sexual preoccupation with children, amount of contact with children) that would seem to be critical for most child molesters and some types of rapists.[139] It would certainly seem that any attempt to predict sexual recidivism must take into account sexually deviant thoughts and behaviors.

These findings regarding the comparative functioning of the VRAG in the prediction of general versus sexual recidivism illustrate several of the strengths of ARA. Unlike clinical risk assessment, in which the ability of the examiner must be taken on faith, ARA allows a quantification of its accuracy, and a comparative examination of accuracy. Although the VRAG has *some* relation to sexual recidivism, it is not as good at predicting sexual recidivism as it is at predicting

[134] *See* Frank E. Harrell, Jr. et al., *Regression Modeling Strategies for Improved Prognostic Prediction*, 3 STAT. MED. 143 (1984). AUC is the "area under the curve" in a graph that plots the "true positive rate" (sensitivity) as a function of the "false positive rate" (1-specificity). The more accurate the scale, the greater the area under the curve.

[135] *See* A. S. Ash & M. Schwartz, *Evaluating the Performance of Risk-Adjustment Methods: Dichotomous Measures*, *in* RISK ADJUSTMENT FOR MEASURING HEALTH CARE OUTCOMES 313-346 (L. I. Lezzoni ed., 1994).

[136] Douglas Mossman, *Assessing Predictions of Violence: Being Accurate About Accuracy*, 62 J. CONSULTING & CLINICAL PSYCHOL. (1994).

[137] Rice & Harris, *Cross-Validation*, *supra* note 124.

[138] Rice & Harris, *Violent Recidivism*, *supra* note 124, at 740.

[139] Hanson & Bussiere, *Predicting Relapse*, *supra* note 75, at 348 (finding that "[s]exual offense recidivism was best predicted by measures of sexual deviancy (e.g., deviant sexual preference, prior sexual offenses) and, to a lesser extent, by general criminological factors").

general violence. Courts can use the statistical information about accuracy – both absolute and comparative – to begin to make more informed decisions about risk.

Further, the scientific nature of ARA allows for a progression of technique and knowledge, as researchers seek to overcome the shortcomings in existing ARA tools. This is well illustrated by the work of Vern Quinsey, Marnie Rice, and Grant Harris, who took on the problem of the diminished accuracy of the VRAG with sex offenders.[140] These researchers employed the same construction procedure used with the VRAG. They examined the predictive efficacy of a large number of variables on a sample of child molesters and rapists (predominantly child molesters).[141] In their combined sample, they found support for the predictive validity of ratings on the Psychopathy Checklist, penile plethysmographic assessment,[142] and prior criminal history.[143] Several of the variables from this study were subsequently incorporated into a new scale, called the Sex Offender Risk Appraisal Guide (SORAG). The SORAG includes eleven of the twelve items on the VRAG. One VRAG item (victim injury) was dropped, and one item (male victim) was changed to: Sexual offenses only against girls under 14. The three new SORAG items are: (1) History of violent offenses, (2) Number of prior convictions for sexual offenses, and (3) Deviant sexual preference (phallometric test results).[144]

Recent studies utilizing the SORAG have been quite encouraging.[145] Marnie Rice and Grant Harris examined the predictive efficacy of the SORAG with incest offenders.[146] The SORAG worked as well for incest offenders as it did for non-incest sex offenders.[147] When examining violent recidivism, the AUC

[140] V. L. Quinsey et al., *Actuarial Prediction of Sexual Recidivism*, 10 J. INTERPERSONAL VIOLENCE 85 (1995).

[141] *Id.* at 85.

[142] The penile plethysmographic assessment (PPG) is a physiological test for examining degree of sexual arousal in response to depictions of sexual stimuli. *See* D. Richard Laws et al., *Assessment of Sex Offenders Using Standardized Slide Stimuli and Procedures: A Multi-Site Study, 7* SEXUAL ABUSE: J. RES. & TREATMENT 45 (1995)

[143] Quinsey et al., *supra* note 140, at 85.

[144] *Id.*

[145] *See, e.g.*, Howard E. Barbaree et al., *Evaluating the Predictive Accuracy of Six Risk Assessment Instruments for Adult Sex Offenders*, 28 CRIM. JUST. & BEHAV. 490 (reporting an AUC value on the construction sample at .73, and a AUC value on the first cross-validation sample at .76); N. Belanger & C. Earls, *Sex Offender Recidivism Prediction*, 8 F. CORRECTIONS RES. 22-24 (1996) (reporting an AUC value of .82 when the SORAG was used on a sample of 57 sex offenders released from prison); P. Firestone et al., *A Comparison of the Sex Offender Risk Appraisal Guide (SORAG) and the Static-99*, (Sept. 1999) (unpublished paper, on file with ACLR) (presented at the Annual Meeting of the Association for Treatment of Sexual Abusers); Kevin L. Nunes et al., *A Comparison of Modified Versions of the Statis-99 and the Sex Offender Risk Appraisal Guide*, 14 SEXUAL ABUSE: J. RES. & TREATMENT 253 (2002); Grant T. Harris et al., *A Multi-site Comparison of Actuarial Risk Instruments for Sex Offenders*, 15 PSYCHOL. ASSESSMENT 413-25 (2003).

[146] Marnie E. Rice & G. T. Harris, *Men Who Molest Their Sexually Immature Daughters: Is a Special Explanation Required?* 111 J. ABNORMAL PSYCHOL. 329 (2002).

[147] *Id.* at 329.

value for the entire sample was .76, compared with .80 for incest offenders only. When examining sexual recidivism, the AUC value for the entire sample was .81, compared with .67 for incest offenders only.[148]

The work of R. Karl Hanson and David Thornton further illustrates the evolutionary course and empirical methodology that characterize the development of ARA tools. [149] Hanson used aggregate data from eight follow-up studies that included 2,592 subjects.[150] He examined seven variables that had emerged as important from an earlier meta-analysis. These seven variables included: (1) "officially recorded" prior sex offenses, (2) stranger victims, (3) any prior non-sexual offenses, (4) age (at time of release for those who were in prison and at time of evaluation for those in the community), (5) marital status, (6) any non-related victims (victims not having a biological, step, or foster relationship with the offender), and (7) any male victims (child or adult).[151] From these seven variables, Hanson chose the four that were most strongly associated with sexual recidivism. The resulting scale, dubbed the Rapid Risk Assessment for Sex Offense Recidivism scale (RRASOR) combined just four variables: (1) prior sexual offenses, (2) age at risk less than 25, (3) extrafamilial victims, and (4) male victims.[152] This scale correlated .27 with sexual recidivism using the scale development samples. The AUC value was .71.[153] Using a different "validation sample," the scale correlated .25 with sexual recidivism, and the AUC value was .67.[154]

Meanwhile, David Thornton, working independently, developed the Structured Anchored Clinical Judgment (SACJ) scale.[155] Unlike the RRASOR, the SACJ was rated using a multi-stage process. In the first stage, documented convictions were coded in the following five areas: (1) any current sexual offense; (2) any prior sexual offense; (3) any current non-sexual violent offense; (4) any prior non-sexual violent offense; and (5) four or more prior (distinct) sentencing occasions.[156] If four or five of the above factors were coded as present, the offender was automatically classified as high risk. If two or three factors were present, the offender was classified as medium risk. If one or none of the factors were present, risk was considered low.[157]

[148] *Id.*

[149] *See* R. Karl Hanson & David Thornton, Public Works & Gov't. Serv., Can., *Static-99: Improving Actuarial Risk Assessment for Sex Offenders*, User Report 99-02 (1999), *available at* http://www.sgc.gc.ca/corrections/publications_e.asp (last visited Nov. 18, 2003).

[150] R. Karl Hanson, *What Do We Know About Sex Offender Risk Assessment?*, 4 PSYCHOL. PUB. POL'Y & L. 50, 64-65 (1998).

[151] *Id.*

[152] *Id.* at 64.

[153] *See* Hanson & Thornton, *supra* note 149, at 2.

[154] Hanson & Bussiere, *supra* note 75.

[155] *See* Hanson & Thornton, *supra* note 149.

[156] *Id.* at 2-3.

[157] *Id.*

The second stage incorporated one of two sets of variables that are regarded as potentially aggravating factors. Set A included the following four variables: (1) any stranger victims; (2) any male victims; (3) never married; and (4) convictions for non-contact sex offenses. Set A was relatively easy to code quickly and reliably. Hence, the five Stage 1 items plus the four Set A items comprised the SACJ-Min -- the minimum required for a valid assessment. The four Set B items, several of which are more time-consuming and difficult to code, included: (1) Substance abuse, (2) Deviant sexual arousal, (3) Psychopathy, (4) Placement in residential care as a child.[158]

The SACJ was developed through exploratory analyses on several datasets in England. The SACJ-Min was validated on a different sample of approximately 500 sex offenders released from prisons in 1979. Follow-up data were collected on the complete sample after 16 years. In this validation study, the SACJ-Min correlated .34 with sexual recidivism and .30 with any sexual or violent reoffense.[159].

The Static-99 represents the combined efforts of Hanson and Thornton to integrate the RRASOR and the SACJ-Min.[160] As the name implies, the scale includes only static variables. The year "99" suggests that the scale is a work in progress. The Static-99 includes 10 variables: eight of the nine original SACJ-Min variables (only Current sex offense was dropped) and all four of the RRASOR variables. Since two of the four RRASOR variables were also on the SACJ-Min, only two new variables were added to the eight SACJ-Min variables.

Like the SORAG, the Static-99 has been the subject of many empirical studies.[161] As noted in discussion of the SORAG, the results from the multi-scale comparison study of Barbaree et al. provided comparable support for the Static-99 and the SORAG, with AUC values of .71 (any reoffense), .70 (any serious offense), and .70 (any sexual offense).[162] In a recent cross-validation study on a large Swedish sample of 1,400 male sex offenders, the AUC values were .74 for any violent offense and .76 for a sexual offense.[163] In this study, the Static-99 predicted sexual recidivism comparably for child molesters (C = .76) and rapists (C = .75).[164]

As further illustration of the evolutionary process of science, as it applies to the ongoing development of increasingly accurate actuarial assessment tools, Hanson and Thornton recently released the Static-2002, a revision of the Static-

[158] *Id* at 4.

[159] *Id.*

[160] *Id.*

[161] *See generally* Barbaree et al., *supra* note 145; Firestone, et al., *supra* note 145; Harris et al., *A Multi-Site Comparison*, *supra* note 122.

[162] Barbaree et al., *supra* note 145, at 507.

[163] Gabrielle Sjöstedt & Niklas Långström, *Actuarial Assessment of Sex Offender Recidivism Risk: A Cross-Validation of the RRASOR and the Static-99 in Sweden*, 25 L. & HUM. BEHAV. 629, 637 (2001).

[164] *Id.* at 637.

99.[165] The Static-2002 has 13 risk predictors, three more than the Static-99. With five new items, coding changes to at least four other items, and one Static-99 item dropped,[166] the Static-2002 must be considered a substantially different scale.

