ADVENTURES IN RISK: Predicting Violent and Sexual Recidivism in Sentencing Law

Melissa Hamilton[*]

I. INTRODUCTION

A new arena inviting collaboration between the law and sciences has emerged in criminal justice. The nation's economic struggles and its record-breaking rate of incarceration have encouraged policymakers to embrace a

[*] Visiting Criminal Law Scholar, University of Houston Law Center; J.D., The University of Texas School of Law; Ph.D, The University of Texas at Austin. The author thanks the faculty reading group at the University of Houston Law Center for their helpful suggestions for this work.

new penology which seeks to simultaneously curb prison populations, reduce recidivism, and improve public safety.[1] The new penology draws upon the behavioral sciences for techniques to identify and classify individuals based on their potential future risk and for current best evidence to inform decisions on how to manage offender populations accordingly.[2] Empirically driven practices have been utilized in many criminal justice contexts for years, yet have historically remained "a largely untapped resource" in sentencing decisions.[3] One reason is that sentencing law in America has for some time been largely driven by retributive theories.[4] The new penology clearly incorporates utilitarian goals and welcomes an interdisciplinary approach to meet them.

As criminal justice officials seek more cost-effective solutions to criminal offending, demand for evidence-based[5] sentencing has intensified.[6] Evidence-based sentencing practices have recently adopted data-driven risk assessment tools to predict recidivism. A central idea is that accurate information on risk can inform decisions to reserve prison resources for high risk offenders, while reducing recidivism of low risk defendants by diverting them to less costly, community-based sanctions. Indeed, risk assessment stands at the "leading edge of the next wave of [sentencing] reform" and is considered the "next frontier" in modern sentencing law.[7] At least a majority of states currently use risk-based assessments in their sentencing systems.[8]

[1] Juliene James et al., *A View from the States: Evidence-Based Public Safety Legislation*, 102 J. CRIM. L. & CRIMINOLOGY 821, 823 (2012); Michael A. Simons, *Sense and Sentencing: Our Imprisonment Epidemic*, 25 J. C.R. & ECON. DEV. 153, 166 (2010).

[2] Malcolm M. Feeley & Jonathan Simon, *The New Penology: Notes on the Emerging Strategy of Corrections and Its Implications*, 30 CRIMINOLOGY 449, 452 (1992).

[3] Jordan M. Hyatt et al., *Reform in Motion: The Promise and Perils of Incorporating Risk Assessments and Cost-Benefit Analysis into Pennsylvania Sentencing*, 49 DUQ. L. REV. 707, 714 (2011) [hereinafter Hyatt, *Reform*].

[4] *See generally* Gerard V. Bradley, *Retribution and Overcriminalization*, HERITAGE FOUND. (Mar. 1, 2012), http://thf_media.s3.amazonaws.com/2012/pdf/lm77.pdf.

[5] Stephen Hart, *Evidence-Based Assessment of Risk for Sexual Violence*, 1 CHAP. J. CRIM. JUST. 143, 146 (2009) ("Evidence-based means an action or decision that was guided by, based on, or made after consulting a systematic review of relevant information in the form of observation, research, statistics, or well-validated theory.").

[6] Hyatt, *Reform*, *supra* note 3, at 714; *see also* J. Richard Couzens, *Realignment and Evidence-Based Practice: A New Era in Sentencing California Felonies*, 25 FED. SENT'G REP. 217 (2013) (noting number of counties in California (2010–2011) adopting some form of evidence-based sentencing practices increased nearly 33%); Simons, *supra* note 1, at 166 (recognizing states moving from punishment to reducing recidivism).

[7] Hyatt, *Reform*, *supra* note 3, at 733; *see also* Roger K. Warren, *Evidence-Based Sentencing: Are We Up to the Task?*, 23 FED. SENT'G REP. 153, 157 (2010) (labeling evidence-based sentencing the "new frontier").

[8] Shawn Bushway & Jeffrey Smith, *Sentencing Using Statistical Treatment Rules: What We Don't Know Can Hurt U*s, 23 J. QUANTITATIVE CRIMINOLOGY 377, 378 (2007).

The science of risk prediction continues to progress. Stakeholders now promote as best practices in evidence-based sentencing jurisdictions the use of actuarial risk assessment tools, particularly in preference to unstructured clinical judgments of future dangerousness. Actuarial tools guide evaluators through a regimented process in which they use existing data to score factors associated with recidivism, provide the given weights, and then extract an estimated probability of reoffending.[9] A *New York Times* article describes the practice as "sentencing by numbers," in which the decision-maker is meant to sentence "offenders the way insurance agents write policies, based on a short list of factors with a proven relationship to future risk."[10] In sum, risk prediction is shaping sentencing philosophies, with technological risk instruments entrenched in decisions about punishment.[11]

While evidence-based sentencing practices led by actuarial risk instruments are gaining widespread approval, critical questions must be raised. Empirically derived and mathematically refined actuarial predictions appear to embody a desirable and progressive policy reform. Actuarial models are praised for being objective, reliable, logical, and quantifiable.[12] However, law and justice have suffered too many unfortunate experiences with purportedly scientific evidence which only later was revealed to be no less than junk science.[13] The purpose of this Article is to address the use of "actuarial justice" in sentencing decisions and to question whether reliance upon even "best practices" is justified. Sentencing and punishment, on the one hand, are considered crucial decisions to protect public safety while, on the other hand, they necessarily involve significant infringements upon such individual rights as liberty and privacy. If officials have misplaced their trust in the actuarial model, injustice may have invaded sentencing regimes.

This Article is concerned with the influence of actuarial risk tools in sentencing decisions generally. There are now dozens of actuarial tools available, some of which are applied across the board, while others are designed for specific types of offenders or crimes or for other more discrete

[9] Kelly Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, 30 JUST. Q. 270, 271 (2013) ("Actuarial risk tools guide practitioners through a logical and simple process to itemize, score, and interpret information about offenders.").

[10] Emily Bazelon, *Sentencing by the Numbers*, N.Y. TIMES, Jan. 2, 2005, (MAGAZINE), at 18.

[11] BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE 188 (2007).

[12] *See infra* notes 63–70 and accompanying text.

[13] *See generally* Erica Beecher-Monas, *Blinded by Science: How Judges Avoid the Science in Scientific Evidence*, 71 TEMP. L. REV. 55 (1998).

segments of society.[14] Thus, in order to delve into more detail about the purported scientific properties and realistic abilities of actuarial sentencing, this Article will eventually focus upon the risk assessment tools targeting those offenders for which public safety concerns are paramount: violent and sexual offenders.

The Article proceeds as follows. Section II introduces the ideas and purposes underlying evidence-based practices. Models of actuarial risk assessments are presented and how they are employed in evidence-based sentencing is demonstrated. Section III outlines a host of evidentiary, empirical, and pragmatic issues with these actuarial tools in sentencing matters. Actuarial risk focuses almost exclusively on the proportionate likelihood of recidivism, without providing data concerning other important dimensions of risk, such as imminence, duration, type of recidivism, or severity of harm. The very limited nature of information provided renders actuarial risk results as insufficiently relevant to assist in any factual question in sentencing matters. A host of statistical measures are presented, and new statistics computed and offered herein, to show that the predictive ability of actuarial tools is rather weak, and high error rates are a consequence thereof. In addition, the popular actuarial risk instruments are not generalizable to routine sentencing populations in the United States. The contention herein is that, altogether, actuarial risk models fail to meet the high standards of validity and reliability for admissibility in the law as expert evidence. From a practical perspective, actuarial models are problematic as well, in that they use group-based statistics, which cannot for empirical reasons be directly used for individual predictions of risk. Section IV then responds to the argument that actuarial results ought to be admissible as just one piece of data in a decision-making process in which an array of information is considered. It is questionable whether many evaluators are sufficiently knowledgeable about actuarial risk methodologies to qualify as expert witnesses in the first place. Further, this Article maintains that actuarial predictions are overly prejudicial, confusing, and misleading, and therefore judges ought to act as gatekeepers for the law to exclude or substantively limit actuarial risk results as evidence in sentencing proceedings.

[14] Jay P. Singh et al., *A Comparative Study of Violence Risk Assessment Tools: A Systematic Review and Metaregression Analysis of 68 Studies Involving 25,980 Participants*, 31 CLINICAL PSYCHOL. REV. 499, 500 (2011) [hereinafter Singh et al., *Metaregression*].

II. EVIDENCE-BASED SENTENCING: THE REIGN OF ACTUARIAL RISK ASSESSMENTS

A new penology has emerged in which criminal justice officials are loosening the grip of retributive ideologies by embracing utilitarian objectives as well. The *sine qua non* of the new penology is risk.[15] The idea is to harness the ability to differentiate among offenders based on their likelihood of future recidivism. The assessment of risk hopefully informs utilitarian judgments to more strategically use incarceration, to craft appropriate rehabilitative programming, and to otherwise manage offender populations presently and in the future. Courts and correctional authorities are increasingly using risk estimates for a variety of reasons, including decisions on pretrial release, conditions of probation, parole, civil commitment, and sentencing.[16] The sentencing decision is likely the most critical legal event, comparatively, as it is explicitly intended to be punitive in nature, is meant as an indication of community condemnation for criminal culpability, and it enjoys greater substantive and procedural processes. Thus, even if risk assessment tools are appropriate for use in decisions regarding bail, probation and parole, and civil commitment, their use in sentencing may be a different matter entirely from perspectives of law and justice.

A. The Allure of Evidence-Based Practices

The new penology's goals of curtailing prison populations, reducing recidivism, and protecting the public require a balancing act at times. Officials wish to identify offenders who can properly be diverted from prison either because they pose little threat or seem good candidates for rehabilitation in community-based programs.[17] Reducing reliance on imprisonment makes fiscal sense as community corrections options are far less costly than prison.[18] Still, several proponents assert that evidence-based sentencing is appropriate not only to potentially divert low risk defendants from prison. Risk judgments can also assist decisions in the reverse, e.g., for

[15] Hazel Kemshall, *Crime and Risk*, *in* RISK IN SOCIAL SCIENCE 76, 77 (Peter Taylor-Goodby & Jens O. Zinn eds., 2006).

[16] PEW CTR. ON THE STATES, RISK/NEEDS ASSESSMENT 101: SCIENCE REVEALS NEW TOOLS TO MANAGE OFFENDERS 2 (2011), *available at* http://www.pewtrusts.org/~/media/legacy/uploadedfiles/pcs_assets/2011/PewRiskAssessmentbriefpdf.pdf.

[17] Melissa Hamilton, *Prison-by-Default: Challenging the Federal Sentencing Policy's Presumption of Incarceration*, 51 HOUS. L. REV. 1271, 1284–85 (2014).

[18] John Stuart & Robert Sykora, *Minnesota's Failed Experience with Sentencing Guidelines and the Future of Evidence-Based Sentencing*, 37 WM. MITCHELL L. REV. 426, 461 (2011).

the strategic use of preventive incapacitation for those at highest risk of recidivism.[19] Sentencing also involves proportionality. Optimistically, risk-oriented practices can help prevent judges or juries from over punishing by sending low risk individuals to prison as well as from under punishing by issuing community sanctions to high risk defendants.[20]

There is little doubt that judgments on future dangerousness have been a part of the decision-making process in sentencing for a long time, even in mainly retributive jurisdictions. Its role had been more casual and often mysterious.

> Informally, sentencing judges have long assessed risk of re-offense in crafting a defendant's sentence. Sometimes, the consideration of risk happened through evaluation of a defendant's prior criminal record, whether as part of a fully discretionary decision or as part of a guidelines system that includes enhanced recommended punishments for repeat offenders. Other times, judges relied on their own intuition, instinct and sense of justice to impose more severe sentences upon offenders whom they, based on their frequently unspoken clinical prediction, believed presented an enhanced risk to the public in the future.[21]

Such an unstructured and unregulated method of predicting risk can be unpalatable. Punishment may as a result be viewed as merely representing idiosyncratic, biased, and unreliable preferences of individual judges.

In lieu thereof, the philosophy of evidence-based sentencing embraces the utilization of scientifically derived information and structured methods to assess risk. Evidence-based sentencing is now promoted by judges, legislatures, and policy groups.[22] On behalf of the judiciary, a direct contributor to the emergence of evidence-based practices was a joint project sponsored by the National Center for State Courts, the National Judicial College, and the Crime and Justice Institute which created and publicly promotes a curriculum to educate sentencing judges nationwide about the benefits of considering factors that have been empirically validated as being

[19] *See, e.g.*, Jordan M. Hyatt et al., *Follow the Evidence: Integrate Risk Assessment Into Sentencing*, 23 FED. SEN'G REP. 266, 266 (2011) [hereinafter Hyatt, *Integrate*]; Hyatt, *Reform*, *supra* note 3, at 735; Michael Marcus, *MPC—The Root of the Problem: Just Deserts and Risk Assessment*, 61 FLA. L. REV. 751, 771 (2009).

[20] Hannah-Moffat, *supra* note 9, at 271.

[21] Hyatt, *Reform*, *supra* note 3, at 724–25.

[22] Memorandum from Vera Inst. of Justice to Del. Justice Reinvestment Task Force 9–10 (Oct. 12, 2011), *available at* http://ltgov.delaware.gov/taskforces/djrtf/DJRTF_Risk_Assessment_Memo.pdf.

either criminogenic or protective.[23] Judges in many states are also formally advancing evidence-based sentencing. The Conference of Chief Justices, representing the highest state judicial officers, passed a resolution supporting "efforts to adopt sentencing and correctional policies and programs based on the best research evidence of practices shown to be effective in reducing recidivism."[24] In addition, Utah's Judicial Council publicly supports the use of evidence-based practices,[25] while the supreme courts of Arizona and Idaho ordered probation offices in their states to specifically focus on identifying strengths and needs in presentence reports.[26]

Still, judges are not necessarily acting on their own in embracing what are perceived as reformist innovations in sentencing practices. Evidence-based sentencing has succeeded in introducing substantive change in many jurisdictions through the unique combination of "multidisciplinary input, bipartisan cooperation, the availability of data analysis and information, and the political leadership on all fronts."[27] Commentators have observed, thereby, that risk assessment has recently experienced a "remarkable resurgence,"[28] a "growing interest,"[29] and is now "widely hailed" as a progressive reform in criminal sanctioning.[30]

Evidence-based sentencing practices have rapidly evolved. The nomenclature itself has changed from an ideology of "future dangerousness" as an expansive and nebulous concept to a more refined perspective of "risk assessment."[31] This evolution has mirrored the progression in the forensic

[23] NAT'L CTR. FOR STATE COURTS, EVIDENCE-BASED SENTENCING TO IMPROVE PUBLIC SAFETY & REDUCE RECIDIVISM: A MODEL CURRICULUM FOR JUDGES 1 (2009), *available at* http://cdm16501.contentdm.oclc.org/cdm/ref/collection/criminal/id/185.

[24] Conference of Chief Justices Bd. of Dirs. & Conference of State Court Adm'rs Policy & Liaison Comm., *Resolution 12 in Support of Sentencing Practices that Promote Public Safety and Reduce Recidivism*, NAT'L CTR. FOR STATE COURTS (Aug. 1, 2007), http://www.ncsc.org/~/media/Microsites/Files/CSI/Resolution-12.ashx.

[25] Utah Judicial Council, *Judicial Council Meeting Minutes*, at 4 (2009), *available at* http://www.utcourts.gov/admin/judcncl/min-2009/min07-09.pdf.

[26] Admin. Order No. 2009-01: Budget Reductions in the Judicial Branch of Arizona, ARIZ. SUPREME COURT (2009), http://www.azcourts.gov/portals/22/admorder/orders09/2009-01.pdf; IDAHO STATE JUDICIARY, 2011 ANNUAL REPORT 7 (2011), *available at* http://www.isc.idaho.gov/annuals/2011/2011_AnnualReport.pdf.

[27] James et al., *supra* note 1, at 848.

[28] John Monahan & Jennifer L. Skeem, *Risk Redux: The Resurgence of Risk Assessment in Criminal Sanctioning*, 26 FED. SENT'G REP. 158, 158 (2014).

[29] J.C. Oleson, *Risk in Sentencing: Constitutionally Suspect Variables and Evidence-Based Sentencing*, 64 SMU L. REV. 1329, 1336 (2011).

[30] Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 805 (2014).

[31] KIRK HEILBRUN, EVALUATION FOR RISK OF VIOLENCE IN ADULTS 14 (Thomas Grisso et al. eds., 2009).

sciences involving methodologies for estimating recidivism risk.[32] Previously, forensic evaluations involving future predictions of antisocial behavior existed in the form of unstructured professional judgments, generally conducted by mental health clinicians.[33] The process of unstructured professional opinions is aptly summarized as follows:

> Clinical judgments about dangerousness might incorporate aspects of the professionals' knowledge, personal experience, "gut" feelings and other intuitions, and whatever other information about the situation that seems relevant to the problem. This process is called "clinical" because it mimics how physicians arrive at judgments about their patients' diagnoses and treatments: doctors interview and examine patients, think about what is probably going wrong, and then suggest what patients should do and prescribe treatments.[34]

While clinical judgments of risk present the advantage of being offered by (hopefully) impartial, educated, and experienced professionals, the potential for unconscious bias, unreliability, and lack of transparency remains.[35] The field of mental health risk assessment recognized these issues and, as scientists are wont to do, continues to work toward advancing knowledge about recidivism risk factors and on improving the accuracy of their risk methodologies.[36] As a result, over time, reliance upon unstructured professional opinion has yielded to structured professional judgment, which itself has generally been supplanted by actuarial risk assessments.[37] Thus,

[32] George Szmukler & Nikolas Rose, *Risk Assessment in Mental Health Care: Values and Costs*, 31 BEHAV. SCI. & L. 125, 131 (2013) ("What was conceptualized as a phenomenon that, unlike 'dangerousness,' was not a fixed quality of an individual, but dependent on the co-presence of many factors, including those external and those subject to change, tends to become, in effect, an objective, calculable, and static measure of risk attached to an individual.").

[33] Gina M. Vincent et al., *The Use of Actuarial Risk Assessment Instruments in Sex Offenders*, *in* SEX OFFENDERS: IDENTIFICATION, RISK ASSESSMENT, TREATMENT, AND LEGAL ISSUES 70, 70 (Fabian M. Saleh et al. eds., 2009).

