Kumho Tire and Expert Reliability: How the Question You Ask Gives the Answer You Get

Mark P. Denbeaux[*] *and D. Michael Risinger*[**]

---

[*] Professor of Law, Seton Hall University School of Law. B.A., College of Wooster, 1965; J.D., New York University, 1969.

[**] Professor of Law and Dean's Research Fellow, Seton Hall University School of Law. B.A., Yale University, 1966; J.D., Harvard University, 1969. The authors would like to thank David "Jake" Barnes for his comments on a draft of this Article.

INTRODUCTION

It has become a commonplace none the less accurate for that, to say that *Daubert v. Merrill Dow*[1] precipitated a revolution in the law of expert evidence,[2] the endpoint and exact contours of which are not yet fully worked out. Nevertheless, it is becoming increasingly clear that this revolution is changing the practical realities and results of trial in many cases or classes of cases in which various sorts of expertise play a central role. This is currently most obvious in regard to toxic tort and products liability claims,[3] but potentially the effects of the revolution will almost certainly be felt in a much broader range of cases, including all those criminal prosecutions in which claimed expertise plays a substantial role in the outcome.

As the revolution unfolds, it raises serious issues along a number of axes.[4] All of these threads of controversy are interdependent, and

---

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

[2] *See, e.g.*, David L. Faigman, *The Law's Scientific Revolution: Reflections and Ruminations on the Law's Use of Experts in Year 7 of the Revolution*, 57 WASH. & LEE L. REV. 662 (2000); Thomas R. Freeman, *Guardians at the Gate*, 24 L.A. LAW. 26, 28 (2001) (referring to the movement as "the *Daubert* revolution"); Marc S. Klein, *The Revolution in Practice and Procedure:* Daubert *Hearings*, 1 SHEPPARD'S EXPERT & SCI. EVID. Q. 655 (1994); Brian C. Murchison, *Treating Physicians as Expert Witnesses in Compensation Systems: The Public Health Connection*, 90 KY. L.J. 891, 917 (2001-02) (referring to the "the *Daubert* revolution"); Joseph Sanders, Shari S. Diamond & Neil Vidmar, *Legal Perceptions of Science and Expert Knowledge*, 8 PSYCHOL., PUB. POL'Y & L. 139, 142 (2002) (same).

[3] *Daubert* has led to a rise in summary judgments against plaintiffs in tort cases in federal court resulting from exclusion of proffered expert evidence. *See* LLOYD DIXON & BRANDON GILL, CHANGES IN THE STANDARDS FOR ADMITTING EXPERT EVIDENCE IN FEDERAL CIVIL CASES 56 (2001). Half of those summary judgments involved exclusion of evidence regarding the cause of the plaintiff's injury. *Id.* This is the hallmark of the toxic tort case. In absolute numbers, the plurality of the cases were classified as products liability cases. *Id.* at 21, tbl. 3.3. On the criminal side, with a couple of notable exceptions, *Daubert* has had very little impact. *See* D. Michael Risinger, *Navigating Expert Reliability: Are Criminal Standards of Certainty Being Left on the Dock*, 64 ALB. L. REV. 99, 149 (2000) [hereinafter Risinger, *Navigating Expert Reliability*].

[4] These include: (1) the role of the judge versus the role of the jury in jury trials; (2) the ideal of a uniform standard for establishing the preconditions of evidence admissibility versus the impact of such low standards on the broader promises represented by the case standard of proof as a whole; (3) the tenability of the claim that judicial evaluation of evidentiary sufficiency adequately resolves questions of low standards of admissibility when applied to claimed expertise; (4) judicial competency to evaluate claims of expertise versus judicial deference to expert communities on the validity of such claims; (5) the ideal of faith in juries to handle and evaluate mixed information more satisfactorily than any other institutional arrangement for dispute resolution versus profound suspicion that there are broad categories of information (claimed expertise among them), that juries cannot be expected to evaluate well; (6) loose standards for the scope of an expert's claimed expertise versus tight standards for scope of expertise; (7) concern that like cases be treated

no satisfactory, full examination of the evolving law of expert evidence can fail to touch on each of these issues. However, different foci will yield different insights, and in this Article we focus on the question of how a court is to go about the task of framing the issue to which standards of reliability will be applied in the individual case. This framing process directly implicates the issues of decisional specificity and generality and their interconnection with the uses and abuses of discretion and precedent.

As a preliminary matter, it is interesting to note that this revolution was anything but inevitable. If one examines the record of the federal district courts from the passage of the Federal Rules of Evidence to the decision in *Daubert*, one sees at first more or less what one would expect to see:[5] for the first decade or so, opinions putter along generally in the expected and usual way, with few reliability challenges to proffered expertise[6] worth the label.[7] Then, around

---

alike versus normal notions of appellate deference to trial courts on rulings of evidentiary admissibility; (8) concern that different cases be treated differently versus a systemic interest of all judges in disposing of foundational issues regarding expertise on broad grounds so as to be spared by precedent from having to repeatedly consider the asserted reliability of various sub- and sub- sub-expertises in a potentially very great number of cases.

[5] This was accomplished by piggybacking on the research and research strategy of the Rand study, DIXON & GILL, *supra* note 3. In order to generate pre-*Daubert* data on how federal district courts handled challenges to expertise in civil cases, the Rand researchers used a Westlaw search with a 27-term search string which was overinclusive but unlikely to exclude any such challenge. *Id.* at 17. This string generated 4097 hits from December 31, 1979 through June 1999. A random one-third (1345) of those cases were examined by "coders," who were law students or recent law graduates who had been trained to evaluate the opinions in regard to whether they involved challenges to the admission of expert testimony on reliability grounds in civil cases. *Id.* Their examination produced 399 cases which in their judgment involved such challenges. Of those, 163 were decided before *Daubert*. *Id.* at 20, tbl. 3.2. We contacted Dr. Dixon, and he graciously supplied us with a list of those 163 cases. We then examined them ourselves. It turns out that we found that we had some disagreement with the Rand raters as to the characterization of some of the cases. At any rate, the assertions in the text are based on our re-examination of that set of 163 cases, plus those turned up when we extended the search back to the effective date of the Federal Rules of Evidence (January 1, 1975), a sample universe large enough to yield confident insights.

[6] The Rand search string turns up a total of about 70-75 cases a year of all types until 1985, when the number jumps to 96, followed by 136 in 1986, and 154 in 1987, with that trend continuing thereafter.

[7] One of our major disagreements with the Rand raters concerns various decisions which they treated as reliability decisions which we do not regard as such. The largest such category involves persons proffered as experts who were called upon to give what were essentially legal opinions. This was the largest single category in the set of pre-*Daubert* cases (23 cases, or 14 percent of the total), resulting in 19 exclusions, or nearly a third of all exclusions. While courts are not always scrupulous in guarding the line between legal conclusions and other kinds of statements, a fair

mid-1985, things begin to take off,[8] with serious issues of expert reliability becoming hotter and hotter. However, these challenges, with their attendant dispute and controversy, are largely concentrated in one type of case dealing with one issue—the type of case and issue involved in *Daubert* itself—that we may style "risk increase" causation in toxic tort.[9] Even before the Supreme Court granted certiorari in the *Daubert* case, lower courts had stripped the "novel" requirement out of the *Frye* test in such cases, holding that Federal Rule of Evidence 702[10] required general acceptance at least of the methodology employed in the generation of all expert evidence assertedly based on science when such evidence was proffered on the issue of causation in toxic tort cases.[11] In addition, many of the lower court opinions contained language even more *Daubert*-like than this regarding reliability in general.[12]

When these issues (in particular the role of "general acceptance") were finally faced by the Supreme Court itself, it was not at all obvious that the Court would go beyond fashioning a doctrine

---

proportion of courts have always rejected such testimony, and done so on the basis of role, not reliability. We do not take these cases to deal with "reliability" in the *Daubert* sense.

[8] The number of cases hit by the Rand search string held between 32 and 43 for every six-month period from 1980 until the second half of 1985, when it jumped to 53 and accelerated from there. It is almost as if Judge Weinstein's opinion in *In re "Agent Orange" Product Liability Litigation*, 611 F. Supp. 1267 (E.D.N.Y. 1985), triggered the rise. *See id.* at 1275-76, 1285 (excluding expert testimony on the causal link between various complaints of Vietnam veterans and their exposure to the defoliant Agent Orange after a *Daubert*-like analysis, and granting summary judgment for the defendants). However, the Agent Orange case was not the first *Daubert* precursor. That award must go to Judge Becker's opinion while a district court judge in *Zenith Radio Corp. v. Matsushita Elec. Indus. Co.*, 505 F. Supp. 1313 (E.D. Pa. 1981), which excluded various proffered expertise on reliability grounds and then granted summary judgment. This decision was later reversed by the Third Circuit per Judge Gibbons in *In re Japanese Electronic Products Antitrust Litigation*, 723 F.2d 238 (3d Cir. 1983). The collision of viewpoints between Judge Becker and Judge Gibbons prefigured many of the controversial *Daubert* issues to come.

[9] *See, e.g.*, Viterbo v. Dow Chem. Co., 646 F. Supp. 1420 (E.D. Tex. 1986) (excluding treating physician on issue of causal link between plaintiff's damages and exposure to herbicide). There are 24 such cases, two-thirds of which resulted in exclusion.

[10] Hereinafter in the text Federal Rule of Evidence 702 will be referred to as "FRE 702," or simply "Rule 702." We regard the abbreviation "Fed. R. Evid." as clumsy and as interrupting flow when used in text.

[11] *See, e.g.*, Daubert v. Merrell Dow Pharm., Inc., 727 F. Supp. 570 (S.D. Cal. 1989) (deciding on lack of general acceptance grounds without reference to novelty).