In a recent comparative analysis of the predictive efficacy of five actuarial risk assessment procedures and the PCL-R, Howard Barbaree et al. found strong support for the SORAG.[167] Among all of the examined procedures, the SORAG had the largest AUC value when predicting any *serious* reoffense (.73), compared with .70 for the Static-99, .69 for the VRAG, .65 for the RRASOR, .65 for the PCL-R, and .58 for the MnSOST-R, an actuarial risk assessment tool developed for use in connection with Minnesota's program of community notification.[168] When predicting *any* reoffense, the SORAG (.76) and the VRAG (.77) were better than the other procedures, which ranged from .71 to .60. When predicting *any sexual* reoffense, the best results were from the RRASOR (.77), the Static-99 (.70) and the SORAG (.70), with AUC values of .65 to .61 obtained for the MnSOST-R, the VRAG, and the PCL-R.[169]

In the most recent comparative study, the efficacy of the VRAG, SORAG, RRASOR, and Static-99 were examined in four independent samples totaling 396 sex offenders.[170] The correlation between the SORAG and violent recidivism was .38, ranging from .31 to .37 across samples, while the equivalent correlation for the Static-99 was .21, ranging from .13 to .25 across samples. The AUC values for the SORAG ranged from .69 to .77, while the AUC values for Static-99 ranged from .60 to .67. When the scales were used to predict sexual recidivism, the AUC values for the SORAG ranged from .59 to .71, while the AUC values for the Static-99 ranged from .54 to .67. Both scales performed slightly better for child molesters than rapists. When predicting sexual recidivism for child molesters, the AUC values were .70 for the SORAG and .65 for the Static-99. The equivalent values for rapists were .62 and .59, respectively. Under "favorable conditions" (e.g., fewer missing items and a fixed follow-up time), the AUC values for prediction of sexual recidivism were as high as .79 for the SORAG and .76 for the Static-99.[171]

[165] R. Karl Hanson & David Thornton, Dep't. Solicitor Gen. Can., *Notes on the Development of Static-2002*, (2003)*,available at* http://www.sgc.gc.ca/publications/corrections/200301 (last visited Oct. 19, 2003).

[166] *See id*.

[167] Barbaree et al., *supra* note 145, at 507.

[168] *See generally* Douglas L. Epperson et al., Minn. Dep't Corr., *Minnesota Sex offender Screening Tool – Revised (MnSOST-R): Development, Performance, and Recommended Risk Level Cut Scores*, *available at* http://129.186.143.73/faculty/epperson/mnsost_download.htm (last visited Oct. 19, 2003).

[169] Barbaree, et al*, supra* note 145, at 507.

[170] *See* Harris et al., *A Multi-Site Comparison*, *supra* note 145.

[171] *Id.*

2. *Appellate Decisions on Reliability*

How should courts evaluate this science? Clearly, ARA is a serious enterprise, backed by sophisticated empirical methodology. Yet, on the other hand, critics are quick to enumerate the many ways in which it is imperfect, including this partial list: small sample sizes, lack of cross-validation,[172] inadequate number of peer-reviewed publications, absence of information on standard errors, and absence of manuals with standardized instructions for scoring.[173] To a greater or lesser extent, all ARA instruments have shortcomings, and these shortcomings detract from the reliability of the instruments.[174] Still, the question in the admissibility context is not whether the method is perfectly reliable, but whether it has *sufficient reliability* to be considered by the trier of fact.

Three appellate courts have addressed the issue of reliability.[175] Two of the three admitted ARA,[176] while the third excluded it.[177] While we will discuss these three cases in more detail shortly, we anticipate that discussion by making three general observations from these three cases. First, none of the three courts engaged in a sophisticated evaluation of the science underlying ARA. Rather, the admitting courts appear to be saying, in a general way, "this science seems weighty," while the essence of the excluding court's reasoning was that the evidence about the science seemed rather thin. Second, all three seemed to evaluate reliability in the context of potential prejudice. In other words, the question seemed to be not "how accurate does risk assessment have to be to justify liberty deprivation" but rather, how accurate does it have to be to avoid potential prejudice arising from labeling ARA as "science." This point is bolstered by the third observation, which is that these courts

[172] *See* Richard Wollert, *The Importance of Cross-Validation in Actuarial Test Construction: Shrinkage in the Risk Estimates for the Minnesota Sex Offender Screening Tool-Revised*, 2 J. THREAT ASSESSMENT 87, 95 (2002) (demonstrating that cross-validation of a commonly used ARA tool results in dramatic "shrinkage" of its predictive power); *see, e.g.*, Dawes et al., *supra* note 57 at 1668 (discussing need for cross-validation to avoid artificially inflating accuracy of actuarial instruments).

[173] People v. Taylor, 782 N.E.2d 920, 931 (Ill. App. Ct. 2002) (detailing shortcomings of actuarial tools).

[174] The flaws may have several effects. For example, the failure to do adequate cross-validation may inflate materially the recidivism probabilities associated with certain test scores. The failure to provide careful instructions for administration may inflate the measurement error, allowing recidivism-biased scoring to inflate results. Small development samples may produce large sampling errors, and so forth. *See, e.g.*, James Bonta et al., *The Prediction of Recidivism Among Federally Sentenced Offenders: A Re-Validation of the SIR Scale*, 38 CAN. J. CRIMINOLOGY 61 (1996); Rice & Harris, *Cross-Validation*, *supra* note 124 (stating that actuarial recidivism prediction instruments derived on large samples have been exceptionally stable on cross-validation).

[175] *See In re* R.S., 773 A.2d 72, 74 (N.J. Super. Ct. App. Div. 2002), *aff'd*, 801 A.2d 219 (N.J. 2002) (per curiam); *In re* Holtz, 653 N.W.2d 613 (Iowa Ct. App. 2002); Taylor, 782 N.E.2d at 922.

[176] *See In re* R.S., 773 A.2d at 96; In re Holtz, 653 N.W.2d at 619.

[177] *See* People v. Taylor, 782 N.E.2d 920, 931 (Ill. App. Ct. 2002).

judged the potential prejudice of ARA in part by its relationship to clinical risk assessment. In all three cases, ARA was used in conjunction with a full clinical assessment. The two admitting courts thought that this conjunction was significant in that it would serve to make clear to the jury that ARA was just another piece of information, passed through the judgment of the clinician, and in this way undercut its (undue) influence as "science." The third court turned this logic on its head, opining that the inclusion of ARA in the clinician's information base would transform the (otherwise admissible) clinical judgment into potentially prejudicial "science." We shall return to a discussion of the relationship between clinical and actuarial methods below. Here, it is sufficient to note that all three courts apparently thought that clinical judgments would be routinely admitted, even if ARA were excluded.[178]

In *In re R.S.*, a New Jersey appellate court held that ARA is "reliable for use in [sex offender commitment cases] as an aid in predicting recidivism."[179] In a crucial passage, the court noted that "actuarial instruments are at least as reliable, if not more so, than clinical interviews."[180] It continued:

> Since expert testimony concerning future dangerousness based on clinical judgment alone has been found sufficiently reliable for admission into evidence at criminal trials, we find it logical that testimony based upon a combination of clinical judgment and actuarial instruments is also reliable. Not only does actuarial evidence provide the court with additional relevant information, in the view of some, it may even provide a more reliable prediction of recidivism.[181]

Although the court asserted that "a substantial amount of reliability must be assured before scientific evidence may be admitted,"[182] its discussion of reliability was somewhat superficial. The court seemed impressed with the scientific method without delving directly into the controversy about the adequacy of the science.

Reliability, according to the court, is contextual. "[W]hat constitutes reasonable reliability depends in part on the context of the proceedings involved."[183] For this court, context is determined, at least in part, by a "weighing of reliability against prejudice . . . . Expert evidence that poses too great a danger of prejudice in some situations, and for some purposes, may be admissible in

[178] *See In re* R.S., 773 A.2d at 74; Taylor, 782 N.E.2d at 975; *In re* Holtz, 653 N.W.2d at 619.

[179] *In re* R.S., 773 A.2d at 75.

[180] *Id.* at 90 (citing Matter of C.A., 679 A.2d 1153 (N.J. 1996)).

[181] *See In re* R.S., 773 A.2d 72, 90 (N.J. Super. Ct. App. Div. 2002), *aff'd*, 801 A.2d 219 (N.J. 2002) (per curiam);

[182] *Id.* at 91.

[183] *Id.*

other circumstances where it will be more helpful and less prejudicial."[184] The court's assessment of this balance turned heavily on the fact that in New Jersey commitments are tried to a judge, not a jury.[185] On review, the New Jersey Supreme Court affirmed. Apparently impressed with the weight of the science ("[t]he extensive expert testimony in this matter concerning validation studies, cross-validation studies, reliability studies, correlation coefficients, and clinically-derived factors attests to . . . reliability in this context"),[186] the court nonetheless suggested strongly that its holding might be limited to the use of ARA only as part of a broader clinical evaluation. ARA, said the court "are not litmus tests."[187]

The Iowa intermediate appellate court, sitting en banc, took essentially the same tack in *In re Holtz*.[188] Reliability, the court noted, is contextual: "the amount of foundation necessary to establish reliability depends on the complexity of the testimony and the likely impact of the testimony on the fact-finding process. . . ."[189] Neither the district court nor the appellate court undertook any independent review of the science. Citing the New Jersey case, and expert testimony, the court admitted the ARA-based testimony, but warned that, "[t]he instruments were used in conjunction with a full clinical evaluation and their limitations were clearly made known to the jury,"[190] thus suggesting, like the New Jersey court, that the clinical context inoculated the ARA from potential prejudice.

The only appellate court to reject the admissibility of ARA was the Illinois intermediate appellate court in *People v. Taylor*.[191] The court first determined that actuarially-based testimony is subject to a *Frye* analysis.[192] The court acknowledged that, under Illinois precedents, clinically-based psychological testimony is not subject to *Frye*.[193] The court then rejected the approach of several other courts that exempts hybrid clinical-actuarial testimony from *Frye*.[194] In the court's judgment, expert use of actuarial methods subjects the normally exempt testimony to *Frye* scrutiny. The court proceeded to subject the proffered testimony to a *Frye* analysis, finding that the state had failed in its burden to establish that the actuarial instruments relied upon had achieved the level of validity required for admissibility. The court concluded that the "instruments are

[184] *Id.*

[185] *See id.*

[186] *See In re* R.S., 773 A.2d 72, 91 (N.J. Super. Ct. App. Div. 2002), *aff'd*, 801 A.2d 219 (N.J. 2002) (per curiam);

[187] *Id.*

[188] *See In re* Holtz, 653 N.W.2d 613, 615 (Iowa Ct. App. 2002) (quoting Johnson v. Knoxville Cmty. Sch. Dist., 570 N.W.2d 633, 637 (Iowa 1997) (citations omitted)).

[189] *Id.*

[190] *In re* Holtz, 653 N.W.2d 613, 619-20 (Iowa Ct. App. 2002).

[191] *See* People v. Taylor, 782 N.E.2d 920, 932 (Ill. App. Ct. 2002)

[192] *Id.* at 930.

[193] *Id.*

[194] *Id.* at 932.

still in the experimental stages and that the validity of these instruments has not been established."[195] With respect to one test, the MnSOST-R, the court noted that the developers had not "released the raw data upon which the MnSOST-R was based, and other researchers have not had the opportunity to replicate and scrutinize the study."[196] The state, according to the court, did not introduce sufficient "statistical evidence demonstrating the reliability and accuracy of these instruments."[197] The court also noted "frequent scoring inconsistencies by different evaluators" and the absence of any "rules . . . on the methods or procedures to combine the results of the various instruments and what weight should be placed upon the instruments in evaluating sexual offender recidivism."[198] The court concluded: "Lacking a threshold showing of any indicia of validity, these instruments should not be presented to the jury as 'science.'"[199] The court noted that the state's witness had claimed that "these instruments are more accurate than pure clinical judgment."[200] But the court refused to credit this testimony, reasoning that the state's witness "offered no support for his conclusion" other than his "own assertions."[201]

In sum, the appellate treatment of scientific reliability gives us only the roughest of benchmarks, suggesting that voluminous testimony on the science reassures courts, while skimpy testimony reinforces judicial worries. We will suggest below that the more fundamental question these courts seem to be addressing is mostly focused on prejudice: whether ARA has enough science to justify the imprimatur of the "science" label.

C. Fit

"Fit" is a basic component of admissibility, and measures testimony's "connection to the pertinent [legal] inquiry".[202] "Fit" addresses whether the risk that is measured by the ARA tools is the same as the risk that must be determined under the governing law. As a preliminary matter, we note that "fit" is a concern both in the context of admissibility and in assessing the weight to be given to particular expert testimony. In this section, we point out that ARA might not answer the precise question posed in SVP cases. But, we argue that the lack of precise fit should not exclude ARA, but rather that ARA requires interpretation and judgment to determine its proper place in determining risk.