[34] Douglas Mossman, *Evaluating Risk Assessments Using Receiver Operating Characteristic Analysis: Rationale, Advantages, Insights, and Limitations*, 31 BEHAV. SCI. & L. 23, 33 (2013).

[35] Melissa Hamilton, *Public Safety, Individual Liberty, and Suspect Science: Future Dangerousness Assessments and Sex Offender Laws*, 83 TEMP. L. REV. 697, 744–49 (2011); Jennifer L. Lanterman et al., *Sex Offender Risk Assessment, Sources of Variation, and The Implications of Misuse*, 41 CRIM. JUST. & BEHAV. 822, 834 (2014).

[36] Richard E. Redding, *Evidence-Based Sentencing: The Science of Sentencing Policy and Practice*, 1 CHAP. J. CRIM. JUST. 1, 4 (2009).

[37] Jeffrey C. Singer et al., *A Convergent Approach to Sex Offender Risk Assessment*, *in* THE WILEY-BLACKWELL HANDBOOK OF LEGAL AND ETHICAL ASPECTS OF SEX OFFENDER TREATMENT AND MANAGEMENT 341, 341 (Karen Harrison & Bernadette Rainey eds., 2013); Thomas Nilsson et al., *The Precarious Practice of Forensic Psychiatric Risk Assessments*, 32

evidence-based sentencing practices are no longer as concerned with educating decisionmakers about which characteristics or circumstances have been empirically found to be correlative or causative of either future violence or desistance from crime; instead, evidence-based sentencing is now mostly about ascertaining the numerical scores and rankings produced by actuarial tools. In other words, "actuarial sentencing" may now be the appropriate label for contemporary sentencing law and praxis.

1. Modeling Actuarial Risk

Actuarial risk tools basically rely upon aggregate statistics derived from historical experience. Actuarial tool creators study the statistical relationships between a host of variables and the risk outcome of interest using data from available samples, often referred to as developmental or normed samples. Researchers often select the stronger predictors to include in the final actuarial model. Appropriate weights often apply to to provide additional points to those factors found to offer greater predictive value than others.[38] A table of estimated probabilities of the outcome occurring is created to match to final scores. This is called an "experience table" since it is based on the observed rates of the outcome of interest from the developmental samples. The experience table might, for example, convey that of the subjects in the developmental sample who were assigned a score of six, 35% were observed to have recidivated.

In sum, the developers of actuarial instruments use existing data in an empirical way to create rules that combine highly relevant factors, provide applicable weights, create final mechanistic scores, and provide an experience table of estimated probabilities of the outcome, all based on the development sample data.[39] Developers of actuarial risk tools at times pool

INT'L J.L. & PSYCHIATRY 400, 402 (2009) ("Instead of categorical assessments of 'dangerousness', the 'risk' of violence was measured as the proportion of individuals who relapsed or committed a certain type of crime in a (hypothetical) group sharing similar rating scores on structured or semi-structured rating scales, or 'instruments.'").

[38] Actuarial models presume multiple factors produce a better predictive tool than a few. Joanna Amirault & Patrick Lussier, *Population Heterogeneity, State Dependence and Sexual Offender Recidivism: The Aging Process and the Lost Predictive Impact of Prior Criminal Charges over Time,* 39 J. CRIM. JUST. 344, 344 (2011).

[39] Jennifer L. Skeem & John Monahan, *Current Directions in Violence Risk Assessment*, 20 CURRENT DIRECTIONS IN PSYCHOL. SCI. 38, 39 (2011) (synthesizing process: "(a) identifying empirically valid risk factors, (b) determining a method for measuring (or 'scoring') these risk factors, (c) establishing a procedure for combining scores on the risk factors, and (d) producing an estimate of violence risk").

together risk groupings, referred to as risk bins, based on point totals.[40] As an illustration, a risk bin might pool together scores 10–15, yielding a single risk probability estimate. Sometimes, too, the instruments place a categorical label on a risk bin, such as the scores of 10–15 representing "moderate risk" or perhaps "high risk."

An evaluator using an actuarial risk instrument so conceived begins by scoring the various factors contained in the model. The evaluator then applies the given weights and calculates a total score. The following demonstrates a hypothesized next step involving a sexual recidivism predictive tool:

> This score translates typically into a risk category, where individuals who score positively on a number of items obtain scores placing them in a high-risk group, those who score on some items are placed in a medium-risk group, while those who score on only a few items are placed typically in a low-risk group. In most cases, the scale developers have compiled 'experience tables' from retrospective studies of released sex offenders that indicate a prediction of future risk, based on the percentage of offenders in each risk category who have recidivated. Hence, a value of 45% might be extracted for a high-risk individual over a 10-year period, which means that individuals with similar characteristics (45 in 100) re-offended within this time-period.[41]

2. Actuarial Risk in Sentencing

Across the country, reliance specifically upon *actuarial* risk assessments in sentencing is spreading.[42] Numerous policy institutes advocate they be used routinely.[43] The Pew Center not only promotes actuarial tools, it advocates that their data outputs anchor sentencing determinations.[44]

[40] Jay P. Singh et al,, *Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review*, 31 BEHAV. SCI. & L. 55, 57 (2013) [hereinafter Singh et al., *Systematic Review*] ("In the prediction-focused actuarial approach, weighted scores are assigned to criminal history, sociodemographic, and/or clinical factors empirically associated with the likelihood of antisocial behavior. These weighted scores are used to classify individuals into risk bins that correspond to probabilistic estimates of future antisocial behavior.").

[41] Leam A. Craig & Anthony Beech, *Best Practice in Conducting Actuarial Risk Assessments with Adult Sexual Offenders*, 15 J. SEXUAL AGGRESSION 193, 197 (2009).

[42] David E. Patton, *Guns, Crime Control, and a Systemic Approach to Federal Sentencing*, 32 CARDOZO L. REV. 1427, 1455 (2011) (conceptualizing actuarial risk as "becom[ing] increasingly popular" across sentencing courts).

[43] NAT'L CTR. FOR STATE COURTS, EVIDENCE-BASED SENTENCING TO IMPROVE PUBLIC SAFETY AND REDUCE RECIDIVISM: A MODEL CURRICULUM FOR JUDGES 1 (2005), *available at* http://cdm16501.contentdm.oclc.org/cdm/ref/collection/criminal/id/185.

[44] PEW CTR. ON THE STATES, RISK/NEEDS ASSESSMENT 101: SCIENCE REVEALS NEW TOOLS TO HELP MANAGE OFFENDERS 5 (2011), *available at*

Multiple state legislatures have likewise become convinced, encouraging—even mandating in some jurisdictions—the use of actuarial risk assessments to inform sentencing decisions. By statute, for example, Pennsylvania,[45] Tennessee,[46] and Alabama[47] require the use of validated risk assessment tools in sentencing proceedings. The State of Washington by law permits a judge to order a presentence risk assessment and for her to have access to the results for sentencing.[48] Vermont[49] and Kentucky[50] also target the specific use of sex offense recidivism risk tools for defendants convicted of sexual crimes by statute.

The State of Virginia is perhaps the most blatant in incorporating actuarial risk tools into sentencing. Virginia law mandates the creation and use of an actuarial risk tool to identify nonviolent, low-risk offenders specifically for the purpose of diversion from prison.[51] Another Virginia statute requires the use of a risk instrument tool concentrating on sex offenders, though for a contrasting function: actuarial results indicating higher probabilities of recidivism risk trigger gradated increases in recommended sentences.[52] At its extreme, the Virginia scheme raises the upper end of the sentencing range by 300% with the highest actuarial scores.

In practice, a number of states' probation departments (including California, Colorado, and Washington) have incorporated actuarial tools into presentence investigation routines.[53] In some cases, the actuarial tool is expected to weigh heavily in the adjudicative process. In New York, for example, probation officers are required to use an actuarial scale to assess

http://www.pewtrusts.org/~/media/legacy/uploadedfiles/pcs_assets/2011/PewRiskAssessmentbriefpdf.pdf.

[45] 42 PA. CONS. STAT. § 2154.5(a)(6) (2009).

[46] TENN. CODE ANN. § 41-1-412(b) (2013).

[47] ALA. CODE § 12-25-33(6) (2013).

[48] WASH. REV. CODE § 9.94A.500(1) (2013).

[49] VT. STAT. ANN. tit. 28, § 204a(b)(1) (2013).

[50] KY. REV. STAT. ANN. § 17.554(2) (West 2013).

[51] Hyatt, *Reform*, *supra* note 3, at 723.

[52] *Id.* at 723.

[53] JESSIKA SHIPLEY, COLO. LEGISLATIVE COUNCIL STAFF, ISSUE BRIEF NO. 12-38, PROBATION SERVICES IN COLORADO 1 (2012), *available at* http://www.colorado.gov/cs/Satellite?c=Document_C&childpagename=CGA-LegislativeCouncil%2FDocument_C%2FCLCAddLink&cid=1251634174919&pagename=CLCWrapper; WASH. STATE INST. FOR PUB. POLICY, RISK ASSESSMENT INSTRUMENTS TO PREDICT RECIDIVISM OF SEX OFFENDERS: PRACTICES IN WASHINGTON STATE 2 (2008), *available at* http://www.wsipp.wa.gov/ReportFile/1015/Wsipp_Risk-Assessment-Instruments-to-Predict-Recidivism-of-Sex-Offenders-Practices-in-Washington-State_Full-Report.pdf (discussing actuarial tool for sex offenders); CAL. SARATSO REVIEW & TRAINING COMMS., SEX OFFENDER RISK ASSESSMENT IN CALIFORNIA (2012), *available at* saratso.org/docs/RA_summary_for_judges_attys_rev3_061611.docx (same).

recidivism for sex offenders and the result should "anchor the judgment or impressions."[54] The proffer of actuarial results is clearly not one sided. Case law represents that actuarial predictions of risk are commonly being introduced in sentencing proceedings by various players: prosecution experts,[55] defense experts,[56] and probation officers in presentence investigation reports.[57]

The strong momentum for incorporating actuarial tool results in sentencing practices likely will continue in the future. The influential Model Penal Code was recently revised and it now explicitly addresses evidence-based practices in sentencing. The model legislation envisions a sentencing commission to be instructed as follows:

> The commission shall develop actuarial instruments or processes, supported by current and ongoing recidivism research, that will estimate the relative risks that individual offenders pose to public safety through their future criminal conduct. When these instruments or processes prove sufficiently reliable, the commission may incorporate them into the sentencing guidelines.[58]

The revised Model Penal Code anticipates that actuarial risk assessment will serve as a "regular part of the felony sentencing process."[59]

The preference for actuarial-based predictions of risk as the new form of evidence-based sentencing is largely explained by their guise of empiricism and science.[60] Sentencing experts and judges seem to elevate actuarial

[54] N.Y. STATE DIV. OF PROB. & CORR. ALTS., NEW YORK STATE PROBATION SEX OFFENDER MANAGEMENT PRACTITIONER GUIDANCE 9 (2009), *available at* dpca.state.ny.us/pdfs/sompractitionerguidanceluly2009.pdf.

[55] *E.g.*, United States v. Zobel, 696 F.3d 558, 565 (6th Cir. 2012); United States v. Ellis, 68 M.J. 341, 343 (A.F. Ct. Crim. App. 2010); Artrip v. State, No. 07-01-0201-CR, 2002 Tex. App. LEXIS 1267, at *12 (Tex. Crim. App. Feb. 20, 2002).

[56] *E.g.*, State v. Seward, 217 P.3d 443, 445 (Kan. 2009); Wilhite v. State, 339 S.W.3d 573, 575 (Mo. Ct. App. 2011); State v. Wilson, No. 2013AP415-C, 2013 Wis. App. LEXIS 953, at *3–4 (Wis. Ct. App. Nov. 13, 2013); Brief of Appellant at *7, *10, United States v. Coffey, No. 12-5050, 2012 WL 1268077 (6th Cir. Apr. 6, 2012); Brief of Appellant at *8, United States v. Guntharp, No. 10-4595, 2010 WL 4084584 (4th Cir. Oct. 18, 2010).

[57] *E.g.*, Harral v. Martel, No. EDCV10-1379-AG(PLA), 2011 U.S. Dist. LEXIS 47675, at *19–20 (C.D. Cal. Mar. 22, 2011); People v. Godoy, No. B214003, 2011 Cal. App. Unpub. LEXIS 2045, at *7 (Cal. Ct. App. Mar. 21, 2011); People v. Hillier, 392 Ill. App. 3d 66, 68 (2009); State v. Winters, No. 5-113/04-0575, 2005 Iowa App. LEXIS 147, at * 2–3 (Iowa Ct. App. Feb. 24, 2005).

[58] MODEL PENAL CODE: SENTENCING § 6B.09(2) (Tentative Draft No. 2, 2011), *available at* http://www.ali.org/00021333/Model%20Penal%20Code%20TD%20No%202%20-%20online%20version.pdf.

[59] *Id.* at cmt. a.

[60] Hyatt, *Reform*, *supra* note 3, at 725 ("Risk assessment tools now under consideration are more transparent, rely on data, and attempt to regularize this instinct and subject it to more

estimates over unstructured professional opinions because the former is conjectured to reduce clinical error.[61] Proponents also favor actuarial tools over probation officers' speculations in presentence reports about future dangerousness.[62]

Actuarial risk in sentencing has been lauded for being transparent,[63] mathematical,[64] and logical.[65] "There is a seductive quality to risk assessment: it appears to bring the future into the present and to make it calculable."[66] Statistical calculations of risk have been conceptualized as providing an important foundation for offering consistency in predictions,[67] standardizing sentencing,[68] and "representing hope for a new age of scientifically guided sentencing."[69] In the actuarial model of sentencing, potentially subjective verbal justifications for individual sentences are replaced with (seemingly) more objective statistical measures; on the whole, words yield to numbers.[70]

Whereas sentencing outcomes in general often draw complaints of opacity, bias, and disparity, defenders of transparency, fairness, and justice

scientifically rigorous examinations. Ensuring uniform application and the unbiased use of available data, these modern predictive tools are facilitated by the use of 'structured, empirically-driven and theoretically driven' instruments."); Redding, *supra* note 36, at 4 (quoting Kirk Heilbrun et al., *Risk-Assessment in Evidence-Based Sentencing: Context and Promising Uses*, 1 CHAP. J. CRIM. JUST. 127, 133 (2009)) ("*Actuarial assessment* is 'a formal method . . . [that provides] a probability, or expected value, of some outcome. It uses empirical research to relate numerical predictor variables to numerical outcomes. The *sine qua non* of actuarial assessment involves using an objective, mechanistic, reproducible combination of predictive factors, selected and validated through empirical research, against known outcomes that have also been quantified.'"); Ruth J. Tully et al., *A Systematic Review on the Effectiveness of Sex Offender Risk Assessment Tools in Predicting Sexual Recidivism of Adult Male Sex Offenders*, 33 CLINICAL PSYCHOL. REV. 287, 288 (2013).

[61] Oleson, *supra* note 29, at 1336; Tully et al., *supra* note 60, at 288; Roger K. Warren, *Evidence-Based Sentencing: The Application of Principles of Evidence-Based Practice to State Sentencing Practice and Policy*, 43 U.S.F. L. REV. 585, 603 (2009); Michael A. Wolff, *Evidence-Based Judicial Discretion: Promoting Public Safety Through State Sentencing Reform*, 83 N.Y.U. L. REV. 1389, 1406 (2008).

[62] Jennifer Skeem, *Risk Technology in Sentencing: Testing the Promises and Perils*, 30 JUST. Q. 297, 300 (2013).

[63] *Id.*; Hyatt, *Reform*, *supra* note 3, at 729; Tully et al., *supra* note 60, at 288.

[64] M. Roffey & S.Z. Kaliski, *To Predict or not to Predict—That is the Question*, 15 AFR. J. PSYCHIATRY 227, 227 (2012).

[65] *Id.* at 227 (conceptualizing actuarial risk "rooted in careful data collection, logical analysis and mathematical [rigor]").

[66] George Szmukler & Nikolas Rose, *Risk Assessment in Mental Health Care: Values and Costs*, 31 BEHAV. SCI. & L. 125, 131 (2013).

[67] Skeem, *supra* note 62, at 300.

[68] Hyatt, *Integrate*, *supra* note 19; Tully et al., *supra* note 60, at 288.

[69] Starr, *supra* note 30, at 2; *see also* Warren, *supra* note 61, at 631 (actuarial risk tools fosters "data-driven sentencing decisions").

[70] Rasmus H. Wandall, *Actuarial Risk Assessment. The Loss of Recognition of the Individual Offender*, 5 L. PROBABILITY & RISK 175, 187–89 (2006).

may aggrandize the seeming objectivity of sentences founded upon putatively impartial risk tools.[71] The preference is not just oriented toward quantifiable objectivity; potential ethical and normative benefits have been observed. Actuarial-based punishments may convey greater "moral certainty and legitimacy" than individual predilections and idiosyncratic judgments of individual decisionmakers.[72]

Largely as a result of the acceptance in the law of actuarial models to help inform various legal decisions (including in sentencing proceedings as just described), a cottage industry of actuarial tool developers and forensic evaluators has arisen and flourished.[73] The two most popular actuarial tools to be used in recent years for violent and sexual recidivism are outlined next.

B. Popular Actuarial Tools for Violent and Sexual Recidivism

The Violence Risk Appraisal Guide ("VRAG") is the best known actuarial tool for violence risk assessment[74] and the most researched in terms of replication and cross-validation.[75] VRAG was developed on samples of juvenile and adult patients released from a single maximum security psychiatric hospital in Canada.[76] The incidence of mental illness in the developmental samples is noteworthy. A significant proportion had been found not guilty by reason of insanity and many tested as psychotic.[77] Table 1 comprises the instrument's scoring sheet to illustrate the factors used and the weights they carry.

[71] Hyatt, *Reform*, *supra* note 3, at 729 ("The inclusion of impartial and empirical processes can help to subvert impressions of individualized bias and refocus the sentencing process on the offender's conduct and the characteristics that are most relevant to determining the risk to the community that they may pose.").

[72] Hannah-Moffat, *supra* note 9, at 276.

[73] *See* Monahan & Skeem, *supra* note 28, at 4–8 (discussing the adoption of risk and needs assessments in the criminal sanctioning systems of Pennsylvania, Virginia, and Utah).

[74] Jennifer L. Skeem & John Monahan, *Current Directions in Violence Risk Assessment*, 20 CURRENT DIRECTIONS PSYCHOL. SCI. 38, 39 (2011).