[12] *See* Mendes-Silva v. United States, No. 89-1131 (RCL), 1991 WL 135090 (D.D.C. July 12, 1991). In *Mendes-Silva*, one party's experts claimed that vaccine-induced encephalopathy had been caused by yellow fever vaccine. The court rejected the proffered experts and provided a good summary of the tides of federal judicial opinion on expert reliability as of the date of the opinion.

that was limited to the recurring problems of that narrow band of cases. Nor was it certain that the Court's doctrine would go beyond explicitly "scientific" evidence, a limitation which would have had the imprimatur of tradition, and which would have covered all potential proffers of expert testimony on the issue of risk increase induced by exposure to a claimed toxic substance. However, as we all know, at least by now, this the Court did not do. Instead, Justice Blackmun fashioned an opinion which, although addressed mainly to a view of the right way to judge reliability in the context of evidence claiming the mantle of science, was firmly based on a construction of FRE 702 which was not so limited, but broader and trans-substantive, requiring by its logic some threshold reliability determination for all proffered expertise.[13]

At first, litigants and lower courts were divided on whether the Supreme Court had actually intended what it had apparently implied in those passages of the opinion.[14] Finally, *Kumho Tire v. Carmichael*[15] laid to rest all residual doubt about the breadth of the change in approach being mandated, but left unclear many questions about how the new enhanced gatekeeping approach was to operate in differing contexts. In order to approach these questions with appropriate caution, it is perhaps advisable to start from the bottom up.

DANGERS AND DIFFICULTIES OF EXPERTISE IN AN ADVERSARY SYSTEM THAT USES JURIES

Our legal notions about how to approach information for purposes of doing justice and resolving disputes are inextricably bound up with the institution of the jury, even in non-jury contexts.[16]

---

[13] The Court in *Daubert* said as much: "Rule 702 . . . clearly contemplates some degree of regulation of the subjects and theories about which an expert may testify." 509 U.S. at 589. But the Court then limits itself to a discussion of the nature of that gatekeeping responsibility "in the scientific context because that is the nature of the expertise offered here." *Id* at 590 n.8. A functionally similar gatekeeping responsibility for proffers of expertise not based on science would seem to follow, and, in the event, did.

[14] *Compare, e.g.*, Iocobelli Constr., Inc. v. County of Monroe, 32 F.3d 19 (2d Cir. 1994) (stating that *Daubert* validity analysis applies only to scientific evidence), *with* Berry v. City of Detroit, 25 F.3d (6th Cir. 1994) (explaining that *Daubert* gatekeeping and validity requirements apply to all expert testimony).

[15] Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999).

[16] Professor Damaska sees the three institutional variables that condition the characteristic details of Anglo-American proof law (the "pillars carrying common law evidence") as being the organization of the trial court (most notably the split-function jury system), the temporal concentration of proceedings (the "day in court" model of the trial that starts and runs until it is finished), and party control (the

By accident, or by some process that might be said to embody at least some sort of design, our British forebears and ourselves evolved a decisional system which displays some useful attributes of more specifically purpose-built systems in other areas of endeavor. The most important of these attributes for this Article (and perhaps generally) is the creation of a two-stage split-function system in which the first decision-maker (the judge) controls the information available to the second decision maker (the jury), thus making possible masking and bias filtration.[17] The filtration is functionally analogous to that which has become a norm of the scientific method in the last half century.[18] This system, in turn, makes it necessary to determine by what standards a judge should perform this general "gatekeeping" function in an adversary system.

A general consideration of this topic is of course beyond the scope of this Article. However, it will serve present purposes to observe that a number of competing interests come to bear on the issue. The collision of these interests has resulted in two general positions regarding proper standards of gatekeeping, positions which are at odds with each other and are imperfectly reconciled in both theory and practice. We might call these the "when in doubt, let it in" principle and the "when in doubt, keep it out" principle.[19] They

---

adversary system). *See* MIRJAN R. DAMASKA, EVIDENCE LAW ADRIFT 125 (1997). While these are logically separable, in that one could, at least in theory, utilize a jury without a separate judge, and one could certainly utilize a bias-filtering split function without a jury, have a concentrated trial without a jury, and have a dominantly party-controlled procedure without a jury, the existence of the jury would seem to make the development of the other aspects more likely. This is true even though current interpretations of the historical record place the development of strong adversary control rather late in the game, in the late eighteenth century. *See* T.P. Gallanis, *The Rise of Modern Evidence Law*, 84 IOWA L. REV. 499 (1999); Stephen Landsman, *The Rise of the Contentious Spirit: Adversary Procedure in Eighteenth Century England*, 75 CORNELL L. REV. 497 (1990); John H. Langbein, *Historical Foundations of the Law of Evidence: A View from the Ryder Sources*, 96 COLUM. L. REV. 1168, 1197-1202 (1996).

[17] *See* Allen D. Allen, *Scientific Versus Judicial Factfinding in the United States*, SMC-2 IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS 548-50 (1972). Professor Damaska makes the point that it is this split function that makes the notion of "admissibility" even intelligible. Mirjan Damaska, *Of Hearsay and Its Analogues*, 76 MINN. L. REV. 425, 455-56 (1992).

[18] *See* D. Michael Risinger, Michael J. Saks, William C. Thompson & Robert Rosenthal, *The* Daubert/Kumho *Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion*, 90 CAL. L. REV. 1, 9 (2002) [hereinafter Risinger et al., *Observer Effects*].

[19] One might say that this tension is best symbolized by Rule 403's requirement that probative value be "substantially" outweighed by prejudicial effect before exclusion results on the one hand, and by Rule 104's general position that the burden is on the proponent of evidence to make out its admissibility. In *Kumho Tire*, the Supreme Court reaffirmed that the general burden is on the proponent, but indicated that the opponent has an obligation to point out why there is a reason to

represent different resolutions of issues about the role of parties and judges in an adversary system, the competence and trustworthiness of juries, and the role of rationality in legal decision-making.

To put this in context, it is necessary to say a bit more about the notion of an adversary system and the contrasting alternative, an "inquisitorial" system.[20] An adversary system is one in which the decisions about how contested disputes are to be presented to the ultimate decision-maker are left in the hands of the disputing parties. In an inquisitorial system[21] (not a very apt label, but too embedded to be changed here), these same decisions are all made by a decision-maker who is at least formally neutral between the parties, and who may or may not also be the ultimate decision-maker on the merits of the dispute.

For all the rhetoric that is sometimes thrown up in defense of "our adversary system,"[22] no one argues in favor of a pure adversary system for the simple reason that such a pure system, like direct democracy, could not function except under exceedingly rare conditions. This is because a pure adversary system would have no judge, in the sense we are accustomed to. The parties would be free to present whatever they themselves determined to be helpful to their cause, and the party with the weakest case could filibuster indefinitely, like a member of the Senate reading from a telephone book, limited only by expense and endurance.[23] For this reason at

---

doubt the proffer's admissibility, to call its admissibility "sufficiently into question" before the court is obliged to make any determination. 526 U.S. at 149, 152-53. This further obscures the default position and heightens the tension between the two approaches. Improper reversal of the burden of showing admissibility in the face of serious challenge is one of the common tactics which courts have adopted to arrive at "light-touch" treatment of prosecution-proffered expertise in criminal cases. *See* Michael J. Saks, *The Legal and Scientific Evaluation of Forensic Science*, 33 SETON HALL L. REV. 1167 (2003).

[20] By far the most sophisticated and influential writer in English on these issues, both from a descriptive and a normative perspective, is Professor Damaska, starting with Mirjan Damaska, *Evidentiary Barriers to Conviction and Two Models of Criminal Procedure: A Comparative Study*, 121 U. PA. L. REV. 506 (1973), and culminating in DAMASKA, *supra* note 16.

[21] *See* JOHN HENRY MERRYMAN, THE CIVIL LAW TRADITION 126-28 (2d ed. 1985); *see also* Mason Ladd, *Expert Testimony*, 5 VAND. L. REV. 414 (1952) (contrasting the rise of the English adversary system with what Ladd took to be the "inquisitorial" function of the jury at an earlier time).

[22] As of February 2003, this cliché phrase generated 956 hits in the Westlaw Journals database (JLR, as it is designated).

[23] Wigmore claimed that something approaching this existed among some African tribes. JOHN H. WIGMORE, A KALEIDOSCOPE OF JUSTICE 728 (1941). Even in such a system, there would in most cases be a practical pressure not to try the patience of the ultimate decisionmaker too much with proffers lacking any apparent bearing on the dispute, lest this be held against you when decision time finally came.

least, one would expect little disagreement that such a pure adversarial system would be undesirable and that there must at least be some judge-administered standards of relevance which will both structure and limit the parties' freedom to proffer material and oblige the ultimate decision-maker to wade through it.[24] In addition, most would probably agree that this would apply even when the parties themselves, for whatever their reasons, would be content to reciprocally drown the decision-maker in mountains of information of peripheral (or no) relevance to the issues that the applicable substantive law defines as the material issues of the dispute. Although it is sometimes said broadly that there is party autonomy to make the rules of evidence for the individual case by agreement or failure to object,[25] there is at least a time-efficiency interest which belongs to the dispute resolution system itself that vests the judge with authority to cut off such proffers.

Beyond this minimum, the more pure adversaryists would begin to dig in their heels. They would say that truth best emerges in the clash of self-interested parties packaging whatever relevant information is available[26] in the most persuasive way they can, that juries are (because they are not repeat players and are a group with a range of life experiences to bring to bear in evaluating the meaning of information) the best possible decision mechanism to handle all types of relevant information, and that, essentially, the system would work best if there were no rules of evidence beyond a weak relevance check.[27]

Adversary skeptics have a different view.[28] While few, if any (at

---

[24] "Courts are so organized that there must be some limit to the facts which may be given in evidence, as there must be an end of litigation." BURR W. JONES, THE LAW OF EVIDENCE IN CIVIL CASES 3 (1896).

[25] *See* DAMASKA, *supra* note 16, at 87.