[195] *Id.* at 931.

[196] People v. Taylor, 782 N.E.2d 920, 931 (Ill. App. Ct. 2002)

[197] *Id.*

[198] *Id.*

[199] *Id.*

[200] *Id.* at 932.

[201] People v. Taylor, 782 N.E.2d 920, 932 (Ill. App. Ct. 2002)

[202] *See* Daubert v. Merrell Dow Pharm. Inc., 509 U.S. 579, 592 (1993).

ARA poses two types of fit problems, which we will refer to as the outcome-measure problem and the group-based problem. Both problems, we suggest, are also present in clinical risk assessment, but hidden by the opaque nature of CRA. ARA provides the transparency that makes the fit question even worthwhile asking.

The outcome-measure problem is quite concrete. ARA tests report on the probability of a certain outcome; if this outcome is defined differently from the outcome of interest in the SVP law, then fit is imperfect. For example, California law requires an assessment of the risk of "predatory" sexual offenses.[203] None of the existing ARA scales limits its outcome measure to "predatory" crimes, and some of the scales may be better at predicting imminent, relatively minor reoffenses rather than the long-term risk of severe crime.[204] Further, as discussed above, some tools, such as the VRAG, measure the risk of violent recidivism including both sexual and non-sexual crimes.[205] Finally, under some SVP laws, the relevant question concerns risk in the short-term "under close supervision" or risk "with treatment," while current static ARA scales, in general, measure long-term stable risk and do not take changeable environmental factors into consideration.

It is important to note at the outset, however, that these fit questions are possible to raise only because of the relative precision and transparency of ARA. The empirical methodology of ARA requires clear specification of the outcomes measured. This specification makes it possible to ask whether the outcomes measured are the same as, or close enough to, the kinds of sexual recidivism that SVP laws aim at. In the clinical method, by contrast, the clinician translates empirical research into risk assessment testimony. The relationship between the outcomes measured in the research, and the outcomes of interest in the courtroom, may be obscured by the opacity of the clinician's expert judgment.

The transparency of ARA exposes problems of fit that are substantial. Admissibility requires a connection between the output of the actuarial tool and the question at issue.[206] An IQ test, for example, would have sufficient fit only if other testimony showed how IQ results could help to assess risk. The results of the Static-99 or the SORAG, on the other hand, which are based on measurements of sexual recidivism, have a much clearer and more immediate fit to SVP

[203] *See* Cooley v. Superior Court of Los Angeles, 57 P.3d 654, 662 (Cal. 2002).

[204] *See* Gabrielle Sjöstedt & Martin Grann, *Risk Assessment: What is Being Predicted by Actuarial Prediction Instruments*?, 1 INT'L J. FORENSIC MENTAL HEALTH 179, 182 (2002). Sjostedt & Grann provide the first clear evidence that prediction within a short time frame may be substantially more accurate than prediction over a long time frame. These authors reported very high predictive accuracy (AUC's of .92 and .94) for the RASOR and the Static-99 when predicting reoffenses within 30 days. This issue is critical, because the statutory mandate calls for prediction of "lifetime" risk.

[205] *See, e.g.*, *In re* Kienitz, 585 N.W.2d 609 (Wis. Ct. App. 1998). *See also In re* Valdez, No. 99-000045CI, at 4 (Fla. Cir. Ct. Aug, 21, 2000) (order granting motion to exclude evidence) (raising a fit concern when it stated: "[N]o evidence was presented to demonstrate that the instruments predicated any specific act or offense contained within the enumerated offenses.").

[206] *See* FED. R. EVID. 401 (defining "Relevant Evidence").

proceedings. Clearly, the less perfect the fit, the more work the proponent of the evidence will have to tie the evidence in to the precise question at issue. Some might argue that the imperfection of the fit could lead to prejudice. We would argue to the contrary, that ARA simply exposes the potential imperfection of fit that is hidden in all risk assessment.

There is an additional question of "fit" that is more difficult and abstract. As described briefly above,[207] the "risk" of recidivism that is estimated by ARA scales is based on aggregate or group data. It is the frequency of recidivists among the subgroup of a validation sample with the same risk score as the subject being evaluated. To put it another way, actuarial assessment tells us the empirically measured rate of recidivism among a group of sex offenders who share a set of characteristics with the subject of the evaluation.

Opponents claim that ARA's group-based information is not relevant to the individual risk assessment required by law. After all, it would hardly be fair to lock a person up merely because he "looked like" others in his reference group (i.e., others who got the same score). As the late Associate Justice Coyne of the Minnesota Supreme Court explained:

> Not only are the statistics concerning the violent behavior of others irrelevant, but it seems to me wrong to confine any person on the basis not of that person's own prior conduct but on the basis of statistical evidence regarding the behavior of other people.[208]

The group-based objection has a broad and a narrow form. The broad form claims that group probabilities are inherently different from predictions of individual behavior. The narrow form acknowledges that all risk assessment is inherently group-based, but complains that ARA is fixed or immutable, classifying people into predefined bins that are too rigid and fail to account for significant individual differences.

Turning first to the broad objection to the group-based nature of ARA, we note that there is a deep philosophical dispute about whether it makes any sense to speak of probability when applied to a single individual as opposed to a group.[209] After all, a given individual, released from prison, either commits another crime (in which case his risk is 100%) or does not (in which case his risk is 0%). Nonetheless, we frequently speak of probabilities that fall in between these two extremes. What would it mean to say that an individual has a 75% risk of reoffending? We assert that this kind of probability can have two different meanings. On the one hand, it could represent empirical information about the frequency of recidivism of a group to which this individual belongs. Alternatively, "75%" may be a way of characterizing our own level of certainty about our forecast about this individual. In either case, the "75%" figure does not refer directly to the individual.

[207] *See supra* Part IV.B.1.

[208] *In re* Linehan ("Linehan I"), 518 N.W.2d 609, 616 (Minn. 1994) (Coyne, J., dissenting).

[209] *See* Janus & Meehl, *supra* note 20, at 36-37.

We need not delve into this philosophical quandary. As a practical matter, it is clear that *all* real-world predictions – at least those that are not simply *guesses*—are based on formal or informal awareness of relevant group behavior. Clinical prediction, at best, is based on perceived commonalities with similarly-situated others – i.e., comparisons to group characteristics and outcomes—ascertained by clinicians in their training and experience. A typical example is *In re Wilson*, in which the court based its commitment order in part on clinical testimony that the defendant was "in a group of offenders whom research studies predict are likely to commit future violent acts . . . ."[210]

In short, being "group-based" does not distinguish ARA from clinical risk assessment, because *all* prediction – including clinical – must be group-based, or at least group-informed; otherwise it would be merely a guess. A clinician who testified that he based his conclusions on experiences with 1,000 prior sex offender evaluations (his reference group) should have vastly more credibility than a clinician who acknowledged that he had never seen a sex offender before evaluating the defendant (no reference group).[211] Clinical prediction clearly relies on reference groups, but they are largely invisible and, at best, vaguely defined. It would be nearly impossible for the clinician who reported 1,000 prior sex offender evaluations to describe for the court the precise nature of that reference group, or, even more importantly, precisely how that reference group guided and informed the evaluation of the defendant.[212] By contrast, ARA uses clearly and readily identified reference groups and equally clearly articulated decision rules about how the conclusions were reached. ARA makes explicit what clinical risk assessment obscures: that prediction and risk assessment are inherently group-based exercises.

It is tempting to respond to the "group-based" objection by pointing out that such group-based assessments of risk are ubiquitous in modern life, and that we act on such group-based assessments of risk in variety of consequential settings.[213] Imagine a 60-year-old, obese chain-smoking man with a family history of heart disease visiting an internist for an annual checkup, and the internist saying to him: "You know, group data say that you're at awfully high risk for a heart attack, but that's group data and you're an individual, so we won't worry about it." If the internist failed to warn the patient of his high-risk status, the internist could easily be found negligent. The advisory from the internist would be based on ample research documenting clear links between the aforementioned characteristics of the patient (i.e., high risk factors) and a high probability outcome (heart attack). Needless-to-say, insurance companies rely on these

[210] *In re* Wilson, No. C3-00-434, 2000 WL 1182807, at *5 (Minn. Ct. App. Aug. 22, 2000).

[211] *See In re* Coffel, No. ED 79989, 2003 WL 716682, *12 (Mo. Ct. App. Mar. 4, 2003) (rejecting expert testimony on the grounds that expert had never diagnosed or treated similar offenders).

[212] *See id.* at **11* (discounting testimony of expert who had interviewed similar offenders, but "did not know whether any of the women she interviewed …had reoffended. She had no idea whether any of the characteristics of female sex offenders she identified from her interviews had anything to do with the likelihood of reoffense.").

[213] *See* Grove & Meehl, *supra* note 67, at 305.

statistically-determined risks to calculate premiums for all forms of coverage. The principle is always the same: In many important areas of our lives, we deduce individual risk from group risk.

But this argument fails to address the full force of the group-based objection. Insurance companies use actuarial tables, because it enables them to make a profit. Although it might be irrational to do so, the high-risk heart patient retains the right to reject or ignore his doctor's group-based advice. Although these examples involve important and consequential decisions, they do *not* involve the long-term, comprehensive and involuntary deprivation of liberty.

In our judgment, the morality of depriving people of long-term liberty based on predictions of future crimes is questionable, in significant part because all prediction is ultimately based on group membership. If we were discussing the wisdom of SVP laws, this would be a powerful argument in opposition.[214] The question we address, however, is different. Given the existence of SVP laws, and their routine use of clinical prediction, the fact that ARA is group-based provides no basis for rejecting its use, because all prediction is group-based.

The narrow formulation of the "group-based" objection also carries considerable force. All prediction necessarily treats individuals as abstractions, isolating "essential" features that are similar to the "essential" features of the group. But critics argue that ARA is especially defective, because the predictive scales are limited to a few, pre-determined items. Thus, while a clinician's expertise presumably allows her to choose the factors that she deems to be most salient for the individual (i.e., to mentally construct the most relevant reference group), ARA rigidly restricts its assessment to the pre-set factors. CRA, this objection goes, may necessarily categorize individuals, but at least it uses the most salient factors about the individual to construct its categories. In contrast, the pre-selected risk predictors of ARA may fail to account for some significant fact about the individual.

As an example of this type of objection, in *In re Valdez*, a Florida trial court points out that none of the ARA tests "seem to include whether the person has been or is being treated, whether he has been or still is incarcerated, is under house arrest, or is comatose, although to the unsophisticated, one or more of those factors would seem to bear heavily on future conduct."[215] Other commentators criticize ARA scales on the grounds that they give no, or too little, weight to dynamic factors, such as treatment-response and post-confinement supervision.[216]

[214] One of us has raised this, among other, objections to SVP laws. *See, e.g.*, Eric S. Janus, *Preventing Sexual Violence: Setting Principled Constitutional Boundaries on Sex Offender Commitments*, 72 IND. L.J. 157 (1996).

[215] *In re* Valdez, No. 99-000045CI, at 6 (Fla. Cir. Ct. Aug, 21, 2000) (order granting motion to exclude evidence).

[216] *See generally* Robert T. Schopp et al., *Expert Testimony and Professional Judgment: Psychological Expertise and Commitment as a Sexual Predator After Hendricks*, 5 PSYCHOL. PUB. POL'Y & L. 120, 137 (1999) (noting that research has shown a positive correlation among factors predicting recidivism and those predicting treatment-response); Harris et al., *Violent Recidivism*, *supra* note 124, at 332-33 (recommending reliance on dynamic factors such as treatment

These putative defects may seriously undermine the assessment power of ARA tools. Yet, despite these defects, ARA retains its general superiority to clinical judgment. Although excluding ARA, while admitting CRA, is incoherent, it is *not* incoherent to take flaws and limitations into account in evaluating the risk assessment testimony as a whole, and in judging whether it is legally sufficient to support massive curtailment of liberty.