[75] Michael H. Fogel, *Violence Risk Assessment Evaluation: Practices and Procedures*, *in* HANDBOOK OF VIOLENCE RISK ASSESSMENT AND TREATMENT: NEW APPROACHES FOR FORENSIC MENTAL HEALTH PROFESSIONALS 41, 57 (Joel T. Andrade ed., 2009).

[76] Stephen D. Hart & David J. Cooke, *Another Look at the (Im-)Precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments*, 31 BEHAV. SCI. & L. 81, 81 (2013).

[77] Marnie E. Rice et al., *Validation of and Revision to the VRAG and SORAG: The Violence Risk Appraisal Guide—Revised (VRAG-R)*, 25 PSYCHOL. ASSESSMENT 951, 953 (2013).

Table 1. VRAG Scoring Sheet[78]

Criminal history score for nonviolent offenses prior to index offense: -2 = score 0 0 = score 1 or 2 3 = score 3 or above	*Age at index offense*: -5 = 39 or over -2 = 34–38 -1 = 28–33 0 = 27 2 = 26 or less
Failure on prior conditional release: 0 = no 3 = yes	*Lived with biological parents to age 16*: -2 = yes 3 = no
Victim injury: -2 = death 0 = hospitalized 1 = treated and released 2 = none or slight	*Marital status*: -2 = ever married 1 = never married
Any female victim: -1 = yes 1 = no	*Elementary school maladjustment*: -1 = no problems 2 = slight or moderate problems 5 = severe problems
Meets DSM criteria for any personality disorder: -2 = no 3 = yes	*History of alcohol problems (by count)*: · *Parental alcoholism* · *Teenage alcohol problem* · *Adult alcohol problem* · *Alcohol involved in index offense* · *Alcohol involved in prior offense* -1 = no boxes checked 0 = 1 or 2 boxes checked 1 = 3 boxes checked 2 = 4 or 5 boxes checked

[78] Vernon L. Quinsey et al., Violent Offenders: Appraising and Managing Risk 237–38 (1998). The Psychopathy Checklist is a multifactor psychological assessment used to rate psychopathy. The DSM is the Diagnostic and Statistical Manual of Mental Disorders.

Meets DSM criteria for schizophrenia: -3 = yes 1 = no	*Psychopathy Checklist score*: -5 = 4 or under -3 = 5–9 -1 = 10–14 0 = 15–24 4 = 25–34 12 = 35 or higher

Static-99 is the most widely used actuarial instrument to predict sexual recidivism.[79] The word "static" in the title highlights that the instrument depends on static, not dynamic, factors, while the "99" merely signifies the year—1999—the scale was introduced.[80] Static-99 was created using four different samples.[81] The first three samples were composed of sex offenders released from Canadian institutions: two samples were discharged from secure psychiatric institutions and one sample comprised offenders released from a maximum security prison.[82] The fourth included a sample of sex offenders released from a prison in England.[83] Table 2 provides the Static-99 scoring sheet.

[79] Daniel J. Neller & Giovanni Petris, *Sexually Violent Predators: Toward Reasonable Estimates of Recidivism Base Rates*, 31 BEHAV. SCI. & L. 429, 432 (2013); *see also* Astrid Rossegger et al., *Current Obstacles in Replicating Risk Assessment Findings: A Systematic Review of Commonly Used Actuarial Instruments*, 31 BEHAV. SCI. & L. 154, 155 (2013).

[80] R. Karl Hanson & David Thornton, *Improving Risk Assessments for Sex Offenders: A Comparison of Three Actuarial Scales*, 24 LAW & HUM. BEHAV. 119, 122 (2000).

[81] *Id.*

[82] *Id.* at 123–24.

[83] *Id.* A revision in the newer Static-99R creates additional categories for the age variable and new proportion tables, though the original version remains the popular version in use today. Leslie Helmus et al., *Absolute Recidivism Rates Predicted by Static-99R and Static-2002R Sex Offender Risk Assessment Tools Vary Across Samples: A Meta-Analysis*, 39 CRIM. JUST. & BEHAV. 1148, 1150 (2012); Rebecca E. Swinburne Romine et al., *Predicting Reoffense for Community-Based Sexual Offenders: An Analysis of 30 Years of Data*, 24 SEXUAL ABUSE 501 (2012).

Table 2. Static-99 Scoring Sheet[84]

Number of prior sex offenses: 0 = none 1 = 1–2 charges or 1 conviction 2 = 3–5 charges or 2-3 convictions 3 = 6 or more charges or 4 convictions	*Age at assessment*: 0 = 25 years or older 1 = between 18 and 25 years
Any convictions for a non-contact sexual offense: 0 = no 1 = yes	*Having lived with an age-appropriate intimate partner for 2 years*: 0 = yes 1 = no
Any convictions for an index non-sexual violence: 0 = no 1 = yes	*Any nonfamilial victims*: 0 = no 1 = yes
Any convictions for non-sexual violence before index offense: 0 = no 1 = yes	*Any stranger victims*: 0 = no 1 = yes
Number of prior sentencing dates: 0 = 3 or less 1 = 4 or more	*Any male victims*: 0 = no 1 = yes

Together, the VRAG and Static-99 remain the favored vehicles for statistics-derived predictions for violent and sexual reoffending. The present and potential future of the mathematical model of actuarial sentencing has now been established and explained. The next Section begins a critical analysis and these two instruments are a focal point.

[84] Hanson & Thornton, *supra* note 80, app. at 133–34.

III. JUDGING EMPIRICAL VALIDITY

Predictions about an individual defendant's level of risk might well be envisaged as an essential consideration for criminal sentencing in modern society. Policymakers, judges, and scholars staunchly promote actuarial assessments as best practices, representing the appropriate use of science in the law.[85] Notwithstanding the groundswell of support, there are strong reasons to question whether statistical risk models are adequately established for their use in such a critical area of criminal law as sentencing and punishment. The potential specter of unreliable science in the law calls for an analytical inquiry. Although actuarial evidence has been admitted in sentencing matters across the country to date, justice should not remain blind to its own potential errors in judgment. This Section outlines a variety of troubling issues—evidentiary, empirical, and pragmatic—with the use of actuarial assessments of risk in sentencing proceedings.

At its core, the introduction of actuarial assessment results in sentencing proceedings is an evidentiary matter. Certainly, the quality of evidence introduced in the law carries foundational importance.

> In our adversary system, the truth-seeking rationality goal of the rule of law forms the basis for evidentiary rules. The basic idea is that the methodologies of the justice system should have truth-generating capacity—a notion of due process. A second consequence of the aspiration to rationality is a concern for accurate evidentiary input: in order to reach a justifiable decision, courts must base reasoning on trustworthy information. A third consequence is that even trustworthy facts must have some logical tendency to prove or disprove an issue in the case. This framework for justice is the inspiration for the rules of evidence, and a fundamental tenet is that only facts having relevance—rational probative value—should be admissible in the search for truth.[86]

Notably, risk assessment results do not represent merely ordinary circumstantial evidence about a defendant's potential future behavior. Whether introduced through the testimony of a forensic clinician or via a presentence investigation report written by a probation officer, risk assessments are acting as, and accepted as, a form of expert evidence. Even though most probation officers would not likely be qualified as expert witnesses in forensic mental health evaluations, much less in actuarial risk assessment technologies, their scoring individual defendants on actuarial

[85] Redding, *supra* note 36, at 2–3.

[86] Erica Beecher-Monas, *The Epistemology of Prediction: Future Dangerousness Testimony and Intellectual Due Process*, 60 WASH. & LEE L. REV. 353, 356–57 (2003).

tools and deriving results fundamentally are being understood as grounded in the scientific method. Thus, it is appropriate to consider whether the actuarial risk tools for violent and sexual recidivism meet the high legal standards required for their admission as expert evidence. The initial question in this adventure concerns the relevance of the information.

A. Fitness

A primary hurdle for the introduction of any evidence in a legal proceeding is one of relevance. Also known as fitness, relevance requires that the proffered evidence should assist the trier of fact in understanding a fact at issue in the case.[87] Proponents of evidence-based sentencing advocate the use of actuarial risk tools as instructive for the utilitarian functions of sentencing. They presume that actuarial results are relevant to a factual determination of the individual defendant's future potential to cause harm. Unfortunately, such a premise may be naïve, even inimical to the interests of justice. For several reasons, the data and other information that current actuarial tools provide appear to be a poor fit for such purposes.

First, even promoters of evidence-based sentencing acknowledge that a key question is: measuring "the risk of what?"[88] Major goals of evidence-based sentencing practices include the ability to detect low risk defendants deserving short prison terms or potentially diverting them to community sanctions, while at the same time to sort out high risk defendants where preventive incapacitation might be justifiable. Presumably, the idea of risk for this purpose is not some unitary characteristic focused solely on an abstract likelihood of antisocial behavior sometime in the distant future. Instead, at least five different dimensions of risk are conceivably pertinent. Probability is one of them, but it may not even be as important as the other four. The additional dimensions of risk include imminence of antisocial acts, type of offense (e.g., violent/sexual/other, contact/noncontact, victim/victimless, child/adult victim), severity of harm, and frequency and duration of offending.[89]

In contrast to this more relevant multidimensional perspective on risk, developers of risk assessment tools generally have addressed only two dimensions. Many instruments count any illegal act, though the ones addressed more specifically herein at least differentiate violent and/or sexual recidivism from more general offending. Otherwise, the instruments tend to

[87] Daubert v. Merrell Dow Pharm. Inc., 509 U.S. 579, 591 (1993).

[88] Hyatt, *Reform*, *supra* note 3, at 743.

[89] Fogel, *supra* note 75, at 43.

operationalize recidivism as a simple dichotomous measure. Actuarial tool developers tally one recidivist as soon as any individual in the developmental sample committed a qualifying act during the period of observation.[90] Thus, actuarial tools likely count identically these two hypothesized individuals: (1) the sample subject who immediately upon release began a long-term crime spree involving heinous violent or sexual offenses which caused significant harm to a variety of victims, and (2) another sample subject who once attempted a noncontact sexual offense a decade after release. But when risk is a basis in a decision for preventive detention or probationary release, the important matters are the probability of some future harm *and* an understanding of the magnitude of the potential harm.[91] Clearly, the danger caused by these two hypothetical offenders is quantitatively and qualitatively disparate. Actuarial tools usually fail to differentiate. VRAG and Static-99, the popular risk tools highlighted herein, do not.[92] In sum, currently available risk tools are uninformative about much of what preferably should be a multifaceted picture of risk.

Second, the goal of identifying *low risk* offenders cannot, including from a scientific standpoint, be informed by current actuarial risk assessments. These tools have not directly, or even indirectly, been developed or modeled to detect non-recidivists or to predict desistance from reoffending.[93] Instead, developers generally have tested and chosen factors that were positively correlated with future recidivism.[94]

Pragmatically, it makes sense that risk tool developers have focused upon factors that can forecast recidivism rather than non-recidivism because violent and sexual recidivism are, contrary to popular belief, low rate events, except in extraordinarily high risk populations.[95] A fixation on positively predicting recidivism helps explain the absence in static risk tools of variables that would potentially be predictive of non-recidivism, such as dynamic factors (e.g., treatment successes, alcohol/drug abstinence, prosocial contacts), circumstantial factors (e.g., loss of opportunity, community services), or idiosyncratic variables (e.g., physically debilitating injury). Further, risk tools typically include a relatively small number of variables, thereby omitting a plethora of potential explanatory or correlative factors.

[90] Scales use differing definitions for recidivism, such as convictions, arrests, probation/parole violations, or self-reports. Instruments may or may not limit recidivism to serious types (such as felonies).

[91] Christopher Slobogin, *Prevention as the Primary Goal of Sentencing: The Modern Case for Indeterminate Dispositions in Criminal Cases*, 48 SAN DIEGO L. REV. 1127, 1135 (2011).

[92] *See* Rice et al., *supra* note 77, at 951; Rossegger et al., *supra* note 79, at 155.

[93] Craig & Beech, *supra* note 41, at 206.

[94] *Id.*

[95] *See infra* notes 158–62 and accompanying text.

Actually, the questions scored in the final models often constitute variables of convenience, items that evaluators will likely be able to score from available institutional or public files.[96] Thus, many individuals assessed in a purportedly "low risk" grouping may simply fall there because the tool used lacks those statistically significant factors that are otherwise relevant to them. Notice from Tables 1 and 2, for instance, that each of VRAG and Static-99, respectively, includes variables found to statistically correlate with violence recidivism that the other omits.

The third issue of fitness for sentencing decisions is specific to actuarial tools utilizing risk bins. Risk bins often classify groups in an ordinal ranking and use categorical labels; designations of low, moderate, and high risk are commonplace.[97] Yet these categorizations are meaningless except as a rather crude ranking system. Clinicians have no commonly agreed definition of risk categories,[98] statisticians have no accepted metric,[99] and there are no normative legal distinctions for such labels.[100] The categorical risk bin technique is merely a comparative and rhetorical device to differentiate the accumulation of risk factors amongst members of the relevant developmental sample. One particular study highlights this concept. Researchers scored a sample of sex offenders using five standard violence and sexual recidivism actuarial tools and found disparate uses of high and low risk labels.[101] The authors of the study explain:

> [W]hen we attempted to identify sub-samples of high and low risk offenders using the [five] instruments, common sub-samples were not identified. An alarmingly high number (55% of the sample) were identified by at least one instrument as being high risk; an alarmingly small proportion of the sample (3% and 4%, respectively) was identified as either high or low risk by all [five] instruments.[102]

Thus, these categorical labels have only relative meaning—not absolute value. This limitation is often ignored. Indeed, the use of such labels can have

[96] QUINSEY ET AL., *supra* note 78, at 143.

[97] J.C. Oleson et al., *Training to See Risk: Measuring the Accuracy of Clinical and Actuarial Risk Assessments Among Federal Probation Officers*, 75 FED. PROBATION 52, 53 (2011).

[98] *See* Daniel J. Neller & Richard I. Frederick, *Classification Accuracy of Actuarial Risk Assessment Instruments*, 31 BEHAV. SCI. & L. 141, 142 (2013).

[99] Jay P. Singh et al., *Rates of Sexual Recidivism in High Risk Sex Offenders: A Meta-Analysis of 10,422 Participants*, 7 SEXUAL OFFENDER TREATMENT 1, 183–84 (2012).

[100] J.C. Oleson et al., *supra* note 97, at 55.

[101] Howard E. Barbaree et al., *Different Actuarial Risk Measures Produce Different Risk Rankings for Sexual Offenders*, 18 SEXUAL ABUSE 423, 429–31 (2006).

[102] *Id.* at 437.

particularly problematic consequences in the law. A risk assessment could inappropriately subsume the standard of proof in law. If the decisionmaker presumes a label of "high risk" equates to meeting the burden of a "more likely than not" standard, the risk tool unfortunately appropriates the ultimate issue.[103] It is troublesome as well if the sentencer equates a score designated as "low risk" as being sufficient evidence under a preponderance of evidence standard to justify a less punitive or non-incarcerative sentence.

Fourth, actuarial tools are relatively unhelpful in the decision as to whether a defendant's sentence should include any period of incarceration. No standard or agreement, formally or informally, exists on the appropriate cutoff threshold for such a yes/no decision. Should the threshold for a decision on incarceration be linked only to a risk bin with a 100% estimated recidivism rate, or, more plausibly, is the threshold lower, such as 50% or 20%?[104] Or is a 5% probability reasonably sufficient to trigger a sentence involving incarceration? One might suggest the categorical rankings of, say, low, moderate, and high risk, could be useful in a jurisdiction with a policy of incarcerating only those at high risk. But, again, considering these labels have little meaning other than to rank order subgroups based on the developmental sample, reliance upon them for determining the need for incarceration remains a dubious lark at best.

Fifth, assuming the decisionmaker determines that a term of imprisonment is necessary, actuarial results fail to, directly or indirectly, assist in understanding how the length of a prison sentence will impact the risk of recidivism the tool projects.[105] Suggest the defendant's actuarial score is matched with a risk bin in which 25% sexually reoffended. This number provides no data about what length of incarceration would be helpful to prevent the hypothetical future crime from occurring. It bears mentioning, too, that proponents of the preventive incapacitation argument often lose sight of the fact that imprisonment is not entirely successful in preventing reoffending as prisoners commit crimes in prison (victimization of fellow prisoners and staff is not uncommon) or from prison (inmates have found ways to victimize the outside public).[106] In any event, by using actuarial

[103] Daniel A. Krauss & Nicholas Scurich, *Risk Assessment in the Law: Legal Admissibility, Scientific Validity, and Some Disparities Between Research and Practice*, 31 BEHAV. SCI. & L. 215, 226 (2013).

[104] Roffey & Kaliski, *supra* note 64, at 229.

[105] Starr, *supra* note 30, at 3 ("For example, if a judge is deciding between a one-year and a two-year prison sentence for a minor drug dealer, it is not very helpful to know that the defendant's characteristics predict a 'high' recidivism risk, absent additional information that tells the judge how much the additional year in prison will reduce (or increase) that risk.").

[106] *See* Nancy Wolff et al., *Sexual Violence Inside Prisons: Rates of Victimization*, 83 J. URB. HEALTH 835, 835 (2006); Rhonda Cook, *Inmates Extort Money from Outside Prison*,

scores to justify imprisonment at all, the scheme also tends to ignore the potential that such an outcome may be further endangering public safety since incarceration is itself often criminogenic.[107]

Alternatively, a plausible argument could be made that the information actuarial instruments is capable of providing is, in any event, unnecessary and improperly invades the province of the factfinder. Recidivism actuarial models rely heavily upon variables involving criminal history, prior social maladjustments, poor family relationships, and mental disorders.[108] Is it that unlikely that awareness of the existence of a relationship between those factors and future antisocial behavior is beyond the ken of judges and jurors?

In sum, actuarial risk tools for the assessment of violent and sexual recidivism appear to be poor fits to answer factual issues about future dangerousness in sentencing. This argument may not yet be convincing inasmuch as perceptions of future risk tend to be commonplace considerations in sentencing matters. Many supporters of risk tools concede some of these weaknesses yet still contend that at least the information obtained from actuarial scoring is better than nothing at all.[109] Consider these issues of fitness, though, along with the criticisms that follow regarding the failure of actuarial risk tools to comply with other prerequisites for expert evidence.