[26] This Article's text has concentrated on aspects of adversary control of the *presentation* of information, mainly because the Article is about judging the reliability of *proffers* of expert testimony. One of the most powerful criticisms of an adversary system, however, is that it leaves both the gathering and presentation of information to partisan adversaries, and this may result in the *non-presentation* of important information. This situation is exacerbated by limitations on information-sharing, especially in the criminal context. *See* DAMASKA, *supra* note 16 at 98-101; *see also* ALVIN I. GOLDMAN, KNOWLEDGE IN A SOCIAL WORLD 300-04 (1999).

[27] For a description of this position, see Dale A. Nance, *Reliability and Admissibility of Experts*, 34 SETON HALL L. REV. 191 (2003). As Professor Nance notes: "[T]he *locus classicus* for this argument is Lon L. Fuller, *The Adversary System*, *in* TALKS ON AMERICAN LAW 34 (Harold J. Berman ed., 1961)." *Id.* at 195 n.13; *see also* GOLDMAN, *supra* note 26, at 296.

[28] For a careful examination of the arguments concerning the relation of partisan adversary control to the truth-seeking function, see DAMASKA, *supra* note 16, at 74-103, and GOLDMAN, *supra* note 26, at 295-300. *See also* JEROME FRANK, COURTS ON

least in our legal culture), would advocate judicial gathering and winnowing of dispute-relevant information and its presentation to the ultimate decision-maker in edited and summary form with no input from or consultation with the parties, adversary skeptics press two kinds of arguments in favor of heavy judicial and rule-structured control of what is given to the jury: adversary excess and jury weakness in dealing with certain kinds of information.[29] The two arguments dovetail. If juries can be persuaded to abandon reason or overvalue certain information, advocates will not only not hesitate to do it,[30] they will arguably have an *obligation* to do it on behalf of their clients.[31] Upon reflection, there appear to be certain commonsense classes of information that are subject to such abuse, such as hearsay or "character" evidence. We must both authorize and require judges to filter such information.[32]

The formal rules of evidence give judges plenty of authority to filter what the judge determines to be iatrogenic information and to exclude it.[33] However, the long tradition of the system, at least in the last century, favors judicial restraint except in areas of mandated categorical exclusion.[34] What the *Daubert* and *Kumho Tire* decisions

---

TRIAL 82-87 (1949); MARVIN FRANKEL, PARTISAN JUSTICE (1980); Warren Berger, *Agenda for 2000 A.D.—A Need for Systematic Anticipation, Address at the National Conference on the Causes of Popular Dissatisfaction with the Administration of Justice* (Apr. 7-9, 1976), *in* 70 F.R.D. 79, 83-96 (1976); Steven Landsman, *Who Needs Evidence Rules, Anyway?*, 25 LOY. L.A. L. REV. 635 (1992); Carrie Menkel-Meadow, *The Trouble with the Adversary System in a Post-Modern, Multi-Cultural World*, 1 J. INST. FOR STUDY LEGAL ETHICS 49 (1996); Roscoe Pound, *The Causes of Popular Dissatisfaction with the Administration of Justice*, 40 AM. L. REV. 729 (1906), *reprinted in* 35 F.R.D. 273 (1964); Gordon Van Kessel, *Adversary Excesses in the American Criminal Trial*, 67 NOTRE DAME L. REV. 403 (1992).

[29] DAMASKA, *supra* note 16, at 84.

[30] It is adversary excess—"the old razzle dazzle," in the terms of the recent depiction of the trial system in the Academy Award winning motion picture "Chicago"—that undermines many of the arguments which are sometimes made concerning the desirability of "informational completeness," and the claim that rules of exclusion underestimate and disrespect juries. It is not what juries would do with information in a vacuum, but rather what lawyers will do to juries using the information, that justifies rules of exclusion.

[31] See the obligation of zealous representation reflected in the ABA Model Rules of Professional Conduct. *See, e.g.*, MODEL RULES OF PROF'L CONDUCT R. 1.3 cmt. 1 (2003) (requiring the lawyer to act with "zeal in advocacy").

[32] *See, e.g.*, FED. R. EVID. 404, 801, 802.

[33] *See, e.g.*, FED. R. EVID. 403 (providing leeway even in the face of the word "substantially"). The term "iatrogenic" was originally a medical term meaning "caused by the physician" and refers to a situation made worse by attempts to improve it. By extension, it is a very useful word in regard to proof rules of various kinds that on balance do more veritistic harm than good.

[34] This is perhaps best represented symbolically by the requirement of Rule 403 that probative value be "substantially outweighed" by prejudicial effect before a judge

did was to move expert evidence generally from the area of "light touch control" and party autonomy to the area of heavier judicial evaluation and control, in the name of "reliability."

So what's so wrong with unreliable expertise anyhow? The commonsense fear is that factfinders will defer to the unreliable expert and treat the unreliable expert's testimony as reliable. One could respond that this danger exists in regard to all evidence. However, at least as to fact witness testimony, and various forms of documentary, physical, or circumstantial proof, the assumption is that average people have developed, through the process of living in society, sufficient knowledge about the world of humans and its workings that they have a fair chance to evaluate and accurately weigh and discount information coming from such familiar sources. Problems of misleading specialized social context are dealt with by having a group of factfinders from across a range of social backgrounds and experiences (that is, a jury), one or more of whom it is hoped will be familiar with the specialized context from living in it in their ordinary lives. Whatever the empirical realities of that assumption,[35] it becomes increasingly tenuous as the information presented to the factfinder becomes more and more removed from any socially common experience.[36] And such claims of specialized knowledge or skill beyond common experience are the essence of asserted expertise.[37]

THE TRADITIONAL APPROACH TO EXPERT RELIABILITY

As previously noted, at least until what we might call the run-up period immediately preceding *Daubert*, judges were not compelled by doctrine and rarely undertook in practice to evaluate the asserted warrant to believe claims of expertise directly, in the terms put forth by the practitioners of the claimed expertise. In those days of "light touch" evaluation, about all that was generally required was that the proposed testimony be facially relevant, usually by virtue of some conclusion ("opinion") put forth by the putative expert, and that the

---

should exclude evidence proffered by a party, a signal not to indulge too fine an exclusionary instinct to which most judges adhere even in fairly extreme circumstances. *See* D. Michael Risinger, *John Henry Wigmore, Johnny Lynn Old Chief and "Legitimate Moral Force"—Keeping the Courtroom Safe for "Heartstrings and Gore,"* 49 HASTINGS L.J. 403, 429-31 (1998).

[35] It only helps in regard to the common experience of significantly large communities in the venire pool, and even then there may be no representative of the salient community on a given jury where one is needed.

[36] *See* GOLDMAN, *supra* note 26, at 308-09.

[37] *See* D. Michael Risinger, *Preliminary Thoughts on a Functional Taxonomy of Expertise*, 31 SETON HALL L. REV. 508, 510-11 (2000) [hereinafter Risinger, *Taxonomy*].

claimed basis of the expertise not be affirmatively discredited as invalid in some generally known and accepted way.[38] In part because of this low threshold of admissibility, even the concept of expertise itself was ill-defined and under-examined in any kind of defensible taxonomic detail. About the only subcategory of expertise which was generally recognized was "scientific" expertise, and the demarcation between "scientific" and other kinds of expertise was not at all clear.[39] As a result of the famous case *Frye v. United States*,[40] scientific expertise was further divided between "novel" scientific expertise and the rest, with only "novel" expertise being subject to any special admissibility consideration bearing on validity. Non-novel "scientific" expertise was thrown back into the "light touch" pool along with all other claims of expertise.

Even in regard to novel scientific evidence (however defined), judges were generally not asked to evaluate the reliability of expert claims in the claims' own terms, but merely to determine if others in an appropriate reference community (the "pertinent field"[41]) accepted them. Hence, judges were spared the necessity of learning anything about criteria that might be applied to evaluate reliability directly, instead merely trusting whatever reference community was chosen to properly select and apply such criteria. And for everything in the "light touch" pool (including "non-novel" "scientific" expertise), facial relevance and minimum plausibility, backed by commercial respectability, remained the only conditions of admissibility.[42]

This is not to say that there were not some judges who were more aggressive than the average in attempting to filter out unreliable expertise of whatever stripe. In so doing, they sometimes used a tool which, though somewhat indirect, has important analogues in, and implications for, current practice: definition of the scope of a particular individual's expertise.[43] Defining the

---

[38] D. Michael Risinger, Mark P. Denbeaux & Michael J. Saks, *Exorcism of Ignorance as a Proxy for Rational Knowledge: The Lessons of Handwriting Identification "Expertise,"* 137 U. PA. L. REV. 731, 766-67 (1989) [hereinafter Risinger et al., *Exorcism*].

[39] Risinger, *Taxonomy*, *supra* note 37, at 509.

[40] 293 F. 1013 (D.C. Cir. 1923).

[41] *Id.* at 1014.

[42] David L. Faigman, Elise Porter & Michael J. Saks, *Check Your Crystal Ball at the Courthouse Door, Please: Exploring the Past, Understanding the Present, and Worrying about the Future of Scientific Evidence*, 15 CARDOZO L. REV. 1799, 1803-09 (1994). The authors refer to this as the "commercial marketplace" test, but the respectability of the market seems also to have been an important factor.

[43] *See, e.g.*, Globe Indem. Co. v. Highland Tank & Mfg. Co., 345 F. Supp. 1290 (E.D. Pa. 1972) (holding that neither an electrical engineer nor an industrial

appropriate scope of a given person's claimed expertise raises considerations which will become more generally important later in this Article, to wit, the relative dangers of over-generality and over-specificity in framing such claims.