We think that there are three basic approaches for dealing with the "fit" objection to ARA. First, there are some, relatively rare, circumstances in which ARA should be disregarded in favor of clinical judgments.[217] Grove and Meehl argued, however, that clinicians are bad at identifying the presence of characteristics that may justify the prepotency of clinical judgment. Grove and Meehl asserted that clinical judgments are superior to the actuarial tests only in those rare situations in which the trumping characteristic is objectively ascertainable and clearly linked with the predicted outcome.[218] Some of the factors noted by the *Valdez* court (e.g., being comatose or incarcerated),[219] would clearly qualify as trumping factors.

Second, some commentators and practitioners advocate the "adjustment" of ARA scores to account for individualized risk factors.[220] Under this method, the examiner adds or subtracts percentage points from the ARA results to reflect risk factors that (in the examiner's judgment) are not adequately reflected in the ARA result. Most commentators believe, however, that this form of "adjustment" transforms ARA into CRA, depriving ARA of its advantage over clinical methods.[221] There is, in our judgment, a special problem that arises with such adjustment --an undeserved veneer of science to what is essentially a clinical judgment.[222]

We dub the third way of approaching the fit problem for ARA the "weight" method. This method recognizes that sometimes ARA simply does not answer the precise question asked by the SVP court. The proper approach is to recognize that the actuarial information is relevant to, but not dispositive of, the

response); Tony Ward & Lynne Eccleston, *The Assessment of Dangerous Behavior: Research and Clinical Issues*, 17 BEHAV. CHANGE 53, 56-57 (2000).

[217] *See* Grove & Meehl, *supra* note 66.

[218] *Id.*

[219] *In re* Valdez, No. 99-000045CI, at 6 (Fla. Cir. Ct. Aug, 21, 2000) (order granting motion to exclude evidence)

[220] *See, e.g.*, Cooley, 57 P.3d at 660 (relying on the following expert testimony that made use of this 'adjusted actuarial method:' "[u]sing the Static-99 test and '[a] dash of clinical judgment,' [the expert witness] estimated Marentez's likelihood of reoffense over a 15-year period at '52 to 55, 57 [percent], something like that.'").

[221] Randy K. Otto & John Petrila, *Admissibility of Testimony Based on Actuarial Scales in Sex Offender Commitments: A Reply to Doren*, 3 SEX OFFENDER L. REP. 1, 15 (2002).

[222] *See id.* (arguing that "it is inappropriate and logically inconsistent to use research on the superior accuracy of actuarial methods to support clinical use of an adjusted actuarial approach, which essentially is a clinical approach.").

legal question. The lack of precise fit is accounted for in the reduced weight given to the ARA information, but not in a "modification" of that information.

Let us address the fit issue somewhat more concretely. Recall that risk may be analyzed in independent subfactors, including probability, imminence, and severity.[223] In their present state, actuarial scales address only probability, and, in fact, are limited to relatively long-term probability, based primarily on static factors. Questions of imminence – i.e., "when" the risk is high or low – are not presently well addressed in ARA instruments, because these tools do not address such important factors as the impact of treatment and available community supervision.[224]

There is good evidence that both treatment,[225] and optimally designed, multi-disciplinary, team approaches to community supervision can reduce recidivism.[226] Some research suggests that actuarial instruments that consider such dynamic variables perform better than instruments that do not.[227] At this point, however, the development of dynamic risk assessment scales that take into consideration such issues as treatment and supervision are strictly in the experimental stage.[228] Thus, the results from most actuarial risk assessment scales

[223] See *supra* text accompanying note 33.

[224] *See* Ward & Eccleston, *supra* note 216 at 63 (distinguishing among assessments for long-term, short-term, and imminent violence; suggesting that long-term assessments may use "group" estimates, but that short-term assessments need to look at the interaction of long term disposition and "short-term triggers").

[225] R. Karl Hanson et al., *First Report of the Collaborative Outcome Data Project on the Effectiveness of Psychological Treatment for Sexual Offenders*, 14 SEXUAL ABUSE: J. RES. & TREATMENT 169 (2002) (reporting on meta-analysis which found that the "relative" reduction in recidivism associated with treatment completion was 40%. The authors defined the "relative reduction" rate as the difference between the treatment and non-treatment rates expressed as a percentage of the non-treatment rate).

[226] *See* ROBERT A. PRENTKY & ANN W. BURGESS, FORENSIC MANAGEMENT OF SEXUAL OFFENDERS 236, 243 (2000) ("[T]he most effective known technique for reducing risk of relapse is intensive supervision" in the community; community "aftercare can be made sufficiently 'tight' to reduce risk to a minimum for many offenders."); Kim English, *The Containment Approach: An Aggressive Strategy for the Community Management of Adult Sex Offenders*, 4 PSYCHOL. PUBL. POL'Y & LAW 218, 219 (reporting on supervision methods that "can exert significant control over offenders' opportunities to commit new crimes").

[227] *See* Anthony Beech et al., *The Relationship Between Static and Dynamic Risk Factors and Reconviction in a Sample of U.K. Child Abusers*, 14 SEXUAL ABUSE: J. RES. & TREATMENT 155 (2002); Rebecca Dempster & Stephen D. Hart, *The Relative Utility of Fixed and Variable Risk Factors in Discriminating Sexual Recidivists and Non-Recidivists*, 14 SEXUAL ABUSE: J. RES. & TREATMENT 121 (2002).

[228] R. Karl Hanson & Andrew J.R. Harris, *A Structured Approach to Evaluating Change Among Sexual Offenders*; 13 SEXUAL ABUSE: J. RES. & TREATMENT 105 (2001); David Thornton, *Constructing and Testing a Framework for Dynamic Risk Assessment*, 14 SEXUAL ABUSE: J. RES. & TREATMENT 139 (2002).

must be interpreted as reporting risk without consideration of treatment[229] or state-of-the-art supervision.

Risk that takes into account treatment and intense supervision may well be the relevant question posed in SVP cases.[230] At this stage, there are no actuarial scales that address the specific question of *controlled* risk, and courts may have to rely on clinical judgments.[231] In these cases, ARA scores are still not irrelevant. ARA risk scores should be interpreted to reflect long-term risk under conditions of unknown supervision.[232] Dvoskin and Heilbrun argue that the lack of actuarial scales addressing the question of controlled risk does not support the modification or replacement of ARA with clinical judgment, but rather require the careful articulation of the "strengths and limitations" of ARA, and the placing of it into context.[233]

D. General Acceptance

As we have noted, "general acceptance" is at the heart of the original *Frye* test, although under *Daubert*, and in many *Frye* jurisdictions, "general acceptance" has become but a part of a broader "reliability" inquiry.[234] Thus, while SVP courts have sought to examine the degree of acceptance of ARA among mental health professionals, they have done so in a broader context in which they have examined other, more immediate indicia of reliability.[235] We

[229] *See* Andrew Harris, et al., Dep't Solicitor Gen. Can., *Static-99 Coding Rules: Revised – 2003* (indicating that the "original samples and the recidivism estimates should be considered primarily as "untreated"), *available at* http://www.sgc.gc.ca/corrections/publications_e.asp (last visited Nov. 18, 2003).

[230] *See, e.g.*, Cooley, 57 P.3d at 671 (noting that relevant question under California SVP Act is whether "the person presents a substantial danger of reoffense if released without conditions, or whether instead he is safe only if restrained, supervised, and treated involuntarily [in] custody").

[231] *See, e.g.*, Sjöstedt & Grann, *supra* note 204, at 180 (noting that "well-informed management interventions depend also upon information about the imminence, nature, frequency, and severity of the outcome in question," information that is not readily available through ARA); Joel A. Dvoskin & Kirk Heilbrun, *Risk Assessment and Release Decision-making: Toward Resolving the Great Debate*, 29 J. AM. ACAD. PSYCH. LAW 6, 7 (2001); La Fond & Winick, *supra* note 19, at 316 (describing need for graduated levels of freedom as a means of judging "safety" among sex offenders).

[232] *See* Hanson & Harris, *supra* note 228 (noting that the "best method of incorporating this information is unknown, but some adjustment seems justified when the offenders' dynamic needs are substantially higher (or lower) than would be expected from their scores on established actuarial measures).

[233] Dvoskin & Heilbrun, *supra* note 231, at 9. *But cf.* Dennis Doren, Evidentiary Issues, Actuarial Scales, and Sex Offender Commitments 10 (2001)(unpublished manuscript, on file with the American Criminal Law Review) (noting that "adjusting people's estimated risk downward from what the actuarial instruments indicate based on the successful completion of meaningful sexual offender treatment programming is empirically supported").

[234] *See, e.g.*, Knapp, *supra* note 104, at 1001.

[235] *See, e.g.*, People v. Taylor, 782 N.E.2d 920, 931 (Ill. App. Ct. 2002) (noting that the Illinois general acceptance test is not concerned with the number of experts who endorse the method, but

suggest that questions of general acceptance are fundamentally indeterminate, and really cannot be decided without some implicit or explicit reference to reliability, and, relatedly, prejudice.

The general acceptance test entails four discrete determinations: first, courts must determine what scientific principle is involved; second, whether the principle or some aspect of the principle is novel; third, what the relevant scientific community is; and fourth, whether that community accepts the principle.[236] Each of these questions can be asked and answered at various levels of generality, with results dependent on the framing of the question. For example, the scientific principle underlying an expert's ARA testimony might be very generally framed (incorporating the general principles of actuarial risk assessment) or very specifically framed (focused on the specific instrument relied on by the expert). Similarly, some advocates argue that the relevant community consists of mental health professionals evaluating sex offender commitment defendants, while others argue that the community must be broader.

Initially, we can examine whether ARA is considered to be "novel" science, thus triggering *Frye* scrutiny. As we have noted, many (but not all) [237] *Frye*-jurisdiction courts have simply admitted clinical dangerousness testimony without any sort of vetting to insure its scientific *bona fides*.[238] The question here is whether, and how, the addition of actuarially-derived information changes the courts' posture on the test for admissibility.

Courts have taken four positions. Some courts have taken the position that ARA and clinical risk assessment should be analyzed separately for admissibility purposes. Several trial courts, for example, have excluded ARA on the grounds that it is novel science that has not been generally accepted, but have made clear that clinical testimony will be fully allowed.[239]

A second group sidesteps the question of ARA novelty by characterizing the use of actuarially-derived information as just another element of the clinical judgment. On this view, the overall judgment of the expert escapes gatekeeping scrutiny, because the ARA has taken on the character of the clinical assessment.[240] This approach assumes that the expert has "supplemented" the

that it is concerned with the exclusion of "methods new to science that undeservedly create a perception of certainty when the basis for the evidence or opinion is actually invalid.") (quoting Donaldson v. Cent. Ill. Pub. Serv. Co., 767 N.E.2d 314, 324 (Ill. 2002))); *see also supra* notes 108-109 and accompanying text.

[236] *See* Knapp, *supra* note 110, at 1017-18 (discussing how a court determines when science is generally accepted); Peter B. Oh, *The Proper Test for Assessing the Admissibility of Nonscientific Expert Evidence Under Federal Rule Of Evidence* 702, 45 CLEV. ST. L.REV. 437, 441 (1997) (discussing the introduction of nonscientific expert evidence).

[237] *See In re* Coffel, No. ED 79989, 2003 WL 716682, at *12 (Mo. App. E.D. Mar. 4, 2003) (finding that expert testimony regarding clinical risk assessment must be based on scientific principles "generally accepted in the relevant scientific community.").

[238] *See* cases cited *supra* note 118.

[239] *See* cases cited *supra* note 90.