B. Validity & Reliability

A separate fundamental requirement for the admissibility of evidence in the law is that the information be sufficiently trustworthy, which, critically for expert evidence, requires that it be valid and reliable.[110] According to Supreme Court doctrine, for purposes of legal evidence, validity asks "does the principle support what it purports to show?" while reliability asks "does application of the principle produce consistent results?"[111]

AJC.COM (Jan. 1, 2013), http://www.ajc.com/news/news/inmates-extort-money-from-outside-prison/nTj4L/.

[107] *See generally* Martin H. Pritikin, *Is Prison Increasing Crime?*, 2008 WIS. L. REV. 1049; Lynne M. Vieraitis et al., *The Criminogenic Effects of Imprisonment: Evidence from State Panel Data, 1974–2002*, 6 CRIMINOLOGY & PUB. POL'Y 589 (2007).

[108] Oleson, *supra* note 29, at 1399 app.

[109] *See* Oleson, *supra* note 29, at 1397; M. Neil Browne & Ronda R. Harrison-Spoerl, *Putting Expert Testimony in its Epistemological Place: What Predictions of Dangerousness in Court Can Teach Us*, 91 MARQ. L. REV. 1119, 1211 (2008).

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

[111] *Id.* at 590 n.9.

1. Predictive Validity

In regards to actuarial assessments of future events, the requirement of validity is often expressed in the field of forensic sciences in terms of predictive validity.[112] A form of psychometrics, predictive validity represents the ability of the tool to accurately foresee the outcome of interest occurring.[113] Two empirical measures typify predictive validity: calibration and discrimination.[114] Calibration refers to the consistency between predictions and observed outcomes.[115] A well-calibrated tool for recidivism risk is one in which the average predicted recidivism rate is relatively equal to the actual rate of recidivism.[116] For example, a tool is well-calibrated if it predicts that 10% of persons classified in the moderate risk group will recidivate if the actual observed recidivism rate of the moderate risk group is about 10%. Discrimination determines how well a tool can differentiate those who experienced the outcome of interest from those who did not.[117] For violence risk tools, if those who recidivated with a violent offense all were scored at higher risk levels than those who did not, the tool discriminates perfectly. A high degree of discrimination does not require, or even signify, a well-calibrated instrument.[118] Thus, a scale can achieve a high rating for discrimination even when the average predicted risk of violent re-offense is significantly different than the actual percentage of violent recidivists.[119]

a. Discrimination

Despite the importance that a measurement of calibration should have on the acceptability of the tool to inform significant legal decisions, the relevant literature and discussion amongst experts have resorted to preferentially highlighting statistical results of discrimination tests in judging the competency of recidivism risk tools. This myopic focus on discrimination is empirically unsound and has likely led many proponents to overestimate the value of the current violence and sexual recidivism risk instruments. This assertion will be further explained, along with providing measures of

[112] *See, e.g.*, Jay P. Singh, *Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer*, 31 BEHAV. SCI. & L. 8, 8 (2013).

[113] *Id.*

[114] *Id.*

[115] N. Tollenaar & P.G.M. van der Heijden, *Which Method Predicts Recidivism Best?: A Comparison of Statistical, Machine Learning and Data Mining Predictive Models*, 176 J. ROYAL STAT. SOC'Y 565, 569 (2012).

[116] *See* Nancy R. Cook, *Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction*, 115 CIRCULATION 928, 928 (2007).

[117] Tollenaar & van der Heijden, *supra* note 115, at 569.

[118] *See* Cook, *supra* note 116, at 928.

[119] *See id.*

calibration, after an exploration of the levels of discrimination produced by the popular risk instruments.

Several statistical measures of discrimination for actuarial tools are available, yet one of them in particular has come to dominate the relevant literature. The discrimination indictor of popular choice is called the "area under the curve" ("AUC"), which is a fraction obtained from the receiver operating characteristic ("ROC") curve.[120] In scientific terms, the ROC curve is the "plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for every possible cut-off threshold."[121] Originally developed in the communication sciences, the ROC curve essentially is used to distinguish signal and noise.[122] Its utility is to display true positives (i.e., the signals) against false positives (i.e., the noise).[123] The ROC curve is a graphical representation.[124] The AUC is a fraction providing a statistical measurement of the ROC curve.[125] AUC values lie between 0 and 1, with .5 indicating discriminatory ability no better than chance and 1 indicating perfect discrimination.[126] As perfection is impossible to attain when forecasting human behavior, and actuarial tools would presumably not be published without achieving some statistically significant level of predictive ability, AUC values for recidivism risk tools typically lie somewhere in between .5 and 1.[127]

The correct interpretation of the AUC (for a recidivism risk tool) is "the probability that a randomly selected individual who committed an [act of recidivism] . . . received a higher risk classification than a randomly selected individual who did not" reoffend.[128] An AUC of .90, as an illustration, means that if one randomly chooses a recidivist and a non-recidivist, the recidivist's actuarial score would be higher than the non-recidivist's score about 90% of the time.[129] AUC fractions achieved by popular violence and sexual

[120] Paul R. Falzer, *Valuing Structured Professional Judgment: Predictive Validity, Decision-Making, and the Clinical-Actuarial Conflict*, 31 BEHAV. SCI. & L. 40, 43 (2013). Since its introduction in 1994, ROC testing is the dominant predictive validity diagnostic in violence risk assessment. *Id.* at 44.

[121] Singh et al., *Systematic Review*, *supra* note 40, at 64.

[122] Diler Aslan & Sverre Sandberg, *Simple Statistics in Diagnostic Tests*, 26 J. MED. BIOCHEMISTRY 309, 311 (2007).

[123] *Id.*

[124] *Id.* at 311 fig.2.

[125] Martin Rettenberger et al., *Prospective Actuarial Risk Assessment: A Comparison of Five Risk Assessment Instruments in Different Sexual Offender Subtypes*, 54 INT'L J. OFFENDER THERAPY & COMP. CRIMINOLOGY 169, 176 (2010).

[126] *Id.*

[127] *See id.*

[128] Singh et al., *Systematic Review*, *supra* note 40, at 64.

[129] Christopher T. Lowenkamp et al., *The Federal Post Conviction Risk Assessment (PCRA): A Construction and Validation Study*, 10 PSYCHOL. SERVICES 87, 92 n.11 (2013).

recidivism risk assessment tools vary by validation study and sample, but they commonly are reported in the range of .70 to .75.[130] Hence, these risk instruments have been able to classify violent and sexual recidivists at higher levels of risk than non-recidivists about 70 to 75% of the time.

Authors of studies investigating the discrimination ability of the popular recidivism risk tools often hype AUCS in the range of .70–.75 as representing moderate or large effect sizes.[131] An effect size is a generic term to represent the statistical magnitude of the phenomenon studied.[132] Yet these categorical descriptions are far more about improvement on chance than a clear barometer of statistical or practical significance.[133] In this area of statistics, there is no consensus on which numeric AUC scores represent small, moderate, or even large effect sizes. A comparative analysis of AUC effect sizes may be of interest. Authors reviewing a variety of violence risk assessment studies found great inconsistencies in reporting possible benchmarks for determining small, moderate, or large AUCs, even amongst studies citing the same sources.[134] In sum, the labeling of the discrimination ability of an actuarial tool as low or high is merely a social construct that is not only contested within the forensic science field, it does not itself offer sufficient evidence about the predictive ability of the tool.

Clearly, AUCs in the range of .70 to .75 offer discrimination abilities statistically better than chance (AUC of .50). But are they undeniable evidence of the predictive ability of actuarial risk tools sufficient for legal decisions which can have stark consequences to individuals and the public? A variety of empirical and practical reasons exist to conclude in the negative. Even with AUCs in that range, studies are showing a not insignificant occurrence of mistaken rankings. Erroneous rank ordering, then, occurs often, perhaps 25 to 30% of the time.

The AUC offers a rather limited perspective of predictive competence. To be clear, it is imperative for anyone using AUC as a diagnostic indicator to understand what the AUC value does not represent: it is not an accuracy index in terms of correctly predicting the actual occurrence of the outcome of interest; it does not signify the probability that individuals are scored correctly; nor does a high AUC score indicate the potential that a person assessed with a high test score will eventually become a recidivist.[135] Equally

[130] Singh et al., *Metaregression*, *supra* note 14, at 503.

[131] Hanson & Thornton, *supra* note 80, at 129.

[132] Ken Kelley & Kristopher J. Preacher, *On Effect Size*, 17 PSYCHOL. METHODS 137, 140 (2012).

[133] *Id.* at 138–39.

[134] Singh et al., *Systematic Review*, *supra* note 40, at 64.

[135] Cook, *supra* note 116, at 928.

important, the AUC statistic provides absolutely no information on the accuracy of any individual prediction as it is exclusively a group level statistic.[136] Instead, the AUC is simply an index of discrimination; it measures the tool's ability to rank order cases in the aggregate. An AUC can be far above .50 even if the tool is not well-calibrated (e.g., the percentage of predicted outcomes is significantly different than the proportion of actual outcomes).[137] Hence, it does not vouch for the tool's experience table of probabilities.

The exaggeration by many enthusiasts of risk tools in overemphasizing the AUC is partly due to its enigmatic character. This statistic is an inherently difficult concept. Alarmingly, evidence suggests even scientists conducting empirical tests of the predictive validity of recidivism risk tools often provide erroneous definitions of the AUCs calculated within their own studies.[138]

[136] Nilsson et al., *supra* note 37.

[137] Falzer, *supra* note 120, at 46. The following is an example of an instrument with poor calibration and perfect discrimination:

> If all recidivists in a sample had a risk of 10% (as calculated by the instrument to be validated) and all nonrecidivists a risk of 9%, the AUC value of the instrument in question would be 1 (= perfect discrimination), as in all the pairwise comparisons the recidivists would have a higher risk of reoffending than the nonrecidivists. However, the assessment of the risk of reoffending would be poor, because a recidivism rate of 100% is not to be expected for a group for which the calculated risk was 10%. Furthermore, the difference between a risk of 9% and 10% would be too small to be of importance in daily practice and would most likely be disregarded.

Astrid Rosegger et al., *Risk Assessment Instruments in Repeat Offending: The Usefulness of FOTRES*, 55 INT'L J. OFFENDER THERAPY & COMP. CRIMINOLOGY 716, 717 (2011).
Another example of the potential practical insignificance of a discrimination index is conjectured:

> In a prospective cohort that is considered generally low risk, such as many population-based cohorts, there may be a small proportion of individuals who are at high risk, with a preponderance of those at low or very low risk. Rank-based measures such as the [AUC] statistic do not take this distribution into account. Differences between [two] individuals who are at very low risk (eg, 1.0% versus 1.1%) have the same impact on the [AUC] statistic as [two] individuals who are at moderate versus high risk (eg, 5% versus 20%) if their differences in rank are the same.

Cook, *supra* note 116, at 929. A more rational reflection on AUC scores notes that "though the ratings or scores of violent persons are, on average, higher than those of non-violent persons (so that the probability of violence increases as the score increases), the score distributions of violent and non-violent individuals overlap considerably." Mossman, *supra* note 34, at 34. Such overlap means that even if risk tools achieve some success in rank ordering overall, this measurement of discrimination is weak evidence of its ability to correctly distinguish recidivists versus non-recidivists.

[138] Singh et al., *Systematic Review*, *supra* note 40, at 64 (finding erroneous AUC interpretations such as proportion of individuals who committed an antisocial act who received

One final observation about overreliance on a *discrimination* measure of predictive ability precedes an exploration of the more important measure of *calibration*. Authors of a meta-analysis have shown that all the prominent violence and sexual risk tools tend to achieve similar AUCs, even after controlling for differences in study design and random effects.[139] A commentator has referred to the common discrimination effect size as the "dodo bird verdict," meaning that each tool may have some minimal value, but none practically more than the other.[140] A suggested explanation for common discrimination effect sizes is the tendency among the recidivism risk tools to tap common historical factors, such as prior antisocial behaviors and poor socialization skills.[141] Moreover, experts contend that there is a natural limit to predicting human behavior and that actuarial technologies for recidivism risk have likely reached that limit already.[142]

b. Calibration

Calibration statistics arguably offer a superior benchmark for evaluating an actuarial instrument's predictive ability.[143] Calibration values exemplify a reliability dimension of the scale as well.[144] One of the major differences in the tests for calibration and discrimination is that discrimination measures ignore base rates, which is the frequency of a given outcome in the population of interest.[145] If 10% of a sample of sexual offenders were arrested for a new sexual offense within the period of observation, 10% would be the base rate of sexual recidivism for that sample. AUC measures ignore base rates. The AUC may be similar across samples with significantly different base rates as long as the instrument does an equivalent job of rank ordering. For instance, the VRAG was based on developmental samples with a combined 31%

higher risk scores than individuals who did not; proportion judged to be at high risk who committed an antisocial act; proportion whose outcome was correctly predicted; and probability a risk prediction would be accurate).

[139] Min Yang et al., *The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools*, 136 PSYCHOL. BULL. 740, 759 (2010).

[140] Pamela R. Blair et al., *Is there an Allegiance Effect for Assessment Instruments?: Actuarial Risk Assessment as an Exemplar*, 15 CLINICAL PSYCHOL. 346, 348 (2008).

[141] Jeremy W. Coid et al., *Most Items in Structured Risk Assessment Instruments Do Not Predict Violence*, 22 J. FORENSIC PSYCHIATRY & PSYCHOL. 3, 13–14 (2011); Yang et al., *supra* note 139.

[142] Monahan & Skeem, *supra* note 28.

[143] For a contrary analysis from the creators of VRAG, *see generally* Grant T. Harris & Marnie E. Rice, *Bayes and Base Rates: What Is an Informative Prior for Actuarial Violence Risk Assessment?*, 31 BEHAV. SCI. & L. 103 (2013).

[144] Ewout W. Steyerberg et al., *Assessing the Performance of Prediction Models": A Framework for Traditional and Novel Measures*, 21 EPIDEMIOLOGY 128, 129 (2010).

[145] Beecher-Monas, *supra* note 86, at 390 n.201.

violent recidivism base rate. Replication studies may achieve a high AUC even if the base rates of the replication samples were significantly higher or lower than 31%. Thus, the fact that replication studies may achieve AUCs in the range of .70 to .75 on very different samples (diverse jurisdictions, offender types, followup periods, type of recidivism, etc.) does not reflect that the same base rate of reoffending is consistent throughout. In fact, as will be shown later, base rates fluctuate greatly across different groups. Again, relative agreement on AUCs for the same risk tool just signifies some achievement on its rank ordering system.

Only very recently have a few researchers focused on computing and reporting calibration statistics for the most popular violent and sexual recidivism actuarial tools. This Article adds to this small body of research by calculating a few additional statistics which can be used to evaluate the predictive validity of the two most popular risk tools used today for violent and sexual recidivism. Calibration statistics are founded upon the calculation of a variety of measures, in the Bayesian probability tradition,[146] as listed and defined in Table 3.

[146] *See generally* Andreas Mokros et al., *Assessment of Risk for Violent Recidivism Through Multivariate Bayesian Classification*, 16 PSYCHOL. PUB. POL'Y & L. 418 (2010).

Table 3. Measures of Discrimination and Calibration

Measure	Definition
Sensitivity	The proportion of recidivists correctly predicted to recidivate.
Specificity	The proportion of non-recidivists correctly predicted not to recidivate.
True Positive Rate	The proportion of recidivists correctly predicted to recidivate. Also known as sensitivity.
False Positive Rate	The proportion of non-recidivists who had been predicted to recidivate. It is the reciprocal of specificity (1-specificity). Also known as false alarms and false positive predictions.
Positive Predictive Value	The proportion of people predicted to recidivate who were observed to recidivate.
Negative Predictive Value	The proportion of people predicted not to recidivate who are not observed to have recidivated.
Number Needed to Detain	The number of individuals judged to be at high risk who need to be detained in order to prevent a single incident of violence or sexual offense in the community.
Number Safely Discharged	The number of individuals judged to be at low risk who could be discharged prior to a single incident of violence or sexual offense in the community.

Unlike the discrimination index, calibration is concerned with the ability of the instrument to predict the actual occurrence of the outcome of interest. Here, the relevant outcome is a recidivist act involving violence or a sexual crime, depending on the instrument. Despite the earlier observation that sentencing decisions likely are interested in various dimensions of recidivism, current actuarial tools generally measure recidivism in a dichotomous manner. With this limitation, then, calibration measures consider the tool's predictive accuracy with respect to recidivism versus non-recidivism.

The calculation measures in Table 3 require the use of a cut-off point in which we designate all those scoring at or above the specified cut-off point as predicted to recidivate and all those below the cut-off point are predicted not to recidivate.[147] We then compare these to the number of recidivsts versus non-recidivists observed in the relevant sample using a 2 x 2 contingency table as illustrated in Table 4.

[147] Singh, *supra* note 112, at 10.

Table 4. 2 x 2 Contingency Table

Outcome

		Recidivist	Non-Recidivist	
Tool Result	Predicted to Recidivate	True Positives	False Positives	*Positive Predictive Value*
	Not Predicted to Recidivate	False Negatives	True Negatives	*Negative Predictive Value*
		Sensitivity	*Specificity*	

Table 5 contains calibration statistics calculated on the VRAG normed samples, while Table 6 provides calibration statistics calculated using data from the original Static-99 normed samples.

Table 5. VRAG (7-year followup)

Bin	n	r	P	TPR	FPR	PPV	NPV	NND	NSD
1	11	0	.00	100%	100%	31%	100%	--	--
2	70	6	.08	100%	97%	32%	100%	1	13
3	99	12	.12	97%	82%	35%	93%	2	9
4	117	20	.17	91%	62%	40%	90%	2	7
5	111	39	.35	80%	39%	49%	87%	2	5
6	95	42	.44	60%	22%	56%	81%	2	4
7	72	40	.55	39%	9%	65%	76%	3	3
8	34	26	.76	18%	2%	81%	72%	6	3
9	9	9	1.00	5%	0%	100%	70%	22	0

Table 6. Static-99 (10-year followup)

Bin	N	r	P	TPR	FPR	PPV	NPV	NND	NSD
0	107	12	.11	100%	100%	22%	100%	5	--
1	150	11	.07	95%	89%	23%	89%	4	8
2	204	27	.13	90%	73%	25%	91%	4	10
3	206	29	.14	79%	52%	29%	89%	3	8
4	190	59	.31	66%	31%	37%	88%	3	7
5	100	38	.38	41%	16%	42%	84%	2	5
6	129	58	.44	25%	8%	45%	82%	2	4

Legend: n = number in bin; r = number recidivated; p = proportion recidivated; TPR = true positive rate (sensitivity); FPR = false positive rate; PPV = positive predictive value; NPV = negative predictive value; NND = number needed to detain; NSD = number safely discharged; -- is used when there was a 0 in the denominator or nominator. NND and NSD numbers have been rounded up as it is not possible to either detain or discharge a fraction of a person.