Consider an auto mechanic who has been trained by Ford Motor Company to work on the engines in Taurus automobiles, and has worked on such engines exclusively for the last five years, since he completed his training. What can we say of the scope of his expertise in regard to diagnosing an auto engine malfunction? We could begin by saying that his only real expertise is limited to the workings of engines he has actually examined physically in his work. However, most of us would probably agree that this is so narrow as to be both useless and unnecessarily grudging, even if his knowledge in regard to those particular engines is marginally more reliable in theory than that same knowledge extended to other engines we take to be functionally similar, because they were manufactured on the same assembly lines or according to the same specifications. But if we are willing to grant him functionally equivalent reliability of knowledge in regard to the similar but not-directly-experienced engines, we have taken the first step in a journey with no clear endpoint. This journey is in some contexts called "extension," in others "generalization," and in yet others "external validity"[44] and is related to the general philosophical problem known as the problem of induction: how far is one justified in generalizing from particular instances to other things, either individually or represented by more general constructs such as categories or the hierarchically interconnected categories and concepts known as theories.[45] To continue with our mechanic, we would probably be willing to grant him knowledge sufficiently reliable to be useful, and more reliable than that of the average person, in regard to—what? All Ford engines, all automobile

---

hygienist was qualified to testify as an expert on the design of molasses storage tank where neither witness had any prior experience or observational knowledge regarding the proper design of a molasses storage tank under the factual setting presented). Defining the scope of the expert's expertise is essentially the approach that Judge Becker, in *United States v. Downing*, 753 F.2d 1224 (3d Cir. 1985), recommended that Judge Wiener explore on remand. Judge Becker, perhaps as a means of salvaging the conviction, recommended that the approach be applied to the proffered "weaknesses of eyewitness identification" expertise that had been wrongly excluded globally at trial. This case generated the label "fit" which was later taken up by the Supreme Court in *Daubert*. *See Daubert*, 509 U.S. at 591. It is possible, however, that the "fit" rhetoric may be ultimately traceable to Wigmore. *See infra* note 150.

[44] WILLIAM R. SHADISH, THOMAS D. COOK & DONALD T. CAMPBELL, EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR GENERALIZED CAUSAL INFERENCE 83-93 (2002).

[45] *See* MICHAEL WILLIAMS, PROBLEMS OF KNOWLEDGE 201-10 (2001).

engines, all gasoline-powered internal combustion reciprocating engines, all engines of any kind, all mechanical principles at work in Taurus engines? What factors are in play when we approach such a problem?

First, it seems clear that the narrower we draw the circle around the mechanic's specific experience and training, the more reliable the allowed exercise of expertise will be (assuming any reliability in the first place). However, we can draw the circle so narrowly that neither the allowed circle of this mechanic, or almost any other, will contain the problem at issue in the case to be litigated. Drawing the circle too narrowly will deprive the litigant of needed expertise, not just from this expert, but potentially from the universe of available experts. This would not be justified if the territory of the circle drawn a bit wider would still represent acceptable reliability and include the subject matter of the case at bar. However, drawing the circle too broadly will give some litigants the right to present faux expertise to which they are not entitled.

A large factor in deciding how broadly to draw the circle in the legal context would seem to be the ease of obtaining more appropriate and reliable expertise. If the case at bar involves the workings of a Wankel rotary engine, a court might appropriately draw the circle narrowly, insisting on the production of a mechanic actually familiar with Wankel engines; however, if a plague had struck the international convention of Wankel engine mechanics and wiped them out, then it might be appropriate to draw the circle more broadly.

These considerations are "best evidence" considerations, or perhaps, more appropriately, "better evidence" considerations.[46]

---

[46] The distinction between "best evidence" requirements, and "better evidence" considerations applied to various decisions on admissibility, emerges from the interchange between Professor Imwinkelried and Professors Faigman, Kaye, Saks, and Sanders reflected in Edward J. Imwinkelried, *Should the Courts Incorporate a Best Evidence Rule into the Standard Determining the Admissibility of Scientific Testimony?: Enough Is Enough, Even When It Is Not the Best*, 50 CASE W. RES. L. REV. 19 (1999), and David L. Faigman, David H. Kaye, Michael J. Saks, & Joseph Sanders, *How Good Is Good Enough?: Expert Evidence Under* Daubert *and* Kumho, 50 CASE W. RES. L. REV. 645 (2000). In contrast with Professor Imwinkelried, Professors Faigman, Kaye, Saks, and Sanders propose a "better evidence" principle for evaluating the admissibility of proffered expert testimony. It was Professor Nance who first pointed out (and embraced) the extent to which the likely availability of better evidence affects admissibility in today's world. *See* Dale A. Nance, *The Best Evidence Principle*, 73 IOWA L. REV. 227 (1988). In *Old Chief v. United States*, 519 U.S. 172 (1997), the Supreme Court recognized a kind of "better evidence" principle when it declared that one important consideration in determining exclusion of prejudicial evidence under Rule 403 is the existence of equally probative but non-problematic evidence. *Id.* at

Historically, the default position of most judicial practice (consistent with the "light touch" approach and the reliance on jury evaluation) seems to have been to draw the circle broadly as a matter of course and only narrow it as the broad definition of scope became seriously out of kilter in regard to exterior boundaries of specialization, and perhaps not even then. For example, the nineteenth century mantra that "any medical man is qualified to testify upon any issue of medical practice" seems to still represent the formal position of many jurisdictions,[47] although the realities of the quality of knowledge represented by boundaries between medical specialties ought to compel a rethinking, and courts are generally spared consideration of the "rule" by the good sense of parties in selecting experts.

Note that to this point we have been discussing situations where an expert is disqualified because an issue is deemed to be outside the scope of her expertise but where it is assumed that there are at least some potential experts for whom the issue is within the scope of their expertise.[48] Where *Daubert* general reliability issues and scope of expertise dovetail most dramatically would be when a court appears to rule that testimony is outside of the proffered expert's area because it is beyond the state of anyone's art, at least anyone within the expert's asserted discipline.

The practical result of the dominant twentieth century practice was that judges generally deferred to experts' claims of expertise, or in the case of "novel scientific" evidence, pursuant to the *Frye* test, to the evaluations of those claims by other putative experts. As already noted, this spared the judge from having to learn anything about the epistemic underpinnings of an expert's claim to expertise, issues that are quite complicated and not entirely well worked out in any area of

---

651-52. Suffice it to say that, along with Professors Faigman, Kaye, Saks, and Sanders, we believe that it is proper to take into account how much better a proffer could have been, in determining whether it is good enough.

[47] *See generally* Roy W. Fouts, *The Medical Expert Witness*, 19 NEB. L. BULL. 213 (1940). Such a position is perhaps less surprising when one considers that no specialization had any institutional recognition within medicine until 1933. *Id.* at 219. McCormick stated in the 1954 first edition of his treatise that no membership in a specialty was required as a condition of giving medical testimony, CHARLES T. MCCORMICK, A HANDBOOK ON THE LAW OF EVIDENCE § 14, at 29 (1954), a statement that has survived through five editions until today, see 1 MCCORMICK ON EVIDENCE § 13, at 24 (John W. Strong ed., 5th ed. 1999).

[48] *See, e.g.*, *Globe Indem. Co.*, 345 F. Supp. at 1291 (stating that while "[t]here are people in the world who would qualify to give expert testimony on this question," the proffered engineer and toxicologist were insufficiently expert "regarding the proper formulation of safety criteria to be followed in the design of molasses tanks in this particular industrial setting").

claimed knowledge.[49] It moved the task of such evaluation, if it was going to be done at all, onto the plate of the jury, guided only by such issues as might emerge from cross-examination, or by each juror's own evaluations of the reliability of the testimony (often based, one must suppose, on the jurors' surface impression of the apparent authoritativeness of the witness and the apparent plausibility of any supporting explanations to which she might testify). Given the difficulty of distinguishing apparent plausibility from validity, it is no wonder that the dominant account of the jurors' usual practical role in resolving issues that have been the subject of expert testimony is either deferential acceptance when only one expert testifies, or selection between the experts as attractive persons and apparently authoritative figures when two experts oppose each other.[50]

Another beneficial side effect of the "light touch" regime, from the judge's perspective, was that even when a *Frye*-type determination had to be made, it was generally thought appropriate to make it in the broadest terms possible, or (which is the same thing) treat any determination of a previous judge as resolving the issue in the broadest terms possible. This is well illustrated by the way in which bitemark identification became generally accepted as a proper subject for testimony in American courts. In the first case to consider the issue, the 1975 California case of *People v. Marx*,[51] the California court of appeals faced a very rare factual circumstance involving both an unusually clear bitemark and an unusually rare pattern of dentition. What that court in fact did was fashion a narrow exception to the California version of the *Frye* test, allowing admission of the evidence in the specific instance in front of it while saying simultaneously: "Concededly, there is no established science of identifying persons from bitemarks."[52] However, as one of us has written before:[53]

> [T]hereafter the *Marx* case was regularly cited by courts dealing with much more questionable applications of bite mark

---

[49] *See* Faigman et al., *supra* note 46, at 655-56.

[50] Or, to be broader and more kind, the resort to "peripheral processing" heuristics which may include these factors. For a summary of the available empirical research in regard to jury processing of expert testimony in complex cases, see Joseph Sanders, *The Paternalistic Justification for Restrictions on the Admissibility of Expert Evidence*, 33 SETON HALL L. REV. 881 (2003).

[51] 126 Cal. Rptr. 350 (Ct. App. 1975).

[52] *Id.* at 353.

[53] Risinger, *Navigating Expert Reliability*, *supra* note 3, at 138. Notes 54-56 are retained from the original work but renumbered sequentially with those of this Article.

> identification, without noting *Marx's* particular facts.[54] In the normal way that courts have worked in defining the parameters of admissibility for proffered expertise, *Marx* came to be read as a global warrant to admit bite mark identification evidence whenever a person displaying apparent credentials chose to testify to an identification. Perhaps the most notorious such case was the very next full-scale examination of bite mark evidence, the Illinois case *People v. Milone*,[55] which, relying at least in part on *Marx*, declared bite mark evidence acceptably reliable under much less clear conditions.[56] After *Marx* and *Milone* there was little serious consideration given to bite mark foundational dependability by subsequent courts . . . .