[240] *See* Arizona *ex rel.* Romley v. Fields, 35 P.3d 82, 89 (Ariz. Ct. App. 2001) ("[U]se of actuarial models by mental health experts to help predict a person's likelihood of recidivism is not the kind

actuarial information with clinical judgment, and suggests that actuarially-derived information used alone might be "novel" and therefore be subject to *Frye* analysis.

Yet a third approach holds, to the contrary, that supplementing a clinical evaluation with ARA undercuts the exemption of "pure clinical" testimony, and subjects the entire judgment of the expert to *Frye*.[241] A fourth approach, apparently required in federal courts as a result of *Kumho*, and adopted in at least one SVP setting,[242] would subject all expert risk assessment – clinical, actuarial and mixed -- to admissibility vetting.

In our judgment, the first approach –subjecting ARA, but not CRA, to admissibility testing – is incoherent because it allows for the possibility that the more reliable ARA would be treated more strictly than the less reliable CRA. For similar reasons, we favor the fourth (*Kumho*) approach, because it treats all forms of risk assessment similarly. However, in view of the ensconced nature of clinical testimony, we doubt that many courts will fully adopt this proposal.[243]

The two middle positions judge the "novelty" of ARA by its relationship to a larger clinical evaluation. These two positions seem to conflate the question of novelty with the potential for prejudice. Both appear to assume that CRA is clean of prejudice, while ARA (given its scientific tenor) engenders prejudice. They disagree about whether CRA "cleanses" ARA, or ARA "pollutes" CRA. Since we take the position (discussed below) that CRA has as much risk of prejudice as ARA, we think that the underlying premise of both positions is wrong. Both analyses have some danger of seeding confusion, so we address each briefly.

The second position has some merit, but verges on a serious misinterpretation. It is possible to understand this position as allowing ARA only to the extent that it has been modified or adjusted by clinical judgment. But as we have noted, commentators believe that such adjustment, at least when done routinely, destroys the advantage that ARA otherwise might achieve.[244] It is, in our estimation, imperative to make a fundamental distinction between

of novel scientific evidence or process to which *Frye* applies."); People v. Ward, 83 Cal. Rptr. 2d 828, 831 (Cal. Ct. App. 1999) ("[T]he testimony of a psychologist who assesses whether a criminal defendant displays signs of deviance or abnormality is not subject to *Kelly-Frye*."); State v. Holtz, 653 N.W.2d 613, 619-20 (Iowa Ct. App. 2002) (stating that ARA should be deemed reliable only when used "in conjunction with a full clinical evaluation"). *But see In re* R.S., 773 A.2d 72, 92 (N.J. Super. 2001), *aff'd*, 801 A.2d 219 (2002) ("[T]he State has established that the actuarial instruments are reliable tools for help in predicting a sex offender's risk of reoffense."); State v. Strauss, 20 P.3d 1022, 1025 (Wash. Ct. App. 2001) ("Scientific literature and secondary legal authority also supports the view that the relevant scientific community generally accepts the [MnSOST, RRASOR and VRAG] as part of an overall risk assessment.").

[241] *See* Taylor, 782 N.E.2d 920 at 930 (2002).

[242] *See In re* Coffel, No. ED 79989, 2003 WL 716682, at *11-12 (holding that psychologist's opinion was inadmissible because the factors she relied upon were solely a product of her "clinical expertise" and were not based on any scientific research or principles generally accepted in the psychological community).

[243] *But see id.*; *cf.* Slobogin, *supra* note 119.

[244] *See, e.g.*, Otto & Petrila, *supra* note 221.

modifications of conclusions deriving from ARA results, which are advisable when warranted, and modifications of the ARA score itself, which are impermissible.

As we have pointed out, ARA cannot be properly used without judgment and interpretation. Expertise is required to properly understand the limits of ARA (and CRA), weigh the significance of its strengths and weaknesses, and judge its relevance to the precise legal question that is at issue (the question of "fit" that we addressed above). In a sense, this context might be said to constitute clinical judgment, and so in this sense, ARA should be embedded in a fuller clinical evaluation. But it is not so much that the clinical judgment modifies the ARA; rather, the clinical information provides context for the ARA to allow it to be properly interpreted, provides information that shows its relevance to the legal questions at issue, and fills in proof that is not addressed by the ARA.[245]

The third variation – the use of ARA in what is otherwise clinical testimony consequently subjects the testimony as a whole to *Frye* scrutiny – seems to us to have a perverse logic that may have undesirable consequences. Most directly, this rule might create a strong disincentive to the use of ARA by experts, who would find their "pure" clinical opinion accepted more readily than a more complete evaluation that made reference to actuarial methods. The rule is perverse because it appears to assume that clinical assessment is free of the problems that limit ARA, whereas, as we have shown, it is likely that CRA suffers from the same problems, even to a greater degree, but that the problems are less visible. Thus, the rule encourages less, rather than more, accuracy.

Turning from the threshold question of "novelty" to the core *Frye* concern of "general acceptance," we begin by noting that there is, to be sure, vehement disagreement about whether actuarial methods in general, or specific instruments in particular, are well-enough developed to be used in the liberty-deprivation context of SVP cases. Respected researchers urge the "complete replacement of existing practice with actuarial methods,"[246] and suggest that the use of clinical methods, where actuarial ones are available, would be "unethical." [247] Yet, other scholars conclude that "even the best studied and validated actuarial tool for assessing dangerousness . . . has not been demonstrated as suitable for practical purposes in many instances, or to be superior to clinical assessments."[248] Nor is there any solid evidence about the degree of acceptance among experts of actuarial methods. Anecdotal evidence reported in appellate cases suggests that the usage rate is fairly high, at least in SVP cases,[249] but others report that "most

[245] *See* discussion *supra* accompanying notes 220-233.

[246] Quinsey et al., *supra* note 124, at 171; *see also* Litwack, *supra* note 89, at 409.

[247] Grove & Meehl, *supra* note 66, at 320 ("To use the less efficient of two prediction procedures in dealing with . . . matters [such as "high stakes" predictions] is not only unscientific and irrational, it is unethical.").

[248] Litwack, *supra* note 89, at 410.

[249] *See, e.g.*, Taylor, 782 N.E.2d at 931 (noting, without citing source, that "many psychologists and psychiatrists utilize these instruments to predict whether a sexual offender is likely to reoffend"); *see also In re* Strauss, 20 P.3d at 1025 (citing testimony that "psychologists who are

professionals continue to use a subjective, clinical judgment approach when making predictive decisions."[250] Some scholars conclude that "there currently are no widely accepted professional standards or guidelines regarding what constitutes the most appropriate approach to conducting sex offender risk assessments."[251]

Given the rather indeterminate outcome of this professional "show of hands," we suggest that the *Frye* analysis ought to focus somewhat more carefully on what, exactly, is "novel" about ARA, and how that novelty bears on the twin concerns in the admissibility arena: reliability and prejudice. Actuarial and clinical methods share a great deal; the novelty of ARA consists mainly in its dual claims to science and predictive superiority.

First, the commonality: Both clinical and actuarial methods claim to identify relevant information about the subject and both *combine* the risk factors, giving potentially varying weights to different factors. They differ in that ARA attempts to determine empirically what factors to select and how to weigh them, whereas CRA relies on the case-by-case judgment of the examiner to make these decisions.[252] Thus, ARA is "nomothetic"[253] – rule-based – whereas CRA claims to treat each case individually. Further, ARA claims to measure and quantify its results empirically and statistically, whereas CRA relies on the experience and credentials of the expert for its *bona fides*.[254] Finally, ARA claims to be predictively "superior" to CRA.

These three "novel" aspects of ARA are interrelated. The empiricism of ARA dictates its rule-based character. The claim of superiority is precisely that

knowledgeable about assessing the risk of recidivism among sex offenders generally accept the actuarial instruments used in this case"); *In re* R.S., 773 A.2d at 80 (citing testimony that of the 150-175 experts in the field of sex offender risk assessment nationwide, "most employ the clinically-adjusted actuarial assessment method").

[250] Grove & Meehl, *supra* note 66, at 299 (listing 17 reasons that might explain why clinicians shy away from the use of ARA and theorizing that the chief among the possible reasons may be the clinicians' concern that ARA may undermine or detract from their professional expertise, or, worse yet, replace them.)

[251] Laura S. Guy & John F. Edens, *Juror Decision-making in a Mock Sexually Violent Predator Trial: Gender Differences in the Impact of Divergent Types of Expert Testimony*, 21 BEHAV. SCI. L. 215, 217 (2003); *See also* Heilbrun, et al., *supra* note 58, at 398.

[252] For example, a Wisconsin court described the clinical assessment of an expert as follows:

> [The expert] explained that he made a clinical judgment as to which factors were the most important, because not all the factors were equally strong as predictors. [He] testified that he weighed the predictive factors [the defendant] possessed against those he did not posses[s] and, in [his] opinion, the former far outweighed the latter.

In re Kienitz, 585 N.W.2d 609, 613 (Wis. Ct. App. 1998).

[253] *See* MEEHL, *supra* note 65.

[254] *See, e.g.*, State v. Keith, 573 N.W.2d 888, 896 (Wis. Ct. App. 1997) (holding that lower court's refusal to allow defense counsel to ask the prosecution expert "whether his past predictions of future dangerousness had been tested for accuracy and validity" on the grounds that such inquiry would be irrelevant was proper).

the trade-off between rules and individuation tips in favor of rules: rule-based judgments do better than ideographic – or individuated – judgments.

We have already reviewed the science underlying the validation of ARA, and the claim of actuarial superiority.[255] The principle of actuarial superiority is not novel. Meehl first proposed it in 1954.[256] It has been tested extensively, and has broad acceptance in the literature, both in general,[257] and in the specific literature concerning sexual offending.[258] Similarly, the science underlying ARA is not new. Statistical decision theory[259] and its application to human judgment[260] have been around for fifty years. The same methodology has been applied in numerous, diverse contexts, including weather forecasting, law school admissions, disability determinations, predicting the quality of the vintage for red Bordeaux wines, and predicting the quality of sound in opera houses.[261] Similarly, the statistics that assess the accuracy of ARA, and allow for the comparison of various ARA scales (e.g., the AUC values)[262] also have a lengthy pedigree.[263]

E. Prejudice

Concern about the potential prejudicial impact of ARA is, arguably, the most coherent of the possible grounds for excluding ARA testimony, and is most clearly at the center of the admissibility issue.[264] We have argued that the comparative reliability of ARA makes its exclusion illogical in a system that admits clinical assessments. However, if ARA is materially *more* prejudicial than clinical assessments, it might not be illogical to exclude the former while admitting the latter.

[255] *See supra* Part IV.B.

[256] *See* MEEHL, *supra* note 65, at 119 (conducting a review of existing empirical studies comparing the efficacy of clinical and actuarial predictions and concluding that in all but one the actuarial methods yielded superior results).

[257] *See* Grove & Meehl, *supra* note 66.

[258] *See* Hanson & Bussiere, *supra* note 75, at 349. *See generally* D.L. FAIGMAN ET AL., MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (1999).

[259] ABRAHAM WALD, STATISTICAL DECISION FUNCTIONS V (1950).

[260] W. P. Tanner & John A. Swets, *A Decision-making Theory of Visual Detection*, 61 PSYCHOL. REV. 401 (1954).

[261] John A. Swets et al., *Psychological Science Can Improve Diagnostic Decisions*, 1 PSYCHOL. SCI. PUB. INTEREST 1 (2000).

[262] *See* discussion *supra* Part V.B.1.

[263] "AUC" is a statistical procedure that is derived from "signal detection theory." A review of the applications of signal detection theory, conducted over thirty-five years ago, revealed approximately 1,000 articles in perceptual and cognitive psychology and several dozen in clinical and abnormal psychology. *See, e.g.*, DAVID M. GREEN & JOHN A. SWETS, SIGNAL DETECTION THEORY AND PSYCHYPHYSICS, (1966).