To provide context for the tables, we can articulate some of the results. Let us first address VRAG. Assume a cut-off score of 7 as it is commonly designated as the beginning of the *contrived* "high risk" category. At the cut-off score of 7, sensitivity is 39% and specificity (1-FPR) is 91%, meaning we can expect that 39% of recidivists to be accurately predicted as recidivists and 91% of non-recidivists to be accurately classified as non-recidivists. The FPR indicates that of the non-recidivists, 9% were falsely predicted to recidivate. At a cut-off of 7, the PPV means that 65% of offenders in the development samples predicted to have reoffended were detected to have reoffended, while 35% predicted to reoffend did not. Thus, the prediction that anyone scoring in risk bin 7 or above would violently reoffend would be wrong 35% of the time. The NPV means that if we predicted that anyone scoring below 7 would not reoffend, we would be right 76% of the time. On the other hand, 24% of recidivists would have been missed.

It is important to be cognizant of the differences between sensitivity and specificity, on the one hand, and PPV and NPV, on the other. Sensitivity and specificity are retrospective in nature; the measures observe the recidivist and non-recidivist groups, respectively, and calculate the percentage that had been predicted to have recidivated or not recidivated, respectively. Sensitivity and specificity are calculated as the columns in the 2 x 2 contingency table (see Table 4). In contrast, the PPV and NPV are base rate dependent and are prospective in nature; the measures consider the groups predicted to recidivate and those not predicted to recidivate, respectively, and calculate the percentage that actually did relapse or did not, respectively. The PPV and NPV are calculated in the rows of the contingency table. Arguably, the PPV and NPV are the more important measures. For one, unlike sensitivity and specificity, PPV and NPV are calibration devices that account for differences in the base rates. For another, in sentencing we are more concerned with whether the actuarial instruments are sufficiently reliable to provide evidence in decisions based on predictions of *future* risk, and such decisions obviously occur prior to that outcome actually occurring. Sentencing, then, is prospective in nature in its relapse analysis. Consequently, the prospective true and false prediction measures appear more pertinent. The NND and NSD data points are also prospectively oriented.

Returning to the example of VRAG with a cut-off of 7, we find that the NND is three, which means that three individuals in VRAG's risk bin 7 and above would need to be detained in order to prevent a single incident of violence from occurring in the community. In contrast, the NSD of three means that three individuals with scores less than 7 could be discharged prior to a single violent incident occurring in the community. The NND and NSD represent moral constructs. One who is sympathetic to the number needed to detain criterion is in favor of detaining that number of offenders in preference for public safety, despite the fact that more individuals than necessary will be effected. In contrast, an NSD adherent would likely believe that detaining too many is unnecessary and injudicious, such that we should seek to release as many as possible to protect civil rights.[148] The NND and NSD are useful barometers in terms of making it even clearer that all of these statistical measures of discrimination and calibration do not exemplify objective numbers divorced from moral choices and ethical consequences. As an illustration, the authors of VRAG interestingly have asserted that "it can be reasonable for public policy to operate on the basis that a miss (e.g., failing to detain a violent recidivist beforehand) is twice as costly as a false alarm (e.g., detaining a violent offender who would not commit yet another violent offense)."[149] Others may at least as reasonably disagree on civil rights grounds and propose a contrasting perspective on the appropriate weighting of false positives and negatives.

The foregoing provided an articulation of the numbers in the VRAG table at just one cut-off point for illustration purposes. There are trade-offs for any chosen cut-off. A higher risk bin as the trigger would likely decrease the true

[148] *Id.*

[149] Harris & Rice, *supra* note 143, at 106.

positive rate, false positive rate, negative predictive value, the number needed to detain, and number safely discharged, while increasing the positive predictive value. Using a lower risk bin would have the opposite effects. The choice is also a moral and ethical one depending on whether one is more concerned with hits or misses.

Formulating a few exemplary statistics from the Static-99 grid in Table 6 may be helpful. If one is more concerned with protecting the public by reducing false negatives, then a lower risk bin would suffice. At risk bin 1, 11% of recidivists would have been missed (using NPV), while at bin 6, 18% would have been missed. If one is more interested in reducing false positives, then a higher risk bin would be of interest. In Static-99, using risk bin 0, 78% of predicted recidivists would have been false, whereas at risk bin 6, the likelihood of false positives is reduced to 55%. Still, risk bin 6 is the top category in Static-99, meaning that of those designated as high risk, half did not recidivate sexually. Static-99 produces a significant number of false positive predictions at its best.

Two issues should be obvious from these worked examples. The first is the significant degree of error rates with these risk scales. The second is the trade-offs that must be made. Not even the most conservative proponent is likely to opt to use preventive detention on the entire sample just to prevent any false negative. At the same time, the most liberal decisionmaker will presumably not advocate for the release of all just to eliminate the chance of incarcerating one false positive. Judgment calls are necessary as to where to weigh false positives and false negatives acceptably.

Tables 5 and 6 used the experience tables in the development samples for those identified scales. One may wonder if the meaningful failure rates for correct predictions are unique to the development samples. Perhaps the instruments perform better in the field? Other research has not supported this possibility. A meta-analysis of VRAG and Static-99 replication studies using new samples shows that at the deemed “high risk” cutoffs of 7 and 6, respectively, the average PPVs were 66% and 33%, respectively, meaning four out of ten false positives with VRAG and seven out of ten false positives for Static-99 in the high risk bins.[150] The alternative violent and sexual recidivism tools do not appear to perform any better.[151]

[150] Singh et al., *Metaregression*, *supra* note 14, at 507 tbl.4.

[151] Seena Fazel et al., *Use of Risk Assessment Instruments to Predict Violence and Antisocial Behaviour in 73 Samples Involving 24,827 People: Systematic Review and Meta-Analysis*, 345 BRIT. MED. J. 1, 10 tbl.3 (2012) (reporting averages from meta-analysis averages in sexual recidivism tool studies: sensitivity (88%), specificity (34%), PPV (23%), NPV (93%), NND (5), and NSD (14); and for violent recidivism: sensitivity (92%), specificity (36%), PPV (41%), NPV (91%), NND (2), and NSD (10)).

At least one advocate of actuarial tools in sentencing derides concerns with false positives. Judge Marcus refers to the false positive critique as a "thinking error" and "propaganda seeking to disparage the use of prison" for incapacitation purposes.[152] As for the "thinking error," he posits an example of a risk assessment tool identifying an offender as "presenting a 30% risk of violent recidivism. That only three of ten . . . will, in fact, commit a new violent crime within the contemplated period does not yield seven 'false positives.' The assessment of risk is by definition (in this hypothetical) precisely accurate."[153] Yet it is difficult to deny the existence of false positives. By incapacitating the ten offenders, seven will, by his own proposal, be unnecessarily impacted. He analogizes the scenario to an unexploded landmine.[154] Such a comparison also appears inapposite. At least with the unexploded landmine, the object is correctly singled out, the dangerous property is known rather than hypothesized, and incapacitating the landmine (presumably by dismantling or exploding it under controlled conditions) does not constitute an infringement on constitutional rights. A landmine is not a human being and enjoys no civil rights.

The next turn is to address the generalizability of empirical risk tools, though the discussion about predictive ability estimates will necessarily carry through the discourse.

2. Generalizability

Significant issues exist with any presumption that a recidivism assessment tool is generalizable outside of the tool's developmental samples. Human behavior is not only difficult to predict as a general matter, criminal acts and their correlates can vary dramatically across groups, times, geographies, environments, and circumstances.[155] Further, recidivism risk tools have generally incorporated variables found to be *associated with* reoffending; researchers did not intend to prove *causation*. The final variables are not, then, shown to be causal to human behavior. Therefore, the factors that were observed to correlate with recidivism in the developmental samples may not replicate to other groups, to other times, etc. "[T]here is no way to tell in the development sample how much of the observed relation between the variables and recidivism is due to underlying associations that will be shared in new samples and how much is due to unique characteristics of the

[152] Marcus, *supra* note 19, at 754–56.

[153] *Id.* at 754.

[154] *Id.* at 755.

[155] Keith Soothill, *Sex Offender Recidivism*, 39 CRIME & JUST. 145, 176 (2010).

development sample."[156] For these reasons, professional ethics require cross-validations before any risk assessment tool is used on any new group.[157] The following provides a good summary of suggested types of cross-validating factors:

> [T]he predictive efficacies of all tools must be eventually subjected to repeated empirical validation with client groups that differ in demographic characteristics (e.g., age, gender, socioeconomic status, ethnicity), level and type of past violence (e.g., criminal histories, sexual vs. nonsexual offenders), psychiatric diagnosis (e.g., presence of personality disorder, psychosis), intervention received (e.g., treated vs. untreated), the specific criterion being predicted (e.g., violent vs. nonviolent behavior or different types of violent behavior), environmental setting (e.g., clients residing in institutions vs. the community), countries of origin of the research, and so forth.[158]

As a result, prior to utilizing a risk tool on any group or individual, the evaluator's initial question should be whether the developmental sample(s) is sufficiently representative of the present group or individual to be examined. It may well not be. For instance, recall that VRAG's normative groups entirely comprised patients discharged from a maximum-security mental health hospital in Canada. Of the developmental samples totaling about six hundred, over two hundred had been adjudicated not guilty by reason of insanity and another one hundred were diagnosed psychotics.[159] The VRAG tool developers concede their intent was to create a risk instrument designed to assess serious offenders likely to have mental health problems in order for counseling professionals to craft appropriate psychiatric patient treatments.[160] Static-99 was also reliant upon significant percentages of forensic psychiatric patients in their developmental samples.[161] This means that the normed samples from these popular tools possessed quite unique group characteristics (e.g., significant numbers of mental disorders and mental health institutionalizations) that are quite unlikely to be shared by many other groups or in other settings. Plus, with these tools' developmental samples

[156] Vincent et al., *supra* note 33, at 81.

[157] STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING § 3.10 (Am. Educ. Research Ass'n, Am. Psychological Ass'n & Nat'l Council on Measurement in Educ. 1999).

[158] Yang et al., *supra* note 139, at 741. Local validation is important, too, as predictive variables of recidivism in a jurisdiction with abundant support services may vary from predictive measures in a jurisdiction without. Monahan & Skeem, *supra* note 28.

[159] Rice et al., *supra* note 77, at 953.

[160] QUINSEY ET AL., *supra* note 78, at 144.

[161] Hanson & Thornton, *supra* note 80, at 122–23.

being entirely Canadian and United Kingdom offenders,[162] they are unlikely to be representative of any group of routine sentencing defendants in the United States. Studies explicitly addressing the issue of differences between countries regularly find that the discrimination ability of actuarial recidivism risk tools for violence and sexual reoffending tends to be lower with samples in the United States as compared to samples in Canada[163] and the United Kingdom.[164]

The lack of representativeness renders the practice of reusing the proportionate estimates of recidivism from the developmental samples (the experience tables) a particularly egregious practice as a result. If the new group is not similar to the developmental sample, the developmental sample is not a representative reference for the individual to be assessed, or the base rates significantly differ, adopting such estimates is specious.[165]

Some studies purport to have cross-validated and upheld the use of the popular recidivism tools on new samples by accentuating that the study found a large effect size for the AUC.[166] Yet, recall that this statistic tells only part of the story about predictive ability. The AUC merely indicates if the instrument's relative ranking of risk conveys some degree of validity, not whether the probability of recidivism remains the same as compared to the developmental sample.[167] Thus, a critical aspect to judging the desirability of relying upon any risk tool's experience table is to either validate that the observed recidivism rates in the new sample appropriately replicates or, in the very likely case that it does not, to either decline to use the tool or perhaps to replace it with one appropriately normed to the new group. Unfortunately, neither option often occurs in practice, whether in clinical settings on in legal contexts.

The fundamental requirements for appropriately and ethically using an actuarial risk tool in real world situations are not merely hypothetical, theoretical constructs. Studies frequently show that base rates of violent and

[162] *Id.* at 122.

[163] Yang et al., *supra* note 139, at 754.

[164] R. Karl Hanson & Kelly E. Morton-Bourgon, *The Accuracy of Recidivism Risk Assessments for Sexual Offenders: A Meta-Analysis of 118 Prediction Studies*, 21 PSYCHOL. ASSESSMENT 1, 7 (2009) (reporting meta-analysis findings of AUCs for Static-99 in samples in the United Kingdom were much higher (average .90) than for U.S. samples (average .60)).

[165] Vincent et al., *supra* note 33, at 81; Nicholas Scurich & Richard S. John, *A Bayesian Approach to the Group Versus Individual Prediction Controversy in Actuarial Risk Assessment*, 36 LAW & HUM. BEHAV. 237, 238 (2012) (Asserting actuarial tools "should only be employed on reference classes similar to those on which such instruments are normed. Applying instruments to different samples/populations is likely to render such estimates spurious.") (citation omitted).

[166] *See generally* Hanson & Morton-Bourgon, *supra* note 164; Rice et al., *supra* note 77.

[167] Vincent et al., *supra* note 33, at 82.

sexual offending vary dramatically across samples.[168] This diversity in base rates underscores that actuarial tools which are developed on relatively small and potentially exceptional samples are unlikely to be representative, at least without local cross validation for both discrimination *and* calibration purposes. It is also important to recognize that criminal offending, violent and sexual offending in particular, is a cultural construct and the incidence and characteristics of crime can be experienced in quite disparate ways in different times, places, and circumstances.

In any event, VRAG's base rate for violent offending was 31% (at seven years) while Static-99's original base rate for sexual offending was 21% percent (at ten years).[169] There is overwhelming evidence that these numbers do not reflect representative base rates outside those samples, and that they are in most cases outliers. A recent meta-analysis of twenty-eight samples and over 6000 subjects, including significant numbers of psychiatric patients, in various countries found an overall recidivism rate for violence (broadly defined) of 25%.[170] Another meta-analysis of studies around the world reported an average recidivism rate for violent crimes of 20%, and an observed sexual recidivism rate of 12%.[171] Researchers reviewing multiple studies acknowledge the wide variation in sexual recidivism rates, observing that the summary statistic is "often in the 10% to 15% range."[172]

Local studies in the United States have found the sexual recidivism rate varying from 3 to 35%, though the upper end appears to be an outlier as it involved a presumably very high risk group in that the sample consisted of offenders being evaluated for sexual predator civil commitment.[173] Another

[168] *See* sources cited *supra* note 156, 157 and accompanying text.

[169] Mark E. Hastings et al., *Predictive and Incremental Validity of the Violence Risk Appraisal Guide Scores with Male and Female Jail Inmates*, 23 PSYCHOL. ASSESSMENT 174, 179 (2011).

[170] Yang et al., *supra* note 139, at 748 (ranging from 5 to 100%); *see also* Singh et al., *Metaregression*, *supra* note 14, at 506 (reporting meta-analysis overall recidivism rate of approximately 31%, inclusive of violent and nonviolent reoffending from 88 independent samples (*n*>5000), a large portion of which were psychiatric patients).

[171] Hanson & Morton-Bourgon, *supra* note 164, at 6 (basing average for sexual/violent recidivism on 50 samples (*n*=17,421), and sexual recidivism on 100 samples (*n*=28,757).

[172] Helmus et al., *supra* note 83, at 1149 (citations omitted).

[173] JILL S. LEVENSON & RYAN T. SHIELDS, SEX OFFENDER RISK AND RECIDIVISM IN FLORIDA 2, 8 tbl.3 (2012) (finding from a sample of 500 sex offenders released from Florida prisons rearrested for a sex crime a rate of 6% in five years and 14% percent after ten years and chronicling sexual recidivism rates of 4% and 7% in South Carolina; 7% and 13% in Minnesota; and 4% and 8% in New Jersey for 5 and 10 years, respectively); Marcus T. Boccaccini et al., *Field Validity of the Static-99 and MnSOST-R Among Sex Offenders Evaluated for Civil Commitment as Sexually Violent Predators*, 15 PSYCHOL. PUB. POL'Y & L. 278, 291 (2009) (finding recidivism rates for Texas sex offenders significantly lower than original and redeveloped STATIC-99 norms); Helmus et al., *supra* note 83, at 1149, 1154 tbl.1; Romine et al.,

meta-analysis yields interesting results. It combined studies of eight risk tools (including VRAG and Static-99), focusing on sexual recidivism rates for the groups that the instruments judged to represent "high risk" of sexual recidivism. The overall mean rate of sexual recidivism for those judged to be at high risk was 33%, with a range of 2 to 75%.[174] The meta-analysis authors concluded:

> One of the assumptions of these instruments is that groups classed as high risk will sexually recidivate at similar rates when sample size, time at risk, and setting are taken into consideration. The findings of the present study suggest that this assumption may not be evidence-based and that recidivism rates amongst those judged to be at high risk vary considerably both within and between instruments.[175]

Perhaps some worked examples will assist in conceptualizing the significance of base rate differences in altering predictive accuracy. I computed the positive predictive values of VRAG (Table 7) and Static-99 (Table 8) using lower base rates ("BR") than their developmental samples considering that most studies have tallied smaller rates of recidivism.

Table 7. VRAG Base Rate Change Impacts

Risk Bin	PPV with BR of 31%	PPV with BR of 20%	PPV with BR of 10%
1	31%	20%	10%
2	32%	20%	10%
3	35%	23%	12%
4	40%	27%	14%
5	49%	34%	19%
6	56%	41%	23%
7	65%	51%	31%
8	81%	71%	52%
9	100%	100%	100%

supra note 83, at 504, 506 tbl.1 (finding sexual recidivism rate of 14% (4% noncontact) in community sample of 744 Minnesota offenders).

[174] Jay P. Singh et al., *Rates of Sexual Recidivism in High Risk Sex Offenders: A Meta-Analysis of 10,422 Participants*, 7 SEXUAL OFFENDER TREATMENT 1, 6 (2012) (noting average follow-up of 81.4 months).