Such an approach yielded a judicially attractive result: a broad admissibility warrant resulting in effortless and time-efficient decisions on admissibility by invocation of precedent without any requirement of thought. The fact that the broad warrant might authorize the admission of much that was very unreliable did not appear to cross the judicial mind. In any individual case where a particular judge might be troubled by questions of reliability, there might be other tools that could be used to achieve an individually satisfactory result (such as manipulation of the scope of expertise applied in the particular case). Or the court might salve its conscience by contemplating the supposed universal solvent of cross-examination (rarely an effective solution in such cases, but what the heck).[57] However, explicit and systematic consideration of reliability

---

[54] *See, e.g.*, People v. Sloane, 143 Cal. Rptr. 61, 69 (Ct. App. 1978) (relying on *Marx* to establish the general reliability of bitemark evidence).

[55] 356 N.E.2d 1350 (Ill. App. Ct. 1976). Notwithstanding the admitted controversy concerning the reliability of bitemark identification among forensic odontologists both at the trial and in the literature of the time, the *Milone* court found the *Frye* general acceptance standard had been met, citing *Marx*. *Id.* at 1359-60. *Milone* remains controversial. The defendant has been released, but continues to maintain his innocence and attack the bitemark evidence. *See* Milone v. Camp, 22 F.3d 693, 697 n.1 (7th Cir. 1994) (noting that Milone was released after serving almost twenty years of his 90- to 175-year prison sentence). In addition, there is good evidence that another person actually committed the murder, a person whose bitemarks have been judged by at least one forensic odontologist to be as good a match for those on the victim as Milone's. *See id.* at 700-01 (noting that the bitemarks found on the victim match the dentition of known serial killer Richard Macek and that Macek confessed to the murder several times prior to his 1987 suicide).

[56] *See Milone*, 356 N.E.2d at 1355-56, 1360 (upholding the trial court's decision to allow bitemark identification testimony even though four forensic odontologists testified to the unreliability of such positive identification).

[57] Compare the faith in cross examination of experts manifested in *Daubert*, 509 U.S. at 593, with James M. Shellow, *The Limits of Cross Examination*, 34 SETON HALL L. REV 317 (2003).

generally was not a fundamental part of the job of judging.

Daubert and its progeny have changed all that, and not just for the federal courts. Evidence is accumulating that *Daubert* has changed the standards applied to expert reliability not only in states explicitly adopting its approach, but also in many states explicitly claiming to eschew it.[58] Exactly how things have changed is less clear, however.

KUMHO TIRE AND THE NEW REGIME OF EXPERT RELIABILITY GATEKEEPING

In many ways, *Kumho Tire*[59] is the most important of the *Daubert* trilogy, not only for making clear the Rule 702 gatekeeping obligation in regard to non-science "experience-based" expertise[60] (the point for which it is best known), but perhaps more importantly for what it says about the general construction and proper approach to the requirements of Rule 702 even in regard to the products of science. Whether one is examining scientific expertise or not, three points come though clearly from *Kumho Tire*: First, a court must review the reliability of the proffered expertise specifically as it applies to the task for which it is being utilized in the litigation in which it is offered, not in some more global sense.[61] Second, a court is obliged to think about and select the most appropriate criteria of reliability for the kind of expertise being proffered, given the circumstances of its generation in the particular case.[62] The authority to honestly make that inquiry (and, to our minds, only that authority) is the essence of the "flexibility" and "discretion" referred to in the *Kumho Tire* opinion.[63] Third, in regard to all expertise (even that

---

[58] *See States Move to* Daubert*, Even When They Say They're Stuck on* Frye, 2 EXPERT EVID. REP. 161 (2002).

[59] 526 U.S. 137 (1999).

[60] "We conclude that *Daubert's* general holding—setting forth the trial judge's general gatekeeping obligation—applies not only to testimony based on 'scientific knowledge,' but also to testimony based on 'technical' and 'other specialized' knowledge." *Id.* at 141

[61] This position is explained at length with extensive quotations from the *Kumho Tire* opinion in D. Michael Risinger, *Defining the "Task at Hand": Non-Science Forensic Science after* Kumho Tire Co. v. Carmichael, 57 WASH. & LEE L. REV 767, 773-75 (2000) [hereinafter Risinger, *Task at Hand*]. *See also* Joelle Anne Moreno, *Beyond the Polemic Against Junk Science: Navigating the Oceans that Divide Science and Law with Justice Breyer at the Helm*, 81 B.U. L. REV. 1033, 1049-60 (2001).

[62] *See* Risinger, *Task at Hand*, *supra* note 61, at 774.

[63] "The objective . . . is to ensure the reliability and relevancy of expert testimony." *Kumho Tire*, 526 U.S. at 152. The Court stated that while the relevant reliability inquiry "should be 'flexible,'" the "'overarching subject [should be] . . . validity' and reliability." *Id.* at 158 (quoting with approval the opinion of the district

based on claims of scientific authority), while a court may consider the famous (or infamous) "*Daubert* factors,"[64] the presence or absence of one or more is not necessarily dispositive of sufficient reliability to gain admission.[65] What is important is the honest exposition by the court of an appropriately strong reason to believe that the proposed product of expertise was generated by a process that will generally yield a result sufficiently reliable so that the official ends of the proof system,[66] will not be undermined by admission of evidence of that level of reliability on that kind of issue in that kind of case.

We say that *Kumho Tire* "says" the above things, but the saying is more explicit as to some things than others. As to the first, we believe that there can be no serious doubt that narrow (rather than global)

---

court). The Court further indicated that the inquiry should be directed toward some "set of reasonable reliability criteria." *Id.*

[64] Referring to "four" factors has become standard, though the real number of factors is subject to debate. The *Daubert* opinion spake thus, without numbering factors: "a key question [in regard to a theory or technique] . . . will be whether it can be (and has been) tested." 509 U.S. 579, 593. "Another pertinent consideration is whether the theory or technique has been subjected to peer review and publication. Publication (which is but one element of peer review) is not a sine qua non of admissibility . . . ." *Id.* "Additionally, in the case of a particular scientific technique, the court should consider the known or potential rate of error . . . and the existence and maintenance of standards controlling the technique's operation." *Id.* at 594. "Finally, 'general acceptance' can yet have a bearing on the inquiry." *Id.* These were summarized in *Kumho Tire* as "several factors" without numbering, but with four bullet points. 526 U.S. at 149-50. However, it is easy to separate whether a claim "can be tested" (its empirical nature or theoretical falsifiability) and the degree to which it has been subjected to actual testing, into two separable but nested factors. In addition, the potential rate of error is arguably always 100 percent in the absence of some kind of testing (though not necessarily the kind of formal testing that would lead to more specific and quantifiable knowledge of an error rate). Knowledge of error rates is thus a product of testing. In addition, can "standards of control" for a technique's operation be a relevant factor if there is no reason to believe such "standards" enhance reliability? This too would seem to be a question of testing, at least in some contexts. Finally, a fortiori "general acceptance" is the product of peer review, so one can argue that there are really eight explicitly referenced "*Daubert* factors" (falsifiability, testing, peer review, publication, potential error rate, known error rate, standards of practice, general acceptance) or only three (falsifiablility, testing which reveals error rate, peer review). In addition, the *Daubert* Court invokes the relevance-based concept of fit, 509 U.S. at 591, which is perhaps best seen as an analogue to "external validity," and which can easily be asserted as a fifth (or ninth, or fourth) "*Daubert* factor." *See supra* note 43 and *infra* note 70. Courts have not always referred to four *Daubert* factors, either. *See, e.g.*, United States v. Crisp, 324 F.3d 261, 266-67 (4th Cir. 2003) (five factors); United States v. Prime, 220 F. Supp. 2d 1203, 1204 (W.D. Wash. 2002) (five factors); United States v. Griffin, 50 M.J. 278, 284 (A.F. Ct. Crim. App. 1999) (six factors).

[65] *See Daubert*, 509 U.S. at 593.

[66] That is, the policies identified as the ends of the system in FED. R. EVID. 102 (truth determination and justice, in the sense of proper application of law to accurately determined facts).

reliability is the explicit theme that drives the entire opinion.[67] The second—the thoughtful selection and application of criteria—is also fairly explicit, given that the court makes it clear that such flexibility and "latitude" as it refers to is to be utilized in the service of determining and utilizing "reasonable measures of reliability."[68]

The third principle requires a bit more exposition, but seems to us a product of fairly necessary implication. The court emphasizes that the "four factors" may be considered in determining the reliability warrant of non-scientific expertise when it would be reasonable to do so given the nature of the expertise under examination.[69] The court further emphasizes (by way of reiterating a part of *Daubert* often unfortunately ignored) that the four factors were not each necessary conditions for a proper reliability warrant, nor were other factors foreclosed.[70]

Thus, on our interpretation of *Daubert* and *Kumho Tire*, the formulation of the reliability issue in regard to proffered expert testimony in every case must take the form, explicitly or implicitly, of the following four-part question, which might be said to embody a different, more general, and perhaps more generally helpful, four-factor approach than the one that has too often been mechanically derived from *Daubert*. These may be best captured initially when set out in the form of a question, thus:

> In regard to any proffer of expertise, is there good reason to believe that the proffered product of the claimed expertise (given its specific form and the methods and conditions of which it is a product) provides the jury with appropriately reliable information on the case-specific question upon which the expert is proffered?

Each of the four parts of this question must be given content-specific consideration with regard to the individual case, but it is the

---

[67] *See* Risinger, *Task at Hand*, *supra* note 61, at 774-76.

[68] *See Kumho Tire*, 526 U.S. at 153. One should note that the Court in *Daubert* also observed that though the inquiry envisioned by Rule 702 was a flexible one, its "overarching subject is the scientific validity – and thus the evidentiary relevance and reliability – of the principles that underly a proposed submission." 509 U.S. at 594-95.

[69] *Id.* at 150.