[264] *See* Taylor, 782 N.E.2d at 931("The *Frye* rule is meant to exclude methods new to science that undeservedly create a perception of certainty when the basis for the evidence or opinion is actually invalid." (quoting Donaldson, 767 N.E. 2d at 324)).

Judges serve as gatekeepers on expert testimony, in part, because it is assumed that juries cannot be counted on to determine the proper weight to be given to such testimony.[265] There are three potential sources of prejudice from ARA testimony. First, there is concern that the scientific and statistical nature of actuarial assessments will unduly influence the fact-finder into giving it more weight and credibility than it deserves, and that the principle of "actuarial superiority" will exacerbate this tendency.[266] The corollary is that the weaknesses of some ARA instruments are too complex for lay fact-finders to apprehend.[267] Second, some worry that juries will ignore the lack of "fit" between the actuarially derived risk and the legally relevant risk, thus giving ARA too much weight.[268] Third, in the words of Professor Tribe, the "incriminating significance"[269] of statistical probabilities is obscure.

We acknowledge and address the potential danger from ARA testimony. Indeed, the cases are salted with examples of courts misciting and misinterpreting actuarial information.[270] Our response is three-fold. First, we join others[271] in

[265] State *ex rel.* Romley v. Fields, 35 P.3d 82, 86 (Ariz. Ct. App. 2001) ("Because 'science' is often accepted in our society as synonymous with truth, there was a substantial risk of overweighting by the jury. The rules concerning scientific evidence appear to have been aimed at that risk").

[266] One trial court relied upon the following statement by Dr. Thomas Grisso in support of its decision to exclude expert testimony based upon ARA estimates:

> Without adequate caveats, the presentation of actuarial data is seductive and particularly prone to mislead. Courts and juries are likely to be impressed not only with the <u>appearance</u> of precision inherent in the use of numbers, but also with the fact that assessment theorists consistently assert that actuarial methods are superior to clinical judgment. In the absence of testimony about the limits of the tools, the potential is great for their results to be given far more weight than the instruments can support.

In re Johnson, No. LACV038974, at 13 (Story County, Iowa, July 13, 2000) (order excluding expert testimony) (quoting Dr. Thomas Grisso).

[267] Dr. Lynn Maskel's testimony in the hearing to exclude expert testimony in *In re Johnson* is representative of this concern:

> [A] lot of the testimony that would rebut the instruments that would poke holes in its reliability and its validity and its soundness is extremely difficult to understand, and I think most members of the jury would not be able to understand the attack on the instruments."

Id. (quoting Dr. Maskel).

[268] *See* discussion *supra* Part IV.C.

[269] Lawrence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329, 1355 (1971).

[270] *See, e.g.*, Cooley v. Superior Court of Los Angeles, 57 P.3d 654, 660 (Cal. 2002) (suggesting that ARA instruments are "given" to a defendant, rather than scored by the evaluator; quoting an evaluator as modifying the Static-99 score with a "dash of clinical judgment," to estimate the defendant's "likelihood of reoffense over a 15-year period at '52 to 55, 57 [percent], something like that.'"); *In re* Seeboth, No. C037185, 2002 WL 31888038, at *8 (Cal. Ct. App. Dec. 30, 2002)

noting that clinical risk assessment testimony carries its own risk of prejudice. Second, we suggest that the concerns raised by Tribe about trial by mathematics are not salient in this context. Finally, we argue that the science of ARA brings a certain transparency that not only has the potential to reduce the misuse of ARA, but also to bring some much needed accountability to the entire process of risk assessment. However, we argue that this increased accountability depends on courts adhering to a set of recommendations that we set out in the final part of the article.

As we have noted, clinical judgment carries significant risks of error, but, for the most part, these errors are rendered invisible by the opaque nature of the clinical judgment process. For example, Borum et al. identify a set of common errors in clinical judgments, including "confirmatory bias," which the authors define as the

> [T]endency to look for evidence that supports one's hypothesis . . . and to ignore, or fail to seek, information that is not consistent with that hypothesis. This bias could also extend into the interpretation phase, where one may interpret the same piece of data in the way that supports one's perceptions /preconceptions when either of two interpretations is equally possible.[272]

Because the critical steps in a clinical assessment occur in the clinician's head, such errors may be difficult to expose through the normal advocacy tools of cross-examination. Where there is a dispute between clinicians, the fact-finder is left with a simple credibility judgment, in which the fears of sexual violence create a strong bias in favor of assessments that are more protective of public safety.[273]

(mis-identifying the RRASOR as the RAZOR); Taylor*,* 792 N.E.2d at 925 (reporting on the mistaken and ambiguous statement of an expert that "the author of the instrument had reported a .71 receiver operator characteristic (ROC). This meant that 71 times out of 100, the Static-99 instrument would correctly identify an individual as a recidivist, while 29 times out of 100 the instrument would incorrectly identify an individual as a recidivist."). In *State v. Oberacker*, the court reported that:

> [T]he Static 99 . . . was only 33% accurate; so 66% of the predictions of Static 99 would be incorrect. . . . The Minnesota Sexual Offender Screening Tool, which Oberacker scored a high relapse quotient on, is a little more accurate with an accuracy rate of between 45% to 50%; so approximately one-half the time it is inaccurate in its predictions.

No. 81093, 2003 WL 125277, at *4 (Ohio Ct. App. Jan. 16, 2003).

[271] *See, e.g.*, Faigman et al., *supra* note 42.

[272] Randy Borum et al., *Improving Clinical Judgment and Decision Making in Forensic Evaluation*, 21 J. PSYCHIATRY & L. 35, 47 (1993). For a similar exposition of judgment errors see Hal R. Arkes, *Principles in Judgment / Decision Making Research Pertinent to Legal Proceedings*, 7 BEHAV. SCI. & L. 429 1989).

[273] Saleem A. Shah has commented that:

In addition to the inaccuracies inherent in clinical assessments of risk, communications about clinical assessment entail potentially serious ambiguities. Clinicians can use either “categorical” descriptors (e.g., “likely to reoffend”), or quantitative terminology (e.g., "52 to 55, 57 [percent], something like that.").[274] In the first case, where clinicians use categorical descriptors, the “cut-off scores” are chosen by the clinician and are hidden from the decision-maker: “[u]ncertainty is ignored and, in effect, the forecaster ‘becomes’ the decision maker, a role for which he/she generally is not well-equipped.”[275] Further, the use and meaning of these categories shows substantial variability.[276] On the other hand, the mere use of probabilistic (quantitative) terms by the clinician does not solve the problem, since research suggests that “probability was not represented consistently and quantitatively in the clinicians’ minds.”[277]

That clinical assessment has its own potential to mislead does not, of course, provide a justification for adding more potential prejudice in the form of actuarial risk assessment. We argue, to the contrary, that the transparency of ARA facilitates more effective trial advocacy, and gives attorneys and courts the tools to bring more accountability to the entire risk assessment process. Two decades ago John Monahan gave a straightforward prescription for improving forensic risk assessment:

> What is necessary . . . is a dramatic increase in the degree to which mental health professionals articulate what it is they are predicting and how they went about predicting it. This involves explicitly enumerating the kinds of acts one takes to be violent, frankly

> [P]owerful social contingencies are associated with . . . [false negative] errors . . . decision-makers will tend to be more careful about avoiding “false negative” errors, *viz*., releasing persons who may later commit some serious or violent crimes, than about “false positive” errors, *viz*., retaining persons who may not be likely to commit serious crimes if released.”

Saleem A. Shah, *Legal and Mental Health System Interactions*, 4 Int’l J.L. & Psychiatry 219, 238 (1981); s*ee also* Thomas R. Litwack & Louis B. Schlesinger, *Assessing and Predicting Violence: Research, Law, and Applications*, *in* Handbook of Forensic Psychology 233 (Weiner & Hess eds., 1987) (“[O]verprediction [of violence] may occur because mental health professionals are extremely fearful of the grave consequences of a false-negative prediction and are determined to be cautious.”).

[274] *Cooley*, 57 P.3d at 660.

[275] John Monahan & Henry J. Steadman, *Violent Storms and Violent People: How Meteorology Can Inform Risk Communication in Mental Health Law*, 51 Am. Psychol. 931, 934 (1996).

[276] *See, e.g.*, Frederick Mosteller & Cleo Youtz, *Quantifying Probabilistic Expressions,* 5 Stat. Sci. 2 (1990); Fenna H. Poletiek, *How Psychiatrists and Judges Assess the Dangerousness of Persons with Mental Illness: An ‘Expertise Bias,’* 20 Behav. Sci. & L. 19, 20 (2002) (reporting that Dutch judges and doctors ascribe different meanings to the term “dangerous” in a civil commitment context); Robert Timothy Reagan, Frederick Mosteller & Cleo Youtz, *Quantitative Meanings of Verbal Probability Expressions*, 74 J. Applied Psychol. 433, 440 (1989) (reporting on variable meanings given to expressions like “likely” and “very likely”).

[277] Monahan & Steadman, *supra* note 275, at 935 n.2.

> stating the factors on which the prediction is based, and being clear on the likelihood with which it is believed they will occur.[278]

Actuarial risk assessment lays out exactly these factors, enabling courts and advocates more clearly to evaluate the reliability and relevancy of the resultant assessments.

There remains to be discussed the potential for the scientific nature of ARA to disable factfinders from fair assessment of its shortcomings. Conventional wisdom holds that "lay jurors are incompetent to evaluate scientific proof critically,"[279] and, perhaps to a less articulated extent, that judges cannot properly evaluate scientific evidence.[280]

To begin, we address the concern raised by Prof. Tribe, denouncing, in his customary eloquence, what he referred to as "trial by mathematics."[281] He maintained that there is an "inherent conflict" between the goals of jurisprudence, particularly with respect to fact-finding, and the uncompromising rationality and objectivity of mathematics.[282]

Warning against the threatened loss of the "presumption of innocence,"[283] Tribe recounted an example in which a palm print found on a murder weapon was sufficiently similar to that of the defendant that an expert could testify that such prints are found in no more than one case in a thousand.[284] The question, as Tribe noted, is "how the jury might best be informed of the precise incriminating significance of that finding."[285] Less than thirty years after Tribe's thesis, we have moved from palm prints and probabilities of 1:1,000 to DNA typing and probabilities that range from 1:500,000 to 1:738,000,000,000,000.[286] Even at the "lower end" of this range, how does a juror maintain a "presumption of innocence" while processing that the probability of a chance match between a DNA sample and the defendant is about one-in-a-million? Or, in Tribe's terms,

[278] MONAHAN, *supra* note 84.

[279] Edward J. Imwinkelried, *Judge Versus Jury: Who Should Decide Questions of Preliminary Facts Conditioning the Admissibility of Scientific Evidence?*, 25 WM. & MARY L. REV. 577, 580 (1984).

[280] *See, e.g.*, Sophia I. Gatowski et al., *Asking the Gatekeepers: A National Survey of Judges on Judging Expert Evidence in a Post-*Daubert *World,* 25 LAW & HUM. BEHAV. 433, 453 (2001) (questioning the "ability of the courts, particularly the state trial courts, to assess the scientific reliability and validity of proffered scientific evidence").

[281] *See generally* Tribe, *supra* note 269.

[282] *Id.* at 1329.

[283] *Id.* at 1370.

[284] *Id.* at 1355.

[285] *Id.*

[286] *See e.g.*, Prentky & Burgess, *supra note* 10, at 104; Richard C. Lewontin & Daniel L. Hartl, *Population Genetics in Forensic DNA Typing*, 254 SCIENCE 1745, 1746 (1991).

what is the "incriminating significance" of one-in-a-million, or for that matter, one-in-a-quadrillion?[287]

The context of SVP cases is, in an important way, different from the criminal trial that Tribe worries about. The plain legislative directive in commitments is to assess future risk, not historical guilt. This is a process that calls explicitly for probability assessment; thus, statistical statements of probabilistic risk fit clearly with the issue of legal relevance. In SVP proceedings (unlike criminal proceedings), the central question is not *whether* or *how* probabilistic evidence is relevant (that question was answered in the passage of the laws), but rather how best to assess the probabilities. Of course, it is precisely this prospective, risk-based aspect of SVP laws that make them constitutionally suspect. But the constitutional doubts appear to have been rejected, and we now focus on the mechanics of the laws.