[175] *Id.*

Table 8. Static-99 Base Rate Change Impacts

Risk Bin	PPV with BR of 21%	PPV with BR of 10%	PPV with BR of 5%
0	22%	10%	5%
1	23%	11%	5%
2	25%	12%	6%
3	29%	14%	7%
4	37%	19%	10%
5	42%	23%	12%
6	45%	25%	13%

The second column in each table uses the base rate in the applicable instrument's development samples and, therefore, represents the tool's original experience table. I posited lower new base rates for Static-99 than VRAG as sexual recidivism occurs less frequently than violent recidivism (violent recidivism instruments often count sexual recidivism, as does VRAG). Let us use as an example VRAG's risk bin 7 (commonly deemed the "high risk" cutoff) where we predict that all offenders scored in risk bin 7 and above would recidivate. With the original base rate of 31%, the positive predictive value was 65%, meaning that of those scoring 7 and above predicted to violent recidivate, 65% did. This correspondingly represents that 35% would have been false positive predictions. Notice the significant drop in PPV statistics when the base rate declines. Using the same cut-off score of 7, the PPV declines from 65% to 51% and 31% with base rates of 20% and 10%, respectively. Thus, with a sample in which the base rate is 10%, using VRAG with a 7 cut-off score, 69% (seven out of ten) would represent false predictions of recidivism.

The loss in predictive value when positing more realistic sexual recidivism base rates with Static-99 is equally as dramatic. Using the top risk bin of 6 as representing "high risk" (according to the developers), the developmental samples' base rate yielded a positive predictive value of 45%. Lowering the base rates to 10% and 5% yielded PPVs of 25% and 13%, respectively. Hence, in a new sample in which the sexual recidivism rate is 5%, correct predictions of sexual recidivism using Static-99's highest bin, 87% would be false positive predictions. The use of 5% here is not just to make a point. It personifies a realistic sexual recidivism percentage. Conducting the most recent nationally representative sample to date, the Department of Justice tracked almost 10,000 sexual offenders released from prisons in the United

States in 1994 and calculated a sexual recidivism rate of 5% (at three years).[176]

The Static-99 developers have issued a revision, Static-99R,[177] with a new normed group which they call routine offenders, with a base rate of 6%.[178] One might then argue that if there is a revision with an updated experience table representing a more realistic base rate, evaluators should just use it as more likely representing a valid tool. The Static-99 developers actually do now suggest that Static-99R norms should replace the original. Nonetheless, the Static-99R's calibration index remains weak. At its best (at risk bin 9), the revised instrument earns a PPV of 33%, meaning two-thirds would be false positive predictions.[179]

As a result of deviations in base rates and sample composition, researchers commonly report concerning levels of diversity of the performance of risk tools.[180] As an example, a recent meta-analysis of studies using Static-99R found that the instrument performed disparately.[181] Across studies, the predicted recidivism rate for a Static-99R score of 0 varied from 0 and 19% (weighted average 5%), a score of 2 varied between 0 and 36% (weighted average of 9%), and score of 5 varied between 1 and 62% (weighted average of 18%).[182] The authors concluded that the predicted base rate fluctuations were likely due to the various impacts of disparities in "cohort effects (i.e., year of release), country, recidivism criteria, quality of recidivism information, offender type, or treatment participation" and the "density of unmeasured risk factors external to Static-99R."[183] The same meta-analysis found great variability from an alternative perspective. The underlying studies associated a predicted five-year sexual recidivism rate of 15% with Static-99R scores ranging from two to eight.[184]

Overall, then, the rather unconvincing calibration statistics produced herein provide confirming evidence that actuarial risk tools are too unreliable

[176] PATRICK A. LANGAN ET AL., U.S. DEP'T OF JUSTICE, RECIDIVISM OF SEX OFFENDERS RELEASED FROM PRISON IN 1994, at 1 (2003), *available at* http://bjs.ojp.usdoj.gov/content/pub/pdf/rsorp94.pdf.

[177] Romine et al., *supra* note 83, at 502–03.

[178] Singh et al., *supra* note 174.

[179] *See Static-99R, Observed and Estimated 5 Year Sexual Recidivism Rates for Static-99R: Routine Sample*, STATIC99 (Nov. 15, 2009), http://www.static99.org/pdfdocs/detailed_recid_tables_static99r_2009-11-15.pdf (using fixed follow up).

[180] *See, e.g.*, Helmus et al., *supra* note 83, at 1157.

[181] *Id.*

[182] Helmus et al., *supra* note 83, at 1164 (weighted averages reported in the text from fixed effects models).

[183] *Id.*

[184] *Id.*

for the purposes of critical criminal law decisions. This opinion is not meant as an indictment of the use in a legal context of VRAG and Static-99 exclusively. Similar issues in predictive ability would generally apply to the other currently available actuarial risk technologies. Concernedly, weak predictive ability plagues risk instruments that were validated on local samples. For example, the Virginia sexual offender recidivism tool used in its sentencing scheme yields a PPV of 57% at five years and 64% at ten years at the state's suggested cut-off point.[185] Hence, it produces about 40% false positives at the state's own official cut-off point. The most recent revision to Minnesota's sexual recidivism instrument (Mn-SOST 3.1) performs even worse: PPVs of 20% and 16% in its top 10% and 15% ranking categories, respectively, leaving 80% false positive predictions at the highest risk levels.[186] In sum, even though these instruments appear to correct the generalizability issue at least with respect to a geographic limitation, the great degree of false positives, four out of ten for Virginia and eight out of ten for Minnesota, reflect the tendency toward exceptional error rates.[187]

The gist of the evidence outlined herein is that the violent and sexual recidivism actuarial instruments appear unsound for use in routine sentencing cases in the United States. The immediate discussion pointed out issues with generalizability and what may be perceived as extreme error rates. So far the analysis has focused on the fit, validity, and reliability of actuarial risk instruments and drew on empirical and logical issues that should provide pause for their use in legal proceedings. The next issue to be addressed is the widespread misuse and erroneous interpretations of the abilities of risk assessments based on actuarial models.

C. Group-Based Statistics: The G2i Problem

Reliance upon actuarial tools to inform legal judgments presents an interpretive quandary that has been nicknamed "G2i."[188] The G2i problem

[185] *See* VA. CRIMINAL SENTENCING COMM'N, ASSESSING RISK AMONG SEX OFFENDERS IN VIRGINIA 89 fig.3 (2001), *available at* http://www.vcsc.virginia.gov/sex_off_report.pdf for PPV calculation data. Cutoff-point is 28 points. *Id.* at 92. The Virginia tool counts as recidivism any misdemeanor or felony crime against a person. *Id.* at 52.

[186] *See* MINN. DEP'T OF CORR., THE MINNESOTA SEX OFFENDER SCREENING TOOL-3.1 (MNSOST-3.1): AN UPDATE TO THE MNSOST-3, at 20 tbl.3 (2012), *available at* http://www.doc.state.mn.us/pages/files/large-files/Publications/MnSOST3-1DOCReport.pdf for PPV calculation data. This tool counts as recidivism reconvictions of hands-on sexual crimes. *Id.* at 9.

[187] *See* VA. CRIMINAL SENTENCING COMM'N, *supra* note 185, at 19–20.

[188] David L. Faigman et al., *Group to Individual (G2i) Inference in Scientific Expert Testimony*, 81 U. CHI. L. REV. 417, 417–18 (2014).

represents a basic disconnect between the scientific method, which operates by studying at the group level, and the law, which focuses on the individual case.[189] Translating from the population, being the group level—the "G" in G2i—to the individual case—the "i" in G2i—is a precarious adventure fraught with errors; but many judges, practitioners, even forensic assessors, fail to notice.[190]

Immanently, actuarial risk tools are scientifically designed at the aggregate level. Actuarial tools are not case studies focused on individuals, nor are they intended to incorporate idiosyncratic traits or qualities of any single person.[191] Whereas developers of actuarial instruments often choose factors that show statistically significant correlations or, alternatively, are statistically significant in regression models, rarely occurring variables naturally cannot achieve the requisite significance.[192] In the actuarial field for recidivism, the nature of study has been to build models for group-based predictions for reoffending, without attention to being able to predict which specific individuals in the group will relapse.[193]

Unlike the attention to generalizable knowledge that science pursues, legal decisions are interested in idiosyncratic traits (to the extent considered relevant) and in making individualized decisions.[194] Scientific studies may properly show that young, undereducated males are significantly more likely to commit violent acts, but in the law the prosecution must still prove beyond a reasonable doubt that this particular young, undereducated male committed the violent crime for which he is prosecuted. Similarly, while scientific studies may find positive correlations between sexual recidivism and variables regarding race/ethnicity, neighborhood, and sexual preference, presumably in sentencing we remain interested in the prosecution's burden to show this individual defendant poses a high risk of re-offense to justify a longer prison sentence.[195] Otherwise, the law is merely profiling in its criminal procedure decisions.

[189] *Id.* at 418.

[190] *Id.* at 420.

[191] *See* Vincent et al., *supra* note 33, at 82.

[192] *See id.*

[193] Nilsson et al., *supra* note 37, at 403.

[194] Fogel, *supra* note 75, at 45 (citation omitted).

[195] Still, a potential difference between these situations is that adjudicating guilt is a retrospective exercise whereas sentencing, at least to the extent it incorporates utilitarian concerns, is forward-looking where future predictions are involved. Others argue predictions of future behavior are always group-based thinking exercises. Eric S. Janus & Robert A. Prentky, *Forensic Use of Actuarial Risk Assessment with Sex Offenders: Accuracy, Admissibility and Accountability*, 40 AM. CRIM. L. REV. 1443, 1478–79 (2003).

A common G2i error is the presumption that group-based data allows for predictions at the individual level. Unfortunately, there is evidence in case law of just this sort of inaccurate attribution in sentencing proceedings. A sentencing opinion has described Static-99 as an objective tool "to predict the danger of future recidivism by [the defendant]."[196] Similarly, defense counsel in another sentencing case is quoted as referring to Static-99 as "a test which is employed and used to predict whether . . . *an individual* poses a risk of sexual assault to the public."[197]

But if group data essentially do not permit individual predictions, one might wonder how group-level data, i.e., nomethic data, are meant to be applied to individual predictions, i.e., on an idiographic level.[198] G2i methods normally operate through inferential reasoning. Usage of actuarial risk tools in clinical and legal realms typically relies on the rhetorical device of analogy, such as "[t]his man resembles offenders who were likely to recidivate, therefore he is likely to recidivate"[199] or some form of relative risk, such as "this offender is riskier than that offender."[200] Often, too, actuarial test results are conveyed in absolute terms, such as "based on the score of *x*, this defendant's risk for violent recidivism over the next *y* years is *z* percent."[201]

Certainly, some attention is appropriate in terms of understanding which type(s) of risk communication methods can best convey actuarial results to fact-finders in legal cases.[202] For purposes of informing legal decisions on appropriate sentences, the individualized and relatively straight forward examples just given are likely preferred by decision-makers. Notwithstanding such desire, these common forms of risk communication are scientifically and logically inaccurate and unfortunately obscure the limitations of using group-based study, which is at the core of the G2i problem.

[196] United States v. Adams, No. 09-2404, 2010 U.S. App. LEXIS 13074, at *116 (3d Cir. Feb. 11, 2010).

[197] People v. Delara, No. D057180, 2011 WL 5826080, at *22 (Cal. Ct. App. Nov. 18, 2011) (emphasis added) (internal quotation marks omitted).

[198] Nicholas Scurich et al., *Innumeracy and Unpacking: Bridging the Nomothetic/Idiographic Divide in Violence Risk Assessment*, 36 LAW & HUM. BEHAV. 548, 548 (2012).

[199] Hart & Cooke, *supra* note 76, at 82 (emphasis omitted) (internal quotation marks omitted).

[200] R. Karl Hanson et al., *Quantifying the Relative Risk of Sex Offenders: Risk Ratios for Static-99R*, 25 SEXUAL ABUSE 482, 484 (2013) (internal quotation marks omitted).

[201] *See* Daniel J. Neller & Richard I. Frederick, *Classification Accuracy of Actuarial Risk Assessment Instruments*, 31 BEHAV. SCI. & L. 141, 141 (2013).

[202] Nicholas Scurich & Richard S. John, *Prescriptive Approaches to Communicating the Risk of Violence in Actuarial Risk Assessment*, 18 PSYCHOL. PUB. POL'Y & L. 50, 52 (2012).

A cognitive error known as an ecological fallacy occurs when one attributes a group characteristic to any individual in the group.[203] Some properties of a group only reside at the aggregate level. For instance, researchers may have observed in the sample studied the occurrence of every type of sexual offense imaginable (e.g., adult rape, statutory rape, child molestation, bestiality, voyeurism, exhibitionism, child pornography viewing). But no one individual in the group is likely to have committed several of them, much less all of them. Thus, the occurrence of a wide variety of sexual recidivism offenses is merely an aggregate statistic; it would be fallacious to describe the study results as evidence that individuals tend not to specialize in their sexual reoffending.

Surely, the group level statistic that actuarial recidivism tools are perhaps most prized for is the proportional statistic tied to the relevant score or risk bin (e.g., 52% of those who scored 6 and higher sexually reoffended). Applying that group proportion to any individual is likewise an ecological fallacy and deceptive. Thus, the communication of risk in absolutist terms ("this defendant is 52% likely to sexually reoffend") is perhaps the worst offender in terms of correctly interpreting the aggregate statistics. Sentencing proceedings unfortunately exhibit a frequent use of risk assessment tools in just such a way. Experts in sentencing hearings have testified that a Static-99 score placed the defendant personally at "an 11% risk for sexual offense recidivism within [ten] years,"[204] or meant that the defendant "presented a 33[%] chance of sexual reoffending within five years, a 52[%] chance after ten years, and a 57[%] chance after fifteen years."[205] In a certification hearing of a juvenile to the adult system for adjudication and sentencing, another expert stated that, based on the VRAG, the juvenile defendant "presents a 48% risk (low to moderate) of recidivating in seven years and a 58% risk (moderate to high) of recidivating in ten years."[206]

As the last example reflects, it is regrettably common for assessors to impose categorizations of predictive risk directly onto individual defendants. For example, a state expert in one case testified that Static-99 "measured [d]efendant's risk for recidivism as low moderate."[207] A state judge sentenced

[203] Scurich et al., *supra* note 198, at 549.

[204] State v. Barnhart, No. OT-10-032, 2011 WL 5317301, at ¶11 (Ohio Ct. App. 2011).

[205] United States v. Adams, 385 Fed. Appx. 114, 116 (3d Cir. 2010).

[206] *In re* D.L.W., No. A12-1112, 2013 Minn. App. Unpub. LEXIS 114, at *6 (Minn. Ct. App. Feb. 11, 2013).

[207] State v. Fults, No. M2004-02092-CCA-R3-CD, 2006 Tenn. Crim. App. LEXIS 520, at *50–51 (Tenn. Ct. App. July 7, 2006); *see also* State v. Seward, 289 Kan. 715, 716 (2009) (denoting defense expert's "report stated that the Static 99 placed Seward in the medium-low risk category, with a 16 percent chance that he would reoffend in the next 15 years"); State v. Winters,

the defendant to a long prison term, citing the results of VRAG, which "placed him in a high risk to re-offend."[208] A presentence investigation report in another case stated: "The results from the Static 99 test suggested [this defendant] posed a 'high risk' for committing another sexual offense in the future."[209]

This type of attribution affirmatively reflects the G2i problem. To be certain, actuarial tools cannot now, or ever, technically operate as a sort of test of an individual's propensity. The ecological fallacy is particularly salient when the group-based study derived correlative factors that were not also shown to be causative. The creators of recidivism risk tools have not proven causation, in part because the tools are generally atheoretical.[210] Altogether, then, actuarial models cannot offer what many unfortunately presume they do, which is the ability to predict which individuals will reoffend. The developers of Static-99 admit that the fundamental attribute of their risk tool is not an absolute measure of risk in which the rate observed for the normed group can be applied to the person assessed; rather they concede their risk tool is designed to provide a relative risk measurement.[211]

Perhaps recognizing the same G2i issues, the literature accompanying the VRAG suggests the following form of relative risk communication in the following exemplary excerpt of a forensic assessment report of a hypothetical Mr. Moore: "Mr. Moore's category for risk of violent recidivism is in the eighth, or second highest, of nine categories. Among offenders in the [developmental] studies . . . , only four percent obtained higher scores, and approximately eighty-two percent in Mr. Moore's category reoffended violently within an average of 10 years after release."[212] This version helps explain the use of actuarial results in the comparative form previously suggested ("this man resembles offenders who were likely to recidivate, therefore he is likely to recidivate"). An example of this style of relative risk

No. 5-113 / 04-0575, 2005 Iowa App. LEXIS 147, at * 2–3 (Iowa Ct. App. Feb. 24, 2005) (scoring Static-99 placed defendant in medium/high risk category).

[208] State v. Gunderson, No. DC 07-0632, 2008 Mont. Dist. LEXIS 868, at *5 (Mont. Dist. Ct. Aug. 22, 2008).

[209] Guidroz v. State, No. 06-03-00239-CR, 2004 Tex. App. LEXIS 2872, at *2 (Tex. Ct. App. Apr. 1, 2004). A probation report in another case likewise stated defendant's "score of zero on the Static-99 assessment placed him at low risk of committing another sexual offense if released on probation." Jati v. Long, No. SACV 12-02073 GAF (AN), 2013 U.S. Dist. LEXIS 151048, at *46 (C.D. Cal. Aug. 13, 2013).

[210] Andrew John Rawson Harris & R. Karl Hanson, *Clinical, Actuarial and Dynamic Risk Assessment of Sexual Offenders: Why do Things Keep Changing?*, 16 J. SEXUAL AGGRESSION 296, 298 (2010) (conceding that with actuarial risk tools, "little attention is paid to the meaning or clinical utility of the risk factors" included).

[211] R. Karl Hanson et al., *supra* note 200, at 484.

[212] QUINSEY ET AL., *supra* note 78, at 357–58 (emphasis omitted).

communication can be found in at least one sentencing document. Scoring the defendant on Static-99 per the judge's order, a presentence investigation reports that the defendant "scored a [six] on this risk assessment. Individuals with these characteristics, on average sexually reoffend at 39% over five years, 45% over [ten] years and 52% over [fifteen] years."[213]

An analogous form of risk communication still has fostered erroneous interpretations, as the previous example illustrates. Lamentably, some academics are repeating this characterization that actuarial estimates provide average recidivism rates for offenders sharing the assessed individual's characteristics.[214] In other words, the assumption seems to be that offenders at each score or in each bin share common characteristics or histories. To the contrary, they may only share equivalent point totals. Because of the variety of factors available in the tools, study subjects may have received the same ending point totals based on completely different factors. To offer an example, two different people may share the same score where one received points on factors relating to criminal history, mental disorder, and trouble with alcohol, and the other for the recidivism predictors involving choice of victim, never being married, and young age. Thus, individuals assessed with the same resulting scores, or combined in the same risk bins, may share none or just a few of the same characteristics. The pair may be more dissimilar than similar.