[70] *Id.* at 151. This is important to keep in mind when one realizes that none of the "four factors" addresses very specifically the criteria by which science itself would evaluate the belief warrant for the reliability of scientific data and its implications in regard to a question different from the narrow question addressed in the experiment or study which generated the data—questions generally referred to by the labels "internal validity" and "external validity." *See infra* notes 79-80. The four factors have too often been deadweights woodenly applied, inert impediments to the development of a sophisticated approach by the courts to belief warrants for scientific evidence.

proper framing of the second and fourth parts of the question which give initial structure to the court's gatekeeping task, and contextually guide the answers to the particular issues raised by the first and third parts of the question. In the proper performance of this judicial function, one must start with the second and fourth inquiries, and it is often most helpful to start with the fourth. For this reason we will number and label the various parts of the complex question set out above as follows:

> 1. Framing the case-specific target issue.
>
> 2. Framing the case-specific claim of expertise.
>
> 3. Determining what available information bears on a rational belief warrant in regard to the reliability of the specifically claimed expertise.
>
> 4. Determining the proper case-specific legal standard of certainty for such a belief warrant.

How these inquiries should be conducted in particular cases can perhaps be best illustrated by starting with a toxic tort hypothetical, somewhat simplified, but not too removed from the real world.

IDENTIFYING THE TASK-SPECIFIC RELIABILITY QUESTION FOR EXPLICIT PRODUCTS OF SCIENCE

Assume that a plaintiff is claiming that her deformed foot was caused by exposure to a compound, legally prescribed to and ingested as a drug by her mother during pregnancy, which goes under the trade name Benediction.[71] In order to establish the legally required element of causation, she intends to call (perhaps alone, perhaps among other intended experts) a toxicologist who will base his testimony on the role of *in utero* exposure to Benediction in causing abnormalities of the feet in humans on animal studies involving the administration of Benediction to lab rats.

1. Framing the case-specific target issue

Somewhat ironically, the problem of defining the appropriate scope of the case-specific reliability issue is least in that troublesome area that precipitated the *Daubert* revolution, "increased risk" causation in toxic tort. In each case, the target issue is more or less self-defining. No judge would be tempted to define the issue in the

[71] The play on the name of the controversial real-world compound Bendectin is obvious, but the reader should not lose sight of the fact that Benediction is a fictitious compound used only to illustrate the process of framing the reliability question.

Benediction birth defect case as, "Is there good reason to believe that the proffered product of claimed expertise provides appropriately reliable information on whether an ingested chemical can cause birth defects?" This part of the target "task at hand" question automatically coalesces around the particular chemical or biological agent attacked as a causal agent by the plaintiff. Nor is it likely that the issue ought to be drawn in regard to a particular subclass of Benediction, given the general fungibility of effect for chemical compounds of the same formula. While there might perhaps be inference issues concerning, for example, what to make of data on causation from compounds closely related to Benediction chemically, those are external validity questions that go to the belief warrant for the relevancy of the proffered evidence, not to a primary part of the target issue itself.

Similarly, the real question under investigation in such a case cannot be "Did Benediction cause this particular plaintiff's birth defect?" (in anything but a purely formal or notional sense[72]), because such a question is unanswerable, at least in the ways we find most satisfying when dealing with causation in everyday commonsense terms (causation thought of by a sort of "mechanical linkage" metaphor which is most comfortably understood in situations of physical trauma[73]). This is for two reasons: first, because it is exceedingly rare that every exposure to such a claimed causal agent is followed by the asserted effect, and second, because such birth defects have a background rate of natural occurrence, and, at least in the current state of knowledge, there is no means of knowing whether the plaintiff's defect was one of the ones which was going to occur anyway even absent exposure to Benediction. In other words, such cases cannot yield satisfying answers to questions of "but-for" causation.[74] At least in the present state of our knowledge, they often manifest what appears to be significant randomness in any causal linkage that exists, and this element of the random might conceivably

---

[72] That notional sense is still embodied in the black letter of the law. *See* RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL HARM (BASIC PRINCIPLES) § 28(a) (Tentative Draft No. 3 Apr. 7, 2003) [hereinafter RESTATEMENT, T.D. 3].

[73] Such a mechanical linkage model works fine for many everyday interactions, but can easily become what Shadish, Cook & Campbell refer to as "a billiard ball model that requires commitment to deterministic causation or that excludes reciprocal causation," observing that such a model is "a caricature of descriptive causation that has not been used in philosophy or in science for many years . . . ." SHADISH ET AL., *supra* note 44, at 465.

[74] And therefore, in such cases, issues of general and specific causation collapse into one another in the individual case, in that, at least once individual exposure has been established, both are to be inferred simultaneously from exactly the same evidence (data).

be found to be part of the linkage phenomenon even if we had perfect knowledge.

So our real expert inquiry must be something like, "How much did Benediction exposure raise the probability of the plaintiff being born with her birth defect?" However, this is still underspecified. Virtually all such claimed relationships, when they are shown to exist, exhibit substantial "dose effects," that is, variations in induced risk depending on the amount of exposure or dose level.[75] Because this is true, we should take it into account in framing the "task at hand" target issue, which thus becomes, "Given the level of exposure of the plaintiff to Benediction as shown by the other evidence, how much does that exposure raise the probability of birth defects?" But this also remains potentially underspecified, since many teratogens raise the risk for some classes of birth defects a lot and other classes of defects almost not at all. So the final iteration of the empirical question becomes, "Given the plaintiff's exposure level as shown by the other evidence, how much does such exposure raise the risk of birth defects of the kind exhibited by plaintiff?"

There are still hard issues left, of course. The first one is a legal issue: What rise in risk (increased probability of birth defects) is enough to fasten liability onto the defendant? This is a "question of law" (or of legal policy) of the normal type. The most popular answer currently is a doubling of risk,[76] because given a large number of cases, at least half the people paid in such cases are paid by those who should pay.[77] A lower required risk increase results in the

---

[75] *See* Bernard D. Goldstein & Mary Sue Henefin, *Reference Guide on Toxicology*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 406-09 (2d ed. 2000) [hereinafter REFERENCE MANUAL]. Even in the so-called "no threshold model," not every exposure actually causes an effect. *Id.* at 407. The model merely posits that any exposure *might* cause an effect. Note that dose is a function both of intensity and duration of exposure. RESTATEMENT, T.D. 3, *supra* note 72, § 28 reporter's note, at 163-64 (citing authorities). Note further that the handiest short summary of current issues in causation in the toxic tort settings is to be found in this reporters note.

[76] RESTATEMENT, T.D. 3, *supra* note 72, § 28 reporter's note, at 180-82 (citing authorities).

[77] This is at least true in regard to "physical deformity" birth defects such as the one in the hypothetical, which by definition have a pre-birth onset. As to other conditions, such as cancer caused by a teratogen, some percentage of the group (which particular individuals are unknown) would have gotten the cancer anyway, and thus they have had their cancer accelerated (maybe a lot, maybe a little) instead of being individuals who got cancer and otherwise would have been free of such cancer throughout their lives. *See* Sander Greenland & James M. Robins, *Epidemiology, Justice, and the Probability of Causation*, 40 JURIMETRICS J. 321 (2000). Whether and exactly how the possibility of "mere" acceleration should affect the legally required standard of sufficient risk increase, or whether it should be viewed as a remedies issue, is unclear.

majority of people who are paid out of defendants' pockets being the people who would have gotten the condition without the defendants' input, and to that degree gives that set of plaintiffs (whose individual members cannot be identified) a windfall. This makes defendants become insurers for damages they did not cause. A higher required risk increase results in injured parties not recovering who ought to be paid by defendant, which raises both justice and efficiency problems. These are persuasive arguments to many on what the standard of the law ought to be, but there are other arguments which can be made in favor of both higher and lower risk increases as standards of legal responsibility.[78] The point here is not which legal standard is best, but merely that this decision has nothing to do directly with control of the reliability of expertise, or the kind of information which experts can justifiably provide.

A related legal issue concerns the question of how sure the factfinder must be concerning the legally designated risk increase in order to reach a verdict, but that is again a legal issue of applicable standard of proof. Difficult as such issues may be, they remain questions of law for the judge (or the legislature) outside the domain of any expert. It will be necessary, however, for the judge to be clear on the applicable legal requirements at the point of framing the reliability question, because that legal requirement forms a part of the question to be answered by reference to the proffered expertise. Thus, the final form of the target issue becomes, "Given the plaintiff's exposure level, did that exposure raise plaintiff's risk of developing the deformity she now has by at least a factor of two (or 'a significant amount' or 'by a factor of ten' or 'to a reasonable certainty,' that is, whatever the applicable legal standard may be that will be treated as establishing causation)?"

2. Framing the case-specific claim of expertise

Having framed the empirical target of the proposed expertise, the court will then face the problem of expressing in exact form the expert claims that are said to bear on the target issue. In risk-increase toxic tort cases, this will almost certainly be information of varying

[78] A lower required risk increase might promote care and safety, a higher required risk increase might encourage the development of more specific evidence concerning sub-populations to separate out the sub-populations in the large-set data that can be identified as having different relative risks. Specifically, this would encourage efforts to tailor research to narrower populations, thus reducing the possibility of "substructuring" (the uneven distribution of risk within a sampled population), and yielding more confidence in the applicability of the risk numbers to the individual case.

quality from a fairly limited number of domains of inquiry, such as epidemiology, experimentation on animals, information on the effects of compounds claimed to be related to the compound in the case, etc. All of these proffers will have a significant claim to being "scientific evidence," that is, the products of some part of normal science practice. Indeed, given the nature of the target inquiry in such a case, it would seem extremely unlikely that any evidence would be proffered that was not claimed to be scientific in this sense.

One characteristic of any testimony that might arguably be deemed the product of science is that it will be traceable back to data. In addition, in most parts of science (and certainly all those that might provide reliable information on risk increase in toxic tort), the testimony will be traceable back to formal data. By formal data we mean the products of some organized regime of observation where both the observation protocols and the results are explicit and objectively recorded in some way, and therefore both potentially available and replicable. The information in the proposed expert's testimony may reflect data only at second hand, or in summary, or in some sort of combination conceptually. It may reflect extension, induction, and abduction away from the data. But to the extent that, upon examination, it cannot be traced back to formal data in some way, it cannot presume to wear the mantle of science.