The question of jury competence is, of course, an empirical one. The available empirical evidence suggests that juries are not, as a general matter, incompetent to assess scientific evidence.[288] Nonetheless, one area in which some scholars do express concern is in jury ability to assess appropriately statistical or probabilistic evidence, and in particular, evidence about risk.[289] Authoritative commentators argue, however, that juries are "more likely to undervalue, rather than overvalue, statistical evidence."[290] Brian C. Smith and his co-authors found

[287] We note that courts have largely overcome this fear of large numbers, and now, for example, routinely admit expert testimony about DNA typing. *See* Brian C. Smith et al., *Jurors' Use of Probabilistic Evidence*, 20 LAW & HUM. BEHAV. 49 (1996) (though DNA evidence remains controversial, "in many contexts, courts readily admit probabilities associated with that evidence.").

[288] Michael S. Jacobs, *Testing the Assumptions Underlying the Debate About Scientific Evidence: A Closer Look At Juror "Incompetence" and Scientific "Objectivity"*, 25 CONN. L. REV. 1083, 1094 (1993) ("The overall picture of the jury that emerges from the available data indicates that juries are capable of deciding even very complex cases, especially if procedures to enhance jury competence are used."(quoting Joe S. Cecil et al., *Citizen Comprehension of Difficult Issues: Lessons from Civil Jury Trials*, 40 AM. U. L. REV. 727, 764 (1991))).

[289] *Id.* at 1096-97 (stating that "[u]nderstanding and evaluating statistical evidence, for example, seems to present real difficulties to most ordinary jurors, as does the task of accurately assessing information about risk.").

[290] Brian C. Smith et al., *Jurors' Use of Probabilistic Evidence*, 20 L. & HUM. BEHAV. 49, 51 (1996) (referring to M. Saks & R. Kidd, *Human Information Processing and Adjudication*, 15 L. & SOC'Y REV. 123-160 (1980-81)); *see also* Jacobs, *supra* note 288, at 1096, n.61; William C. Thompson & Edward L. Schumann, *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy,* 11 L. & HUM. BEHAV. 167, 183 (reporting on experimental findings that people tend to "underutilize associative evidence"). Note that much of the research on this issue has been undertaken in a context that differs in a significant way from the SVP context. Experimental research such as Smith et al.'s, has examined juror findings of criminal guilt in light of probabilistic evidence of a match between characteristics of the defendants (e.g., DNA, blood type, etc.) and people in the population at large. *See* Smith et al., *supra* note 290. In other words, the statistics do not directly measure the guilt of the defendant, but rather might constitute some circumstantial evidence of guilt. The experiments measure jurors' ability to use the statistical evidence properly, i.e., to properly assess the impact on the determination of guilt of the given probability. SVP cases are different in that the statistical evidence purports to be direct evidence of the central material determination: the probability of the

in experimental research in the criminal law context that, "[j]urors' sensitivity to variation in both the statistical and nonstatistical evidence, along with the overall tendency to underuse the forensic evidence, clearly demonstrate that the probabilistic evidence did not 'dwarf' the soft evidence."[291] More to the point, Hilton and Simmons' recent study found that professional fact-finders, confronted with both clinical and actuarial information about future violence, were largely uninfluenced by the actuarial information.[292] A recent empirical study with mock juries drew conclusions that support our argument. Krauss and Sales found that mock jurors were influenced more strongly by clinical assessments of dangerousness than by actuarial assessments.[293] Further, they found that "adversary procedures," such as cross-examination and competing expert testimony, had "significantly less" impact on clinical testimony than on actuarial testimony.[294] A subsequent study, while failing to replicate Krauss and Sales' results, found that mock jurors in SVP cases were not differentially influenced by clinical versus actuarial risk assessment testimony.[295]

Finally, critics might worry that the potentially imperfect fit between ARA and the salient legal questions of risk might lead jurors to give the ARA more weight than it deserves. As we discussed above, current ARA scales assess long-term risk without intensive community supervision or state-of-the-art treatment,[296] whereas the legal question posed by at least some SVP statutes is to assess the risk with these additional factors present. But, contrary to the critics, we suggest that even with this limitation, ARA can improve risk assessment. ARA provides a potentially sound benchmark for historical (i.e., static) risk. Absent a dynamic component to the ARA scale, the examiner is free to incorporate pertinent information about treatment and supervision into summary conclusions. Because the parameters of ARA are clear, attorneys can raise the fit issue in high relief. Without this transparency, clinical assessments appear to fit precisely with the legal question, but this is potentially false precision, resting on the opaque judgment of the expert.

A small but growing number of cases suggest that courts and lawyers are beginning to understand that ARA provides a framework for more accountability in risk assessment. In *Cooley v. Superior Court*, a California trial court dismissed

defendant's recidivism. Thus, the possibility that the significance of the statistical evidence will be misunderstood is of less concern in SVP cases.

[291] Smith et al., *supra* note 290, at 75.

[292] *See generally* N. Zoe Hilton & Janet L. Simmons, *The Influence of Actuarial Risk Assessment in Clinical Judgments and Tribunal Decisions about Mentally Disordered Offenders in Maximum Security*, 25 LAW & HUM. BEHAV. 393 (2001).

[293] Daniel A. Krauss & Bruce D. Sales, *The Effects of Clinical and Scientific Expert Testimony on Juror Decision Making in Capital Sentencing,* 7 PSYCHOL. PUB. POL'Y & L. 267, 300 (2001).

[294] *Id.* at 303.

[295] Guy & Edens, *supra* note 251, at 215 (finding little support for the hypothesis that clinical opinion testimony would be more influential than actuarial based testimony).

[296] *See* discussion *supra* Part V.C.

a petition after a careful review of the evidence.[297] The key finding for the trial court was that the actuarial results were not reliable enough, and that the state's expert had not taken into account dynamic factors.[298] And in *In re Dean*, the court reported a detailed critique of the state's risk assessment focusing on the shortcomings and misuse of ARA.[299]

These decisions suggest that the presence of ARA can provide the tools and the occasion for a more thorough examination of the adequacy of risk assessment testimony by trial courts. At the same time, ARA can be misinterpreted, and lead to more, rather than less, confusion and inaccuracy in risk assessment. In the next part of this article, we offer a number of recommendations to facilitate the salutary, and minimize the harmful, effects of ARA.

VI. INCREASING ACCOUNTABILITY AND ACCURACY THROUGH ARA TESTIMONY: SUGGESTED COURTROOM GUIDELINES FOR THE USE OF ARA

We believe that the controversy over the admissibility of ARA testimony in SVP cases is healthy, but ultimately misplaced. The debate has been healthy because it has exposed, in a rather coherent way, some of the shortcomings and limitations of ARA, and thereby of risk assessment in general. It is ultimately misplaced, in our judgment, because in the real world of SVP cases, where courts are legislatively mandated to make risk assessments and routinely use clinical judgment in that process, it is incoherent to ignore ARA.

Attention should now focus on the relative scientific merits of different ARA scales and the ways in which ARA can provide an optimally positive influence in SVP cases. ARA will be a good influence if it increases accuracy and accountability in the risk assessment process of SVP cases. ARA will have a neutral influence if it is simply subsumed into the opaque and undifferentiated category of "clinical" risk assessment. It will have a bad influence if it is misused[300] or its limitations are ignored, and it thereby contributes, and lends its veneer of science, to an inherently unreliable process.

A. Doing Good: Increasing Accuracy, Exposing Shortcomings, and Enabling Accountability in Risk Assessment

ARA can do good in three ways. First, ARA can increase accuracy when compared to clinical judgment. Second, ARA can increase the clarity and the understanding of the process and the shortcomings of risk assessment. By revealing all of the steps leading to the assessment, ARA makes transparent what has been omitted (e.g., dynamic factors), as well as the potential sources and

[297] 57 P.3d 654, 675 (Cal. 2002).

[298] *Id*. at 673-74.

[299] *See generally In re* Dean, No. 96-2-02973, 2000 WL 690142, at *3 (Wash. Ct. App. May 30, 2000).

[300] Misuse results if ARA's application, interpretation and incorporation into the examiner's finding and conclusions fail to adhere strictly to recommended procedures and guidelines.

magnitude of error. In addition, the enhanced exposure of ARA shortcomings should translate to enhanced exposure of the shortcomings of clinical assessment. An excellent example is *In re Coffel*, in which an appellate court found a clinical assessment wanting after applying the same kind of foundational vetting that many courts are directing at ARA.[301]

Third, ARA permits accountability by quantifying estimates of risk (probabilities) *and* the magnitude of likely error. The quantification gives judges the ability to determine the legal sufficiency of risk assessment testimony – i.e., whether it crosses the likelihood-threshold of the law. Without quantification, courts must rely on opaque categorical labels ("likely," "highly likely"), which hide implicit, unarticulated, standards of likelihood.[302] A small study published in 1999 found a reluctance by clinicians to use numerical probability figures in communicating risk.[303] Because a large proportion of the study subjects thought that "the state of the research literature doesn't justify using specific numbers,"[304] recent advances in ARA might facilitate more widespread use of quantification among experts. Several courts have resisted the call to quantify risk standards.[305] In part, these courts have justified their refusal on the grounds that the probability aspect of risk must be evaluated in connection with the severity of the predicted behavior, and that this sort of reciprocating relationship between risk and severity requires a case-by-case evaluation by the trier of fact. In our judgment, the refusal to set any quantitative guidelines for risk assessments facilitates arbitrary application of the law. Setting such guidelines would not be difficult, and could respect the desire to relate risk to severity. Courts could, for example, divide severity into three roughly defined levels, giving examples for each level, and setting quantitative probability standards for each level. Trial courts would have to explain their findings by reference to this grid.

In order to reap these benefits of ARA, courts must take additional steps to minimize the chance of prejudice from the use or misuse of ARA. We propose three steps to address possible prejudice. First, at the admissibility stage, courts should focus on the qualifications of the examiner rather than the "general acceptance" or reliability of the evidence. Some research indicates that mental health clinicians may lack formal training in risk assessment, and thus may be

[301] *See In re* Coffel, No. 79989, 2003 WL 716682, at *12 (Mo. Ct. App. March 4, 2003) (noting that factors used in forming opinion of likelihood of individual to reoffend was based on "'clinical expertise' that is simply nonexistent").

[302] *Cf.* Monahan & Steadman, *supra* note 275, at 935 (stating that mental helath professionals have long been recommended to communicate risk assessments of violence in probabilistic terms and that as predictions become more valid, they will become more likely to be expressed as probabilities).

[303] Heilbrun, et al., *supra* note 59, at 399-401.

[304] *Id.* at 399.

[305] *See, e.g.*, People v. Ghilotti, 44 P.3d 949, 971-72 (Cal. 2002) (rejecting argument that "likely" means "better than even chance"); *Boucher*, 780 N.E.2d at 50 (holding that "[w]hile the Commonwealth is required to prove beyond a reasonable doubt that a person is sexually dangerous . . . it is not required to prove to any particular mathematical quantum the likelihood of his committing another sexual offense.").

unaware of risk assessment research findings.[306] Courts must be confident that examiners qualified as "experts" are indeed experts in risk assessment with sex offenders—that they are knowledgeable about and qualified to perform all forms of risk assessment (including ARA), and are able to fully and accurately explain and interpret the data generated by these assessments.[307] To this end, we propose that the forensic divisions of the principal professional organizations, the American Psychological Association and the American Psychiatric Association, jointly develop training and certification programs for insuring adequate competence in risk assessment. We appreciate the somewhat paradoxical nature of this recommendation, given that both organizations have taken formal positions in the past opposing the use of "dangerousness predictions" in high-stakes legal settings.[308] We approach this, however, from a practical standpoint. Courts are mandated by law to render these judgments and a small army of practitioners is complying. The optimal response, from our vantage, is not to ignore reality but to confront it by ensuring that those professionals who do these high-stakes evaluations for the courts are properly trained.