The third common form of communication that can misdirect the sentencer is in the form of a relative risk assessment on a hierarchical scale (e.g., "this offender is riskier than that offender"). Indeed, interested parties concerned with the G2i problem suggest that a better approach is to conceptualize actuarial risk assessment as providing assistance in classification of different groups of offenders.[215] For example, an expert in one sentencing hearing testified that the defendant "scored in the lowest risk

[213] People v. Hillier, 910 N.E.2d 181, 184 (Ill. App. Ct. 2009).

[214] *See, e.g.*, Starr, *supra* note 30, at 806 (describing actuarial risk tools as "models [which] provide reasonably precise estimates of the average recidivism rates for the group of offenders sharing the defendant's characteristics." (emphasis omitted)); Beecher-Monas, *supra* note 86, at 410 ("The most that one can say for any actuarial risk assessment instrument is that it can give a probabilistic estimate of the level of risk for people who share characteristics with the person assessed.").

[215] CHRISTOPHER BAIRD, NAT'L COUNCIL ON CRIME & DELINQUENCY, A QUESTION OF EVIDENCE: A CRITIQUE OF RISK ASSESSMENT MODELS USED IN THE JUSTICE SYSTEM 3 (2009) ("Although [actuarial] models are frequently depicted as a means to predict which offenders will reoffend, actuarial risk assessment is more appropriately described in terms of classification. These systems simply apply group statistics to individual decisions to help agencies identify where they should focus their resources. In essence, these tools establish base expectancy rates for offenders who have different profiles.").

category relative to other adult male sex offenders,"[216] and a presentence report in another case documented that, "[b]ased on the Static 99 score this places [this defendant] in the high category or between the top 12% risk category relative to other male sex offenders."[217] This form of risk articulation shares the concerns just addressed for the other types of communication in terms of inexpertly using group-level statistics to adjudge an individual's chance of recidivating. Yet it raises another conceptual issue not yet discussed. The relative ranking to other persons may be practically meaningless without knowledge of the relevant base rate of recidivism. It seems necessary when understanding a relative risk of an outcome to factor in relative to what? If the base rate is 10%, a decision incorporating a risk estimate presumably would be very different than if the base rate is 50%, much less 80%.

In addition, a relative or ordinal categorical ranking may be particularly fraught with misestimations for sexual offenders. Studies consistently show that the public has an erroneous perception that sex offenders are highly likely to sexually reoffend and, as a consequence, overestimate actual rates.[218] Indeed, one study found the tendency for the public to dramatically overestimate the recidivism rate of typical sex offenders at nearly 75%,[219] despite the reality that, at least in the United States, recidivism for sexual offenders is a small fraction of that in most studies.[220] Thus, a relative or categorical risk communication comparing the defendant as higher risk than other sex offenders will likely yield a higher than necessary prediction.[221]

The G2i problem could well be conceptualized as another problem of fitness. The factual issue of a sentencing defendant's risk of reoffending is, or at least should be, an individualized inquiry. Arguably, we should not be overly interested in the average recidivism rate of the group of violent or sexual offenders, as applicable. Instead, the issue at hand is the future risk of the individual defendant at hand, who may vary from the average in ways not

[216] United States v. Robinson, 669 F.3d 767, 770 (6th Cir. 2012) (internal quotation marks omitted).

[217] *Hillier*, 910 N.E.2d at 184 (internal quotation marks omitted).

[218] Jorge G. Varela et al., *Same Score, Different Message: Perceptions of Offender Risk Depend on Static-99R Risk Communication Format*, 38 LAW & HUM. BEHAV. 418, 418(2014).

[219] Timothy Fortney et al., *Myths and Facts about Sexual Offenders: Implications for Treatment and Public Policy*, SEXUAL OFFENDER TREATMENT 1, 9 tbl.3 (2007); *see also* Stacey Katz-Schiavone et al., *Myths and Facts about Sexual Violence: Public Perceptions and Implications for Prevention*, 15 J. CRIM. JUST. & POPULAR CULTURE 291, 300 tbl.3 (2008) (reporting 98% of survey respondents answered affirmatively that "most sex offenders reoffend").

[220] *See supra* note 174 and accompanying text.

[221] Varela et al., *supra* note 218, at 419.

measured by the tool.[222] The inability of group-based statistics to provide predictions at the individual level, as just explored, make the actuarial tools rather unsuitable to answer such factual question. Hence, the arguments made in this Section perhaps have come full circle in a sense, yet lead to a new perspective based on purely legal considerations of evidence law. Are actuarial risk assessments too prejudicial, confusing, and/or misleading for the courtroom?

IV. PREJUDICIAL IMPACT

Empirical and interpretive challenges with statistically driven assessments plague the use of actuarial tools even in clinical environments. Concerns may appropriately be heightened further when they are offered in a legal context, particularly in such a critical proceeding as sentencing which necessarily involves public safety and fundamental deprivations of liberty and privacy. Sentencing individuals to potentially long-term periods of incarceration based on determinations of risk deserves circumspection. To this end, a commentator has suggested that risk assessment evidence which drives more punitive sentences ought to be subject to a stricter legal standard for admissibility.[223] At the same time, preferring certain defendants by reducing their sentences due to lower risk scores from actuarial tools demands consideration as it may cause unwarranted disparity among otherwise similarly-situated offenders, reduce the deterrence value of punishment, and needlessly endanger the public. Instead of caution, however, policymakers and judges seem impressed by the guise of empiricism, and in lieu of critiquing the fitness, validity, and reliability of risk tools, officials are more likely to reify them. This Section offers additional cautionary tales on the use and misuse of risk assessment results in sentencing matters.

A. Experts on Future Dangerousness

Despite multiple and significant weaknesses and misinterpretations, little evidence exists in sentencing law of risk scales being substantively or procedurally challenged. There exists Supreme Court precedent that supports, at least as a threshold matter, the admissibility of expert evidence about future dangerousness in sentencing proceedings. In the case styled

[222] *See, e.g.*, Mark S. Brodin, *Behavioral Science Evidence in the Age of* Daubert*: Reflections of a Skeptic*, 73 U. CIN. L. REV. 867, 911 (2005) (citing cases excluding expert testimony concerning reliability of average eyewitness as the issue is this particular eyewitness' reliability).

[223] Hannah-Moffat, *supra* note 9, at 286.

Barefoot v. Estelle,[224] the Supreme Court addressed the admissibility of certain evidence in a case in which a jury sentenced a capital defendant to death upon a finding that "there is a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society."[225] Barefoot objected to the state's offer of psychiatric expert witnesses to testify about his future dangerousness potential.[226] His challenge was not an evidentiary one per se, but an argument that psychiatric opinion concerning future risk was so unreliable that this type of evidence would produce arbitrary sentences in violation of the Eighth Amendment's Cruel and Unusual Punishment Clause.[227] Interestingly, the American Psychiatric Association ("APA") submitted an amicus brief in support of the defendant's position, declaring psychiatric testimony on future risk could not be reliable and that, in the organization's estimate, two out of three predictions by psychiatrists of long-term future dangerousness were erroneous.[228] Although acknowledging the APA's position, a six-justice majority nevertheless ruled against the defendant.[229] According to the majority, even the APA did not assert that psychiatrists were always wrong and, though many psychiatrists contested the reliability of such predictions, other doctors remained willing to testify and to give their professional opinions about a defendant's future risk of violence.[230]

The *Barefoot* majority appeared to be concerned with a sort of contagion effect. The majority opinion expressed disquiet about the possibility that if expert evidence was ruled inadmissible in capital cases, the whole idea of future dangerousness as a proper criterion in sentencing decisions generally would be in peril.[231] The Court likewise noted that since the state's capital sentencing statute required juries to make this type of factual determination of future risk, jurors should at least get some external assistance.[232] Besides, the majority ruled, any "shortcomings" in expert judgments about future risk could effectively be evaluated during the adversarial process.[233]

The *Barefoot* decision is not necessarily dispositive here. *Barefoot* concerned unstructured clinical judgment (understandably, as the case preceded the development of actuarial risk tools), with the majority endorsing

[224] 463 U.S. 880 (1983).

[225] *Id.* at 884.

[226] *Id.* at 885.

[227] *Id.*

[228] *Id.* at 920 (Blackmun, Brennan and Marshall, J. dissenting).

[229] *Id.* at 906 (majority opinion).

[230] *Id.* at 899–901.

[231] *Id.* at 896 (precluding factfinding on future risk akin to "disinvent[ing] the wheel").

[232] *Id.* at 896–97.

[233] *Id.* at 899.

the testimony of one of the psychiatric witnesses who claimed an ability to make an expert assessment "if given enough background information."[234] This suggests the majority may have only been approving expert opinions based on a holistic review from a comprehensive clinical evaluation. The ruling may not extend to the far more limited review required by current actuarial tools. Indeed, in a later decision, the Supreme Court questioned, albeit in dicta, the reliability of an expert's opinion on future risk if it followed merely a "cursory" review of the individual defendant and his circumstances.[235]

Another reason to potentially distinguish *Barefoot* is on the question of the expert's qualifications. The discussion in Section III should have given the impression that risk assessment technologies are far from intuitive devices. Their foundational methodologies are abstruse and evaluators face challenges in correctly interpreting and communicating results. Christopher Slobogin is a supporter of the use of actuarial risk technologies in the law, yet accepts that to be qualified as an expert to render actuarial evidence, the person must understand the underlying statistical techniques and have access to the specialized knowledge on which the tool is founded.[236] He rightly explains that any potential expert witness should be able to articulate the methodology used in constructing and validating the tool as

> it is unlikely that a layperson would understand the significance of a finding, say, that someone who belongs to a group with a base rate for violence of ten percent has a forty percent chance of recidivating within a given period of time, without some explanation of the significance of base rates, false positives, and follow-up periods.[237]

The experts at issue in *Barefoot*, though giving clinical opinions rather than scoring an actuarial instrument, at least were presumably well-educated psychiatrists knowledgeable about the field of forensic mental health and conversant in clinical risk appraisal methods. Today, many of those in the sentencing world who are scoring and interpreting actuarial results are evidently not so qualified. Quite likely, numerous evaluators are woefully unqualified in the relevant areas of expertise. Regarding one such group, multiple states now permit, even require, probation officers to routinely include actuarial assessment results in presentence reports.[238] One might argue that scoring an actuarial tool itself is a simple task requiring no special skills or education. To the contrary, the tools are not easily scored, some of

[234] *Id.* at 899 n.7.
[235] Ford v. Wainwright, 477 U.S. 399, 415 n.3 (1986).
[236] Christopher Slobogin, *Dangerousness and Expertise*, 133 U. PA. L. REV. 97, 137 (1984).
[237] *Id.*
[238] *See supra* note 26.

them require mental disorder diagnoses, and many factors necessitate judgment calls.[239] Perhaps a few examples will suffice. VRAG includes factors such as a diagnosis of a personality disorder, a complicated Psychopathy Checklist evaluation, and a rating for elementary school maladjustment. All known recidivism risk tools require adjustments for a variety of criminal history variables that often remain vaguely constructed and defined. On the whole, insufficient attention is being paid to whether the experts testifying in sentencing proceedings, or pseudo-testifying through a backdoor method of incorporating risk scores and interpretations via presentence reports, are properly qualified to score actuarial scales or to intelligently explain the methodological attributes as Professor Slobogin suggests.

B. Judges as Gatekeepers

A number of researchers in the mental health field now voice skepticism about the scientific value of actuarial risk. The authors of a recent meta-analysis of violence risk assessment tools observe that the prediction of violence is one of the "most complex and controversial issues in the behavioral sciences" and the significant problems and discrepancies in risk assessment practices that their study revealed led to their conclusion that actuarial tools should not be the sole or primary basis for criminal justice decisions.[240] Other experts allege that, with the base rate of violence so low, "for the foreseeable future, no technique will be available to identify those who will act violently that will not simultaneously identify a large number of people who would not."[241] In addition, researchers in a separate meta-analysis of risk assessment tool studies comment: "One implication of these findings is that, even after 30 years of development, the view that violence, sexual, or criminal risk can be predicted in most cases is not evidence based."[242]

Should sentencing judges defer to actuarial risk as a mere policy choice or must it be subjected to normal evidentiary standards of the law? Bernard Harcourt laments that "[w]hat we have done, in essence, is to displace earlier

[239] Oleson, *supra* note 29; *see supra* Tables 1–2.

[240] Yang et al., *supra* note 139, at 740, 761.

[241] Alec Buchanan et al., *Resource Document on Psychiatric Violence Risk Assessment*, 169 AM. J. PSYCHIATRY 340 (2012 Supp.).

[242] Fazel et al., *supra* note 151, at 5; *see also* Rettenberger et al., *supra* note 125, at 183 (concluding from study of actuarial tools (including Static-99): "One major aim of most criminal justice systems is to calculate risk by predicting the probability of severe sexual crimes. This goal obviously is not yet achieved satisfactorily by actuarial risk assessment, because results are far from ideal" and should probably be considered only as part of a broader clinical assessment.").

conceptions of just punishment with an actuarial optic."[243] The steadfast reliance upon actuarial instruments may simply be pragmatic. The policy of using risk tools in sentencing decisions represents a sort of "better safe than sorry" approach[244] that elevates public safety over individual liberty, conveying the political willingness to withstand false positives over false negatives. The political advantage is evident as "false negatives engender political opprobrium and false positives go virtually undetected"[245] with preventive incapacitation.

Judges are expected, though, to be apolitical and independent arbiters of truthful evidence. They have been tasked to act as gatekeepers interested in excluding unreliable science otherwise disguised as expert evidence. Nonetheless, scant evidence exists of courts restricting, much less questioning, actuarial risk assessments in sentencing proceedings. As a general matter, rules of expert evidence are often now ignored in sentencing. A legal commentator mourns that sentencing hearings have become "an evidentiary free-for-all."[246] Courts in at least nineteen states have expressly ruled that the states' evidentiary admissibility standards do not apply to expert testimony based on structured risk assessments or, if they do apply, most tools are deemed, with little or no review, to meet the appropriate standard.[247] Researchers reviewing the use of VRAG results in American courts concluded, "it is clear that on whole the courts accepted the findings of the risk assessment instruments."[248]

Numerous proponents of actuarial tools in sentencing concede some of the empirical problems, but contend that the answer is for the adversarial process to flesh out any issues or concerns on behalf of the factfinders and/or grant defendants access to their own professional risk experts.[249] This argument appears consistent with the *Barefoot* opinion advocating that any battle be waged in the courtroom. Yet counsel have generally been unwilling, unable, or too enamored of the scientific cloak to use the adversarial questioning process to critically examine risk tools, their underlying methodologies,

[243] HARCOURT, *supra* note 11.

[244] Nilsson et al., *supra* note 37, at 405–06.

[245] Scurich & John, *supra* note 202, at 58.

[246] Beecher-Monas, *supra* note 86, at 357.

[247] Krauss & Scurich, *supra* note 103, at 220.

[248] Michael J. Vitacco et al., *The Role of the Violence Risk Appraisal Guide and Historical, Clinical, Risk-20 in U.S. Courts: A Case Law Survey*, 18 PSYCHOL. PUB. POL'Y & L. 361, 383 (2012).

[249] Michael H. Marcus, *Conversations on Evidence Based Sentencing*, 1 CHAP. J. CRIM. JUST. 61, 105 (2009); Pari McGarraugh, *Up or Out: Why "Sufficiently Reliable" Statistical Risk Assessment is Appropriate at Sentencing and Inappropriate at Parole*, 97 MINN. L. REV. 1079, 1109 (2013).

issues of generalizability, or the true meanings of the G2i interpretations provided.[250] In terms of the latter, a commentator regrets that the "near complete failure to even mention problems of determining individual risk from group data is, perhaps, the single greatest blot on the majority of risk assessments presented to the courts in all jurisdictions."[251] In this respect, perhaps it is not entirely the fault of judges for allowing in this type of evidence. The law maintains

> a requirement that evidence be cogent. This requires that the limitations of [risk] assessments be iterated and subjected to judicial scrutiny. Alarmingly, demonstrable limitations of risk assessments and the instruments or techniques on which they are based are all too often simply ignored by forensic practitioners of various persuasions, if they are comprehended in the first place. And so the courts are denied the very information they should be provided with when considering the prognostications of these practitioners. This lacuna must be remedied to prevent errors in the investigatory processes being relied upon and hence perpetuated in the adjudicative phase with the result that miscarriages of justice are all but guaranteed to occur when preventative detention and supervision of sexual and violent offenders are mooted.[252]

Two notable exceptions exist. A pair of prominent federal judges has publicly questioned actuarial evidence of risk in sentencing proceedings. Judge Posner of the Seventh Circuit in a written opinion chastised the litigants for not even enquiring about the offered Static-99 evidence.[253] He further opined that even though actuarial assessment "may be more accurate than clinical assessments, . . . that might not be saying much."[254] Judge Jack Weinstein of the Eastern District of New York has done a commendable job of critically assessing Static-99, even examining the experts himself in lengthy testimonial exchanges, citing in his written opinion numerous forensic science publications, and listing the instrument's flaws.[255] Among

[250] Browne & Harrison-Spoerl, *supra* note 109, at 1211.

[251] Ian R. Coyle, *The Cogency of Risk Assessments*, 18 PSYCHIATRY PSYCHOL. & L. 270, 274 (2011).

[252] *Id.* at 271–72.

[253] United States v. McIlrath, 512 F.3d 421, 425 (7th Cir. 2008). Judge Posner bemoans that the litigants not only failed to investigate the Static-99 outcome's validity as applied to the individual defendant, they failed even at a most basic level to identify it: "We are not even told what 'Static-99' is." *Id.* at 424.

[254] *Id.* at 425. Though, even Judge Posner makes an error in conceptualizing Static-99 as requiring a conviction to count recidivism. *Id.* at 425. Instead, one of the four samples operationalized recidivism to include arrests or readmissions to the psychiatric institution. Hanson & Thornton, *supra* note 80, at 123 tbl.2.