The first necessary step, therefore, is to describe, at least in general, the kind of data upon which the proposed expert's testimony is based. In our hypothetical, we have said that the proffered expert proposes to testify about the results of (data generated by) studies of Benediction administered to pregnant laboratory rats, the observed rates of deformity of the extremities with and without such exposure in resulting offspring, and also the implications of those studies for forming a judgment about the likely risk increase, if any, in humans exposed to Benediction, as well as the magnitude of such increase. This kind of testimony may immediately be seen to have two significantly different aspects: one dealing with the validity and meaning of the study data in their own terms (i.e., Were the studies undertaken by procedures that allow us to say reliably what the data mean concerning effects or associations in the groups of animals tested?), and the other dealing with reasoning from the data to groups other than the one tested (i.e., Would the data distributions hold for other animals of the same species, other animals of different closely related species, other animals not so closely related, higher or lower dosages as a ratio of body weight, in each of these groups, etc.?).

The first of these aspects of science-based testimony raises

questions about what is called "internal validity"—essentially, "Are the data any good in their own apparent terms, given the way they were generated?"[79] The second of these aspects raises issues of "external validity,"[80]—that is, "How far are we justified in taking the data, assuming them to be accurate in their own terms, and in drawing implications or conclusions in other contexts either very closely related (other members of the same sex and age of the same species) or much further away (say, humans)?" If you hear an echo of the problem of "area of expertise" and the Ford Taurus mechanic, you would not be wrong.

As to internal validity, it is accomplished by adhering to procedures designed in advance to eliminate, to the extent possible, potential confounding factors.[81] Science in general has a few general norms of such good practice, and each particularized special area of inquiry supplements the general norms with others purpose-built to reduce or eliminate the particular threats to internal validity that reflection has suggested as the most dangerous in that particular domain.[82] Internal validity depends on good study design and execution.

The essence of external validity is not so easy to suggest. No one

---

[79] This is close enough for our purposes without getting too deeply into the technical controversies that have surrounded the notion in the land of its birth. The concept was formulated by the late eminent psychologist Donald Campbell in the 1950's in the context of cause-and-effect studies, and in that tradition is generally limited to such studies. SHADISH et al., *supra* note 44, at 37. It was originally contrasted with "external validity" in much the same terms as are used in the text, that is, external validity referred to extension of conclusions to contexts outside the narrow bounds of the actual study. *Id.* In later refinements, the Campbellian taxonomy of validity for causal studies expanded to include two more validity components, "statistical conclusion validity" and "construct validity." Statistical conclusion validity refers to the validity of statistical conclusions derived from data in individual studies. *Id.* Construct validity refers to the validity of constructs suggested by an individual study or studies. A strong argument can be made that statistical conclusion validity can best be viewed as an aspect of internal validity and that construct validity is best viewed as an aspect of external validity. However, Campbell and his school view them as independent. *Id.* It is unnecessary for present purposes to explore this more fully. Also, for present purposes, it makes sense to apply both external and internal validity notions to skills testing contexts as well as cause-and-effect contexts. (There is an argument by which one can equate or convert skills tests to cause-and-effect tests, but it is beyond the scope of this footnote and this Article.)

[80] *Id.*

[81] *Id.* at 39-42.

[82] For instance, in any inquiry using humans as apprehenders or interpreters, one such principle is "keep the process of data collection and analysis as blind as possible for as long as possible." Robert Rosenthal, *How Often Are Numbers Wrong?*, 33 AM. PSYCHOLOGIST 1005, 1007 (1978).

believes that internally valid results of studies have meaning only for the exact entities in the exact universe studied. Everyone is easily convinced that they may be generalized to other universes not studied, as long as they are populated by entities exactly like the ones in the studied universe. One of the bases for the scientific success of physics and chemistry is their good fortune at their foundations to be dealing with universes of entities that are virtually completely alike, and therefore fungible when it comes to generalizing from data (you seen one gamma ray or atom of a hydrogen isotope, you seen 'em all, pretty much literally for most purposes[83]).

When one moves into biological systems, however, such fungibility starts to break down, but in what patterns and for what purposes, it is not usually clear. This raises difficult questions about principles of external validity for study results, such as: "When are results in mice going to track results in humans?" Note that the answer is not likely to be universal. Few would instinctively say "always." One might be tempted to say "never," but the effects of a hydrogen bomb on a mouse at ground zero is likely to correspond to the effects on a human pretty completely. Ultimately, it is an empirical question subject to investigation, but ironically it is a question that can only be answered completely by doing studies that render the question moot for most purposes, since you have to establish the effects on humans by direct empirical study in order to be sure of the correspondence, and then you do not need to reason from the animal model for that effect. In some animals, however, there can be enough experience for suggestive patterns to develop which indicate the tenability and strength of reasoning by extension in regard to at least some classes of phenomena in humans.[84]

On some questions, some kinds of studies have a validity advantage because they come close to dealing with the question under investigation without the necessity of too much extension. This is why epidemiology has been viewed as epistemically privileged on issues of risk increase in toxic tort. A well-designed epidemiology study of the very agent in question at the same exposure levels as are involved in the case at bar comes close to answering the risk-increase question directly. The key term here is "well-designed." Epidemiological studies are at least as difficult to design in internally valid ways as other kinds of studies[85] and may suffer from potential

---

[83] Except for the loose use of the term "see."

[84] Goldstein & Henefin, *supra* note 75, at 410-11.

[85] Michael D. Green, D. Michael Freedman & Leon Gordis, *Reference Guide on Epidemiology*, *in* REFERENCE MANUAL, *supra* note 75, at 354-73.

confounds involving sampling bias,[86] improper control designs,[87] and low statistical power resulting from inappropriately small sample sizes for the potential effect size being studied,[88] among others. The results of bad epidemiology with poor indices of internal validity have no claim to any privileged position just because its bad data would be more directly meaningful to the question under investigation if it were good. (This illustrates that internal and external validity are not theoretically independent, but rather they are nested. Internal validity can exist in theory without external validity, but there can be no external validity without internal validity.)

Good epidemiology is difficult to design and very expensive. In many toxic tort claims, good epidemiology is just not available. It is one thing to say that good epidemiology trumps other sources of information when available. It is another to say that questionable epidemiology should lock out other relevant information from other domains of inquiry, and yet another to say, as some courts have, that only epidemiology can form the basis of sufficient and sufficiently reliable information to take a risk-increase causation issue to the jury.[89]

This is not to deny that the presence of internal and external validity problems can so undermine the reliability of proffered testimony claiming to be the product of science that it is not sufficiently reliable to deserve entry to the courtroom at all. This would be especially true where there was a consensus in the generally relevant scientific communities that the data were unreliable or the extensions unjustified even for provisional belief, or that more apposite evidence rendered the proffered conclusions highly unlikely. It seems prudent, not only to find that such substantial outliers provide insufficient evidence even to support a preponderance verdict alone,[90] but also to exclude them in order to

---

[86] *Id.* at 355-56.

[87] *Id.* at 363-64.

[88] *Id.* at 362.

[89] Such a so-called "epidemiologcal threshold" was first employed in *Brock v. Merrell Dow Pharm., Inc.*, 874 F.2d. 307, 315 (5th Cir. 1989) (construing Texas law). Most courts properly reject it. See, for example, the extensive authorities collected in RESTATEMENT, T.D. 3, *supra* note 72, § 28 reporter's note, at 170-71.

[90] As Christopher Mueller points out, there is nothing wrong with granting summary judgment after a "*Daubert* hearing" because the proffered expert testimony is so unreliable that it would not be reasonable for a jury to base a verdict on it. Christopher B. Mueller, Daubert *Asks the Right Questions: Now Appellate Courts Should Help Find the Right Answers*, 33 SETON HALL L. REV. 987 (2003). Perhaps courts in such cases would be better advised if they simply said this instead of declaring such evidence "inadmissible" under Rule 702, then granting summary judgment for having no evidence. Courts should avoid making admissibility determinations or

insulate the jury from potential reliance on them even when there is other evidence, and this seems to be what was envisioned in *Daubert* and *Kumho Tire*.

However, some residue of questions in both spheres of validity will be present in the most reliable scientific evidence. Studies are rarely perfect, or perfectly on target, and the fact that available studies are not perfect or require some extension to apply to the target causation issue should not automatically result in exclusion of testimony based on them, no matter how forcefully their imperfections are pointed out.

It is not uncommon for causal relationships to be inferred by the convergence of information from various domains at some remove from the target issue, where the product of no single domain could be said to be a reliable indicator of causation by itself. This is not surprising. It is the normal way of circumstantial evidence, building walls by bricks in ordinary trials. When there are interlocking and mutually corroborating results from a variety of domains and studies that individually are all subject to plausible external validity objections, it would seem that exclusion based on external validity grounds ought to be approached with caution and an attempt at sophistication. We are not saying that some external validity leaps are not so great that they would form the proper basis for excluding any proffers of expert testimony based on them. When the leap is closer, however, and there is not much valid affirmative counterevidence from more epistemically privileged domains, it would seem that external validity issues in toxic tort causation cases would be better controlled by sufficiency judgments rather than threshold admissibility judgments, or left to the jury, especially given the preponderance standard of proof obtaining in civil cases.

3. Determining what available information bears on a rational belief warrant in regard to the reliability of the claimed expertise ("good reason to believe")

Thus far, in our hypothetical toxic tort, we have filled in the generally applicable question as follows: "Is there good reason to believe that a witness with training and credentials in toxicology, basing his testimony explicitly in large part on formal data from animal studies involving the exposure of white mice to large amounts of Benediction, can give testimony of sufficient reliability for the purposes of the law on the issue of whether the risk of congenital foot

---

declarations except in cases where admissibility would actually make a difference.

deformities in humans at least doubles as a result of *in utero* exposure to Benediction of the kind involved in this case?" We still must deal with how a court is to approach the issues entailed in the phrase "good reason to believe [that the proffered expert testimony] is sufficiently reliable for the purposes of the law." Let us start by asking what constitutes "good reason to believe," that is, what is sometimes called a proper belief warrant, in regard to information such as this.