Second, both courts and mental health professionals have an obligation to insure that the limitations of risk assessment, including ARA, are fully explored and explained.[309] Though we have alluded to some of these limitations above, a full discussion is beyond the scope of this paper. Here, it suffices to point out that the clearest way to neutralize any undue weight arising from the "scientific" aspect of ARA is to take advantage of the science to expose its limitations.

[306] *See* Eric B. Elbogen, Cynthia Calkins Mercado, Mario J. Scalora, & Alan J. Tomkins, *Perceived Relevance of Factors for Violence Risk Assessment: A Survey of Clinicians*, 1 INT'L J. FORENSIC MENTAL HEALTH 37, 44 (2002). *Compare In re* Seibert*,* 2003 WL 722871, at *2 (per curiam) (rejecting argument that psychologist who referred to ARA in performing risk assessment must be "expert in the use of statistical instruments," reasoning that the ARA "was relevant as part of the clinical assessment."), *with In re* Coffel, 2003 WL 716682, at *11 (holding that risk assessment opinion of psychologist was "rank speculation, not substantial evidence" where psychologist "had no training or experience in assessing the risk of reoffense. She knew of no research" on the subject.).

[307] *See, e.g.*, Jacobs, *supra* note 288, at 1096-97 (arguing that difficulties in "[u]nderstanding and evaluating statistical evidence . . . may be attributable to shortcomings in advocacy or explanatory skills of lawyers and scientific experts.").

[308] *See* Randy Otto, *On the Ability of Mental Health Professionals to "Predict Dangerousness": A Commentary on Interpretations of The "Dangerousness" Literature*, 18 LAW & PSYCHOL. REV. 43, 49-50 (1994) (citing John Monahan et al., *Report of the American Psychological Association Task Force on the Role of Psychology in the Criminal Justice System*, 33 AM. PSYCHOLOGIST 1099, 1110 (1978)) for the following proposition:

> It does appear from reading the research that the validity of psychological predictions of violent behavior, at least in the sentencing and release situations we are considering is extremely poor, so poor that one could oppose their use on the strictly empirical grounds that psychologists are not professionally competent to make such judgments.

[309] *See* Grisso, *supra* note 64, at 6 (noting that it is the obligation of mental health professionals to reveal these limitations clearly to the court).

Third, we urge courts to control the language used to describe the statistical evidence.[310] Both research and commonsense suggest that the way in which risk is communicated affects the way in which it is understood.[311] Since risk is inherently a group characteristic, risk assessments should be ascribed to the relevant group, not to the individual defendant.[312] Thus, courts should insist that experts characterize the risk of recidivism for members of a specified group, describe the development and/or validation sample that the group comes from (e.g., a group of incest offenders in the development sample who scored X on the SORAG and who were followed for X years with X supervision) and describe the way in which the defendant shares (or does not share) characteristics of this group (he is an incest offender who also scored X, etc.).

Accordingly, courts should exclude testimony that directly ascribes a risk to the defendant. For example, in *In re Dean*, the court described the risk testimony as follows:

> Based on the information in [the defendant]'s file, Dr. [H] determined that [the defendant]'s SORAG was 13. According to Dr. [H], this meant that *[the defendant]'s risk to reoffend in the next 10 years was 59 percent* and *[the defendant]'s risk to reoffend* in 7 years was 45 percent.[313]

By ascribing a risk directly to the defendant, this testimony obscured the critical steps that must be taken to link the ARA results to the legally relevant measures: the "fit" of the instruments to the legal categories, the process of generalizing from the development or validation samples to the defendant, and the band of potential error defining the score and its associated probability.

Under our suggestion, the expert would have testified that 59% of the individuals who comprised the development sample of the SORAG – or a validation sample of the SORAG – and who scored 13 were subsequently rearrested for a sex crime within 10 years. Our suggested language underscores the scientific nature of ARA – specifically, the empirical relationship between risk factors and recidivism – without exaggerating the science by extending its patina to judgments that fall in the domain of law. It also emphasizes the need to translate the inherently group-based information of risk assessment to the individual assessment required by law, and resists the reification of the numbers as a "characteristic" of the defendant.[314]

[310] *See* John Monahan et. al, *Communicating Violence Risk: Frequency Formats, Vivid Outcomes, and Forensic Settings*, 1 INT'L J. FORENSIC MENTAL HEALTH 121, 121 (2002) (asserting that "better-informed legal decision-making about risk can be achieved only when violence risk is assessed accurately and communicated so it can be understood by the decision-maker.").

[311] *See generally* Monahan & Steadman, *supra* note 275.

[312] *See* Otto & Petrila, *supra* note 6, at 16.

[313] *In re* Dean, 2000 WL 690142, at *2 (emphasis added).

[314] Some studies use "survival analysis" and report "failure rates," a statistic that is distinct from, and often yields considerably higher numbers, than the recidivism rates determined exclusively by

B. *Avoiding Neutral and Bad*

Though ARA can improve accuracy and accountability of SVP proceedings, this benefit can be easily lost or turned to harm if ARA is misused, or its real limitations are ignored. As we have argued, if ARA is routinely "modified" by clinical judgment, it will lose the advantage it gets from its empirical grounding. But equally, ARA cannot be adopted "automatically" because its fit and weight in making the legally relevant finding require the exercise of judgment. At the very least, the following factors need judgment in order to place ARA in its proper context: (a) the adequacy of the methods by which the tool was developed; (b) the care with which it was scored in the particular case; (c) the fit or match of the defendant on relevant characteristics to the development and/or validation samples; (d) published indices of predictive accuracy; (e) the "fit" with the particular question presented in the case; and (f) the weight of the actuarially-derived assessment, given the degree of fit and shortcomings in development and application.

Finally, we readily acknowledge that the use of ARA can do harm in SVP cases. Harm will come if the principle of actuarial superiority is misinterpreted to mean that ARA is infallible, error-free, or invariably sufficient to meet the state's burden. Harm will occur if the shortcomings of ARA are not brought out in testimony. Finally, harm will occur if courts engage in or permit a selectively biased use of ARA, accepting it to justify decisions already made,[315] rejecting it when it does not.[316]

VII. CONCLUSION

We have argued that ARA is admissible in SVP proceedings under prevailing standards. We would be remiss if we did not press the argument one step further. ARA is not only admissible, but ought to form a key part of the risk assessment in SVP cases. We need to make this argument because courts are not alone is resisting the use of actuarial methods for assessment. As Grove and Meehl note, "[d]espite sixty-six years of consistent research findings in favor of the actuarial method, most professionals continue to use a subjective, clinical judgment approach when making predictive decisions."[317]

the percentage of the sample that were identified as sexual recidivists. For a fuller discussion, see Robert A. Prentky, et al., *1997 Recidivism Rates Among Child Molesters and Rapists: A Methodological Analysis,* 21 LAW & HUM. BEHAV. 635 (1997).

[315] *See* James Bonta, *Offender Risk Assessment: Guidelines for Selection and Use*, 29 CRIM. JUST. & BEHAV. 355, 357 (2002) (offering reasons for use of ARA).

[316] *See* State v. Robertson, 768 N.E.2d 1207, 1214-1215 (Ohio Ct. App. 2002) (upholding SVP commitment despite actuarial assessment showing only 20% risk of reoffense).

[317] Grove & Meehl, *supra* note 66, at 299. *See also* Bonta, *supra* note 315 (reporting survey of correctional psychologists showing small percentage use well-accepted actuarial tools to assess risk).

In an SVP evaluation the stakes typically are very high, involving liberty interests (for the defendant), safety interests (for victims or potential victims), and competing claims on scarce treatment and prevention resources (for policy makers).[318] Although human error and miscalculation are inevitable, the search for the closest approximation of the "truth" should be uncompromising. To that end, courts and mental health experts undertaking the task of rendering a judgment about someone's future dangerousness must exercise utmost care and employ, as indicated above, methods and procedures that reflect the state-of-the-art. Clearly, the state-of-the-art, best-practice method embraces the "second generation" of empirical research on risk assessment.

The same best-practice method also leads to the conclusion that reliance solely on clinical judgment is improper and, under forensic circumstances, arguably unethical. In their lengthy review of the large number of situations in which actuarial prediction is demonstrably superior to clinical prediction, Grove and Meehl concluded tersely that, "[t]o use the less efficient of two prediction procedures in dealing with [‘high stakes’ predictions] is not only unscientific and irrational, it is unethical."[319] By the same token, however, best-practice methodology would not, in our opinion, rely exclusively on the results of an actuarial risk assessment and would never knowingly exclude potentially critical, risk-relevant information that is not reflected in the actuarial risk assessment.

Having made this argument, we, nonetheless, find ourselves in the paradoxical situation clearly articulated by Grisso.[320] Mental health professionals have the obligation to explain fully the limitations of ARA. But if their “testimony candidly and forthrightly stated all of those limitations – questionable standardization, unknown inter-examiner reliability, ambiguous base rates, and only the first steps in validation – one is in danger of causing courts to discount the tools’ results altogether.”[321] If these results are discounted, courts will fall back on clinical assessments, whose limitations are more severe, but also more hidden, than ARA.

The solution to this paradox is simple: rather than dumbing down risk assessment by insisting that its limitations be hidden behind the opacity of clinical judgment, courts should raise the bar by insisting that clinical assessment be held to the same standard of reliability and validity as ARA, and by taking advantage of the quantification of ARA to set enforceable and uniform legal standards for risk assessments.

There are no magical solutions to the problem of assessing highly idiosyncratic facets of human behavior. Our only recourse is to persist in our best efforts to find empirically-based answers to assist in assessing risk in different situations, for different types of offenders, and most importantly, with a consistent

[318] *See* Janus, *supra* note 16, at Part III.C. (arguing that poor risk assessment exacerbates an already severe misallocation of resources for SVP programs).

[319] Grove & Meehl, *supra* note 66, at 320.

[320] Grisso, *supra* note 64, at 7 (advocating a duty to inform the court as to the risks involved with ARA methods, yet noting that this information may cause the court to exclude such evidence).

[321] *Id.* at 7 (emphasis omitted).

and uniform approach. Those best efforts have resulted in marked progress over the past ten years, yielding a number of empirically validated risk assessment scales. Since we may fairly assume that the problem of sexual violence is unlikely to abate in the near future, the demand for assessments of sexual dangerousness also will not abate, and with it the apparent tension over how these assessments should be conducted. In most other realms of inquiry, the answer would be obvious: Assessments should be conducted in the most scientifically credible and reliable fashion possible. The weight of the evidence points to the superiority of actuarial assessment of risk over clinical assessment of risk.

ARA, in our opinion, offers the best hope for transparency, accountability and uniformity in the risk assessment process. As we have described, its misuse and misunderstanding carries the risk of substantial prejudice. But ARA has dangers even if, and perhaps especially if, it is used well. With improved ability to assess risk, lawmakers may seek to expand risk-based deprivations of liberty, allowing civil-commitment-style "dangerous person laws" to displace the normal means of dealing with antisocial behavior, the criminal justice system.[322] And these new laws may, in turn, generate an even more sophisticated and far-reaching science of prediction, which may, in an escalating spiral, facilitate yet more preventive detention. Yet the moral qualms associated with preventive liberty-deprivation will be strong even if the assessment of future risk approaches perfection. Our advocacy for ARA is, for this reason, undertaken with considerable trepidation.

[322] *See* Foucha v. Louisiana, 504 U.S. 71, 80-81 (1987) (discussing the State's right to confine persons who pose a danger to the community).