[255] United States v. C.R., 792 F. Supp. 2d 343, 445–66 (E.D.N.Y. 2011).

Judge Weinstein's astute observations are that it "is essential in using risk assessment tools to consider the appropriateness of the population used to create the base for assessment,"[256] it is inappropriate to use a risk tool on a defendant for whom it was not normed,[257] use of the instruments may still reflect biases of the evaluator,[258] and excessive reliance upon them is a real and serious concern.[259]

To be sure, the issue herein is not whether the use of risk assessment tools for violent and sexual recidivism constitutes poor science in any holistic sense. The standards of law and science are not synonymous.[260] The error rates underlying risk tools may be acceptable to scientists or in clinical settings, while at the same time too high in a legal context.[261] VRAG and Static-99, for instance, may be completely acceptable in a mental health situation where the results may be part of a broader clinical assessment on which a psychologist will base a treatment plan. Or the risk scores may assist a psychiatrist in determining which patients potentially can be safely transferred to a less secure area of a mental health institution. The argument here does not intend to infringe upon those experts' balancing of interests in their own professional settings. The law can learn from, and embrace, knowledge from the sciences, but the law also still stands on its own merits. Particularly when the stakes are so high, an independent weighing by legal minds should be perforce. The overall thesis, then, is that the actuarial risk tools for violent and sexual recidivism are too unreliable for the purposes of sentencing decisions, with the confusion about them amongst the experts themselves as further confirmation thereof.

Importantly, some proponents argue that actuarial scores and their corresponding rankings and/or probabilistic interpretations do not need separately to meet legal standards of admissibility because they constitute

[256] *Id.* at 449.

[257] *Id.* at 446.

[258] *Id.* at 462.

[259] *Id.* at 461.

[260] Frederick Schauer, *Can Bad Science be Good Evidence? Neuroscience, Lie Detection, and Beyond*, 95 CORNELL L. REV. 1191, 1219 (2010) ("[T]he evaluative standard to be used by the law, even when it is science that is being evaluated, must be based on law's goals, law's purposes, and law's structures, and as is so often the case, what is good outside of law may not be good enough inside it.").

[261] *Id.* at 1214 ("Science can tell us that a certain scientific process has, say, a 12 percent error rate (or specific rates of Type I and Type II errors or false positives and false negatives). And scientists must decide for their own scientific purposes whether such rates are sufficient, for example, to assert that something is the case, conclude that a finding is adequate for publication, or find a research program promising enough to renew a research grant. But whether such an error rate is sufficient for a trier of fact to hear it, put someone in jail, keep someone out of jail, justify an injunction, or award damages is not itself a scientific question.").

just one piece of information in a multidimensional decision.[262] This framing appears too simplistic and dismissive. The notion that unreliable science (even junk science) should somehow be protected because it might constitute simply one source of information in a multi-factor decision should offend any strong adherent to the principles of law and the desire to admit only truthful evidence. A plethora of other independent authorities would claim to have the knowledge and ability to predict future behavior and would honestly assert a conviction that the foundations of those predictions lie in science and based on reliable methods. Envision astrologists, numerologists, and palmists who purportedly predict the future through objective and standardized means. Consider those trained in psychology and psychiatry who have in our history promoted prognostications of antisocial behavior founded on such "scientific" theories as phrenology, physiognomy, and somatotypes. The "only one piece of evidence" rationalization would admit as expert evidence each of them.

Besides, even if actuarial risk assessment has some minimal relevant value or represents merely a morsel of external aid to assist in complicated decisions, an additional query should be whether it will do more harm than good? Federal Rule of Evidence 403 states: "[t]he court may exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, [or] misleading the jury . . . ."[263] It seems at least reasonable to conclude these risk assessments are overly misleading, confusing, and prejudicial.

C. Science is Fallible

Information offered as expert evidence and portrayed as founded upon the scientific method is necessarily accorded a higher status in the minds of recipients.[264] The use of the word "expert" deploys all of the positive and

[262] *E.g.*, Michael A. Wolff, *Evidence-Based Judicial Discretion: Promoting Public Safety Through State Sentencing Reform*, 83 N.Y.U. L. REV. 1389, 1404 (2008) (framing it as informed sentencing via risk assessment tools); *see also* Browne & Harrison-Spoerl, *supra* note 109, at 1212 (arguing judges benefit from risk assessment experts with different professional opinions); Schauer, *supra* note 260, at 1205 (contending scientific evidence need not be shown reliable beyond a reasonable doubt where it accounts for one piece of evidence in decisions on incarceration).

[263] FED. R. EVID. 403.

[264] Hyatt et al., *Integrate*, *supra* note 19, at 267 ("Recent advances in the science and statistical methodologies of prediction have allowed higher degrees of automation for actuarial risk forecasting than ever before."); McGarraugh, *supra* note 249, at 1105–06 (noting actuarial tools developed by scientists through empirical methods); Redding, *supra* note 36, at 4, 7 (noting actuarial "approaches rely on empirically identifiable criminogenic risk and protective factors and/or scientifically validated tools for assessing those factors, providing a quantifiable prediction

superior connotations the English language has given it: "[o]ne whose special knowledge or skill causes him to be regarded as an authority; a specialist."[265] In the law, expert testimony is given an authoritative and privileged status. Such a perception is regrettable here. The abilities of actuarial tools can be misleading in being draped in the guise of empiricism, with its attendant "aura of scientific infallibility."[266] Further, the results of risk assessments here are reified in sentencing decisions as issuing from objective calculations completed in a scientific "test."[267] This conceptualization of actuarial instrument predictions is misleading and likely accounts for the few questions or criticisms being raised in legal cases about the significant rates of error, critical lack of generalizability to American sentencing populations, and erroneous interpretations of actuarial results, all as discussed herein. "Science" has obscured what is really a matter of art. To foretell a person's future antisocial actions, a heavy dose of imagination is intrinsically required. The connotation of the descriptor of "scientific tests" masks both reality and common sense, luring the audience into forgetting about the inherent incompetence to effectively predict human behavior, especially in the long-term.

The entirety of Section III herein should establish how actuarial instruments and their resulting interpretive results confuse factual issues regarding future risk. Supplemental support exists. A growing body of research from the forensic mental health field further illustrates just how confusing (and manipulable) communications of risk are to factfinders. Researchers in a trio of studies found that judges and mock jurors believed the categorical format of communication (high versus low) to be more probative and led them to rate more offenders at higher likelihoods of reoffending than when the risk format used numerical statistics.[268] The

of risk (stated in probability or categorical format)" and representing a "scientific assessment of the offender's recidivism risk"); *but see* Margareth Etienne, *Legal and Practical Implications of Evidence-Based Sentencing by Judges*, 1 CHAP. J. CRIM. JUST. 43, 60 (2009) (using actuarial risk tools, "judges must beware of the ability of science to seduce them out of judging").

[265] "*Expert" Definition*, OXFORD ENGLISH DICTIONARY (2d ed. 1989), *available at* www.oed.com/.

[266] Jack B. Weinstein, *Scientific Evidence in Complex Litigation*, *in* ALI-ABA COURSE OF STUDY: TRIAL EVIDENCE, CIVIL PRACTICE, AND EFFECTIVE LITIGATION TECHNIQUES IN FEDERAL AND STATE COURTS 709, 723 (1991).

[267] *E.g.*, United States v. Zobel, 696 F.3d 558, 565 (6th Cir. 2012); United States v. Robinson, 669 F.3d 767, 770 (6th Cir. 2012); State v. Barnhart, No. OT-10-032, 2011 WL 5317301, at ¶12 (Ohio Ct. App. 2011); People v. Delara, No. D057180, 2011 Cal. App. Unpub. LEXIS 8881, at *63 (Cal. Ct. App. Nov. 18, 2011).

[268] *See generally* Stephanie A. Evans & Karen L. Salekin, *Involuntary Civil Commitment: Communicating with the Court Regarding "Danger to Other"*, 38 LAW & HUM. BEHAV. 325 (2013); P.P. Kwartner et al., *Judges' Risk Communication Preferences in Risk for Future Violence Cases*, 5 INT'L J. FORENSIC MENTAL HEALTH 185 (2006); Varela et al., *supra* note 218.

finding is reminiscent of the concern previously discussed related to the dangers posed by using amorphous categorical rankings that overtake any sense of absolute risk estimation. Judges and jurors are confounded even more by numerical results. A pair of studies discovered that risk estimates given in frequencies (e.g., 5 of 10) were perceived by jurors to amount to a greater risk of recidivism than equivalent probabilistic risk estimates (e.g., 50%),[269] which, of course, is illogical. Another study found that framing a risk estimate as the probability of violence (e.g., 26% likely to be violent) leads to a greater assessment of risk than when the equivalent risk estimate is framed as the probability of no violence (e.g., 74% likely to be nonviolent),[270] which is likewise incongruous. A suggested explanation for these results is that judges and jurors have issues with innumeracy, denoting a lack of ability with numbers,[271] particularly with statistics,[272] as buttressed by this body of research.

One of those experiments just mentioned has additionally confirmed the potential for courtroom manipulation. The researchers sought to compare judgments when risk assessment results were packed or unpacked. Unpacked results simply communicated the result as high risk or low risk, while packed results included the high or low risk attribution and provided additional contextual information to explain the relevant factors that contributed to the evaluee's high or low risk result. When the description of a risk assessment's high or low risk results were unpacked, study subjects were more likely to agree with the rating of high or low risk, respectively.[273] As a consequence of the findings, the researchers offered certain basic legal strategies: attorneys for the government would want to unpack high-risk but not low-risk results, while attorneys for the defendant would wish to do the opposite.[274] These suggestions clearly play on the apparent opportunity to exploit confusion in understanding the information that actuarial tools can provide.

A counterargument may be that judges and jurors are generally not entirely awed by science or by expert witnesses and that they affirmatively have the capability of critically assessing, even disregarding, expert evidence which

[269] Craig & Beech, *supra* note 41, at 205; Paul Slovic et al., *Violence Risk Assessment and Risk Communication: The Effects of Using Actual Cases, Providing Instruction, and Employing Probability Versus Frequency Formats*, 24 LAW & HUM. BEHAV. 271 (2000).

[270] Nicholas Scurich & Richard S. John, *The Effect of Framing Actuarial Risk Probabilities on Involuntary Civil Commitment Decisions*, 35 LAW & HUM. BEHAV. 83 (2011).

[271] Scurich et al., *supra* note 198, at 549.

[272] MORRIS E. CHAFETZ, THE TYRANNY OF EXPERTS: BLOWING THE WHISTLE ON THE CULT OF EXPERTISE 103 (1996); Joseph Sanders, *The Merits of the Paternalistic Justification for Restrictions on the Admissibility of Expert Evidence*, 33 SETON HALL L. REV. 881, 906 (2003).

[273] Scurich et al., *supra* note 198, at 551.

[274] *Id.* at 552.

they suspect is not up to par.[275] Whatever the merit that argument may have with other types of expert evidence, it seems unsuitable here. How can judges or jurors possibly comprehend the abilities and flaws of actuarial instruments when, as shown earlier, "experts" themselves are often mistaken about them?[276] Plus, any transparency offered by actuarial methods, for which they are widely lauded, is by and large a myth. The biased focus on the discrimination measure to judge predictive ability effectively conceals issues of poor calibration performance and lack of generalizability. Assessment and scoring practices, too, are clouded in secrecy. There are no suggested guidelines or limits on what types of evidence or witnesses should be used to gain the necessary information to score the worksheets. Evaluators often rely on hearsay and additional evidence that would otherwise be inadmissible in court and for which the truthfulness is unknown. Issues with the quality of the underlying information are often ignored as actuarial results are conveyed in an objective and mathematical manner with little enquiry into the sources from which the data and impressions were obtained. Just ponder, if you will, the likely sources on which evaluators base their deductions to score factors such as elementary school maladjustment, parental alcoholism, victim injury (VRAG), or the existence of any prior female victims of reported or unreported assaults (Static-99).

A supplementary consideration undermines the vision that actuarial risk assessment is simply another piece of information in a complex decision. The potential for undue prejudice can be realized by drawing upon several psychological constructs. One is anchoring bias, which occurs in any decision-making process when one places too heavy a weight—the anchor—on a single piece of information. Often the anchor is the starting point and it can maintain an overly influential effect on the final decision, particularly in the face of uncertainty.[277] Anchoring bias is a cognitive heuristic; when faced with particularly complex judgments, reliance upon an anchor can be seen as a useful mental short-cut, but it also tends to unconsciously produce significant errors.[278]

The potential for anchoring bias to occur, and the errors in decision-making that likely result, is evident as supporters often forthrightly advocate that actuarial tools should dominate sentencing decisions.[279] Thus, prejudicial

[275] Sanders, *supra* note 272, at 907.

[276] *See supra* note 138, 205–09 and accompanying text.

[277] Singer et al., *supra* note 37, at 347.

[278] *Id.*

[279] N.Y. STATE DIV. OF PROB. & CORR. ALT., NEW YORK STATE PROBATION SEX OFFENDER MANAGEMENT PRACTITIONER GUIDANCE 9 (2009) (directing probation officers that the actuarial results ought to "anchor the judgment or impressions").

impact occurs when the actuarial "results" potentially taint both the evaluators and triers of fact.[280] They may become biased toward the anchor—the actuarial result—and fail to adequately reassess that anchor even in light of contradictory information, which may be unwisely discounted.[281] A recent study is illuminating. Using judges as sample subjects, researchers found that when given actuarial predictions of sexual recidivism risk, judges became more conservative in their decisions; that is that they were more likely to order detention than without such information.[282] However, the rate of false positives also increased as a result.

Another prejudicial effect is related to general misestimations regarding the propensity of violent and sexual offenders to reoffend. Criminological research typically demonstrates that the public believes that recidivism rates of violent and sexual offenders are much higher than they are in reality.[283] The psychological construct of confirmation bias is informative here, indicating "the tendency to unwittingly select and interpret evidence in a manner that confirms a previously held belief or hypothesis, while minimizing or failing to recognize contrary evidence."[284] Further, cognitive dissonance occurs when, given conflicting evidence, people tend to select that evidence which reinforce decisions they have already made and downplay the contrary signs.[285] Thus, judges and jurors may overvalue a risk tool prediction that confirms their preexisting inclination toward assuming high risk.

Several studies are on point with respect to the potential bias toward believing in high recidivism potential. Using judges and/or mock jurors in their samples, researchers revealed that sample subjects given an actuarial prediction of high risk were more likely to adopt that valuation and to devalue a low risk actuarial prediction, possibly because they had a tendency to believe violent offenders routinely are dangerous.[286] A reasonable hypothesis is that the subjects valued opinions consistent with their hypothesis supporting dangerousness and discounted expert information that was

[280] Hart & Cooke, *supra* note 76, at 97.

[281] Fogel, *supra* note 75, at 46.

[282] Evans &. Salekin, *supra* note 268.

[283] Stacey Katz-Schiavone et al., *Myths and Facts About Sexual Violence: Public Perceptions and Implications for Prevention*, 15 J. CRIM. JUST. & POPULAR CULTURE 291, 299 (2008); Justin T. Pickett et al., *Vulnerable Victims, Monstrous Offenders, and Unmanageable Risk: Explaining Public Opinion on the Social Control of Sex Crime*, 51 CRIMINOLOGY 729 (2013).

[284] Ryan W. Scott, *The Skeptic's Guide to Information Sharing at Sentencing*, 2013 UTAH L. REV. 345, 374.

[285] Beecher-Monas, *supra* note 86, at 395–96.

[286] *See generally* Evans &. Salekin, *supra* note 268; Kwartner et al., *supra* note 268; Varela et al., *supra* note 218.

contrary thereto as it contradicted their assumption about sex offender recidivism.

There is confirming anecdotal evidence of sentencing judges discounting low risk assessments.[287] A sentencing judge in a particularly enlightening exchange explicitly rejected defendant's proffer of low risk assessment from a Static-99 scoring, describing the evidence as just "what some scientist or some academic guru might think about the likelihood of re-offending at this stage" and concluding: "[q]uite frankly, I don't care about Static[-]99."[288] Perhaps this judge is perceptive, indeed.

V. CONCLUSIONS

Risk assessment is envisioned as a progressive criminal justice policy. Optimistically, risk-informed practices permit officials to make smarter decisions in managing criminal populations to achieve cost-effective solutions by reducing reliance upon imprisonment while at the same time still protecting the public. Actuarial risk instruments are the modern face of the new risk penology, purportedly offering objective, reliable, and empirically validated predictions of recidivism potential. Whatever merit actuarial assessments may have for a variety of criminal justice decisions (such as bail, probation, and parole), they are far too problematic for use in sentencing matters. Sentencing is a critical stage in criminal proceedings. In sentencing adjudication, untoward and unquestioning reliance upon a potentially error-ridden source of information undermines the standards of evidence law, offends the principles of justice, and potentially thwarts the goals of deterrence and reducing recidivism.

Actuarial risk tools are unfortunately reified in sentencing proceedings as epitomizing an infallible application of scientific principles and the empirical method. The guise of science unwittingly convinces many that actuarial scales allow us to accurately and precisely predict the immanently unpredictable—human behavior. The nonpartisan qualities of numbers and statistics can be both seductive, allowing sentencers to feel that risk can be corralled, and powerful, seemingly insulating sentencing decisions in a veil of science. Nonetheless, the almost complete failure to question, critically analyze, or challenge the actuarial model of sentencing has potentially permitted unreliable science to invade sentencing law and remain undetected as such. The naturalistic fallacy is present, whereby proponents of actuarial

[287] State v. Perrine, No. 99534, 2013 Ohio App. LEXIS 5738, at *20–22 (Ohio Ct. App. Dec. 26, 2013) (discounting Static-99 low risk score).

[288] State v. Zink, No. 2011AP2684-CR, 2012 Wis. App. LEXIS 892, at *3–4 (Wis. Ct. App. Nov. 14, 2012).

justice confuse what is with what ought to be. We might want, for all sorts of justifiable and honorable reasons, to be able to objectively and reliably differentiate likely recidivists from non-recidivists, and to sentence accordingly. But no matter how much we wish for science to solve our problems, current actuarial methods for predicting risk are not the panacea advocates imply. To the extent that sentencing includes utilitarian concerns involving future risk, science cannot save the legal system from a heavy measure of uncertainty. Certainly, unreliable science will not be the savior.