We are lucky in that this rather deep question of philosophy is comparatively easy to answer satisfactorily in regard to proffered expert evidence about risk-increase causation in toxic tort, and doubly lucky in that addressing the question in this context will give us leverage which will be useful in dealing with questions of proper belief warrant in regard to other kinds of claimed expertise.

A reasonable determination of the proper factors that go into a warranted belief in risk-increase causation claims is comparatively easy because virtually any such information will of necessity be the product of science narrowly defined, as explained above. And therefore:

> A. If we believe that information properly generated by the methods required of practitioners of science by the applicable practice norms of the area in which they operate has a high claim to reliability,
>
> B. Then it is appropriate to look to that science practice itself for the proper variables affecting warranted belief for such a claimed product.

These sound a bit circular, but they are not. The first merely asks whether science has a privileged claim to reliability on some types of questions. We need not undertake an extensive discussion to conclude that an affirmative answer is judicially noticeable in regard to issues of empirically observable fact and the structures of generalization (theories) that are built on them. Indeed, the *Daubert* Court's invocation of the concept of falsifiability seems to have been intended to suggest the scope of the domain of inquiry that can in theory have a claim to being "scientific." Therefore, all that remains is to determine the factors that go into a proper belief warrant, both as to internal and external validity, in the area of science which generated the data upon which the proffered expert is relying. Of course "all that remains" is easy to say, but the task is far from trivial. It is, however, doable. Each area of real science will have an associated literature dealing with proper methodology and issues raised by various threats to validity. Some of these will be common to most science (for instance, the requirement of procedures to guard against observer effects when human perceptors, raters, or evaluators

must be used).[91] Others may be specific to a given area of inquiry (dose level issues in toxicology, especially as they relate to generalizations from animal data, for instance). In every case, the court is of necessity obliged to become acquainted with these criteria sufficiently to evaluate the general strength of the testimony's claim to reliability, that is, why and to what degree one would be warranted in believing what the expert asserts.

This was what the *Daubert* Court appears to have believed it was attempting generally in the outlining of the (in)famous "four" factors:[92] (1) falsifiability or testability, and testing (whether a claim "can (and has been) tested"); (2) establishment of (potential or actual) error rates; (3) peer review and publication; and (4) general acceptance. Of those factors, only the general notions of "the extent to which a claim has been tested" and the establishment *vel non* of "known error rates" (which is just a byproduct of what would be required to count as adequate testing in many areas of scientific inquiry) approach directly an evaluation of validity. We take these to mean little more (and no less!) than the proposition that any area unconcerned with testing and error rates has no claim to scientific validity in the sense of being a proper product of science. Whether there can be some other basis for a belief warrant for such non-science-based expert claims is an issue we will come to in due course.

The other two *Daubert* factors attempt to use peer evaluations of validity (either by reference to a small group of pre-publication reviewers, or a larger community) as an alternative source of information to the court's own evaluations. Unfortunately, like the *Frye* test before them, these factors provide weak warrants in areas with low claims to validity under the general norms of science but with a guild structure that allows them to claim peer acceptance both in regard to publication and specific review, and in regard to community acceptance.[93]

The "*Daubert* factors" are usually supplemented by *Daubert*'s further requirement of proper "fit" between the data and the problem presented by the case.[94] Here, the Court seems to have been getting at something akin to problems of external validity and extension. Taken together, these general "*Daubert* criteria" have

---

[91] *See* Risinger et al., *Observer Effects*, *supra* note 18, at 9.

[92] For a fuller analysis of the three or four or five or eight or nine "factors," see *supra* note 64.

[93] The Court in *Kumho Tire* recognized as much when it said that the general acceptance factor was not a good indicator of reliability "where the discipline itself lacks reliability." 526 U.S. at 151.

[94] *See supra* notes 64, 70.

often been treated as a mechanical checklist by both lawyers and judges with all the skill displayed in the product of an undistinguished paint-by-number picture.[95]

Perhaps it is unduly harsh to be too critical of the *Daubert* Court's less-than-perfect framing of these reliability considerations. After all, it was a brave maiden voyage into waters uncharted at least by the law, being undertaken by admirals unfamiliar with even the kind of hazards that might lie there. To its credit, the Court did say that none of the factors were either necessary or sufficient, but this was lost by many who followed what they took to be their sailing directions. It might yet be possible that *Kumho Tire*'s emphasis on seeking the best criteria of validity reasonably applicable to the particular proffer of expertise under challenge (perhaps supplemented by utilization of neutral experts to educate the judge on the reliability problems specific to particular areas of scientific inquiry involved in the case) can lead in the direction of more defensible consideration and use of information actually bearing on the validity of proffered scientific evidence (when actual scientific evidence is really what is being proffered). However, even this may not always lead to better rulings on "*Daubert* motions," because there is still one criterion to be examined which, poorly handled, can lead to questionable results.

4. Determining the proper case-specific legal standard of certainty for such a belief warrant ("sufficient reliability for the purposes of the law")

Even if the whole validity issue is evaluated in a sophisticated and rational manner, there remains the question of "how reliable is reliable enough" for the purposes of the law, in order for the proposed testimony to be admissible. It seems to us that that issue must of necessity depend on a variety of factors, such as whether the expert will testify to a conclusion or as an educational witness to supplement the factfinder's general knowledge, whether the issue involved is a specific fact or a magnitude judgment, and so forth.[96] One centrally important factor would seem to be the underlying case standard of proof and its attendant distribution of burdens. We do not intend to discuss this extensively here, except to say that if courts

---

[95] For a particularly egregious recent example, see the two-paragraph attempt by Judge Bownes regarding the reliability of handwriting identification expertise in *United States v. Mooney*, 315 F.3d 54, 62-63 (1st Cir. 2002).

[96] A range of such variables is discussed in Risinger, *Taxonomy*, *supra* note 37, throughout, but particularly summarized at pp. 535-36.

adopt reliability standards (as opposed to reliability-affecting criteria) from science, they run the risk of excluding proffered testimony according to too high a preliminary standard when the applicable case standard is low.[97]

Take the following hypothetical: Let us assume that toxicology as an enterprise displays a great fear of ever making an affirmative claim of causation which turns out to be wrong. Toxicologists pride themselves on being able to say, "When we declare that 'A' causes 'B,' you can bank on it." As a result, they refuse even to consider basing any notion of causation on any relationship shown by data, if the relationship might have occurred by chance even one time in a thousand. They may be said to have a terror of false positives, and not to care much about overlooking relationships that others might find persuasive. (It would be hard to call these false negatives, because the toxicologists simply remain agnostic about such relationships, but they are errors of another sort, perhaps.)

Now assume the only evidence available on causation is from toxicology, and it shows an association which could not be accounted for by random occurrence one time in a hundred. The toxicologist would say the information was insufficient to draw a conclusion and, therefore, fundamentally useless. The usual scientist (who would probably find information meaningful that would only be the product of random occurrence one time in twenty, since this corresponds to the conventional .05 level adopted for statistical significance) would accept the information, and might base decisions on it that would be justified in her field. If a gatekeeper stood between the toxicologist's information and the usual scientist, and applied the toxicologist's criterion for "reliable enough for use," the usual scientist would be deprived of the information even though it was good enough for her purposes, so she would be cut off from doing with the information what it was proper for her to do within her sphere. The admission decision would have improperly prevented her consideration of information, by importing without thought the standard of "reliable enough" from the domain of the information's original generation, just because it was there.

Similarly, in the legal context, the conventional level of certainty required to say a relationship is established in the sciences is conservative compared to that represented by the normal tort standard of proof (a preponderance of the evidence), which equates

---

[97] A similar point is made more extensively in Neil B. Cohen, *The Gatekeeping Role in Civil Litigation and the Abdication of Legal Values in Favor of Scientific Values,* 33 SETON HALL L. REV. 943 (2003).

the disvalue of false positives and false negatives. Hence, adopting too high a standard, even by claiming to import it from the science that gave rise to the data, runs the risk of depriving the person with the burden of production and persuasion reasonably reliable information on the issues of the case. Such concerns seem at least in part to have driven the New Jersey Supreme Court in its creation of special rules for dealing with reliability in risk-increase causation cases, starting with *Rubanick v. Witco Chemical Corp.*[98] and proceeding through *Landrigan v. Celotex Corp.*[99] to the recent case of *Kemp ex rel. Wright v. State.*[100]

We are not saying that there is always an easy solution to the issue of what is "reliable enough for the purposes of the law."[101] All we can do is call on courts not to be too quick to exclude proffered expertise in civil cases based on unsophisticated acceptance of the proposition that a relationship shown by the data upon which it is based is not "statistically significant," especially when there is information from multiple domains being proffered on the same issue. In addition, we think it proper to observe here that the corollary of our position is that information reliable enough to be admitted for one legal purpose, with a low attached standard of proof, is not necessarily reliable enough to be admitted for a different legal purpose with a high attached standard of proof. To put it bluntly, we believe that information properly admitted in civil cases is not necessarily reliable enough for admission by the prosecution in criminal cases. We know this position is currently looked upon as something of a heresy,[102] though there are plenty of legal contexts in which courts have applied such differential

---

[98] 125 N.J. 421 (1991).

[99] 127 N.J. 404 (1992).

[100] 174 N.J. 412 (2002).

[101] Professor Neil B. Cohen showed quite a while ago that there is plenty of room to use analogues to the notion of confidence in probability theory in approaching the concept of preponderance. Viewed this way, we may think of preponderance sufficiency as requiring fair certainty that the true value of the probability derived from the evidence falls within a range all of which is above 50 percent. *See* Neil B. Cohen, *Confidence in Probability: Burdens of Persuasion in a World of Imperfect Knowledge,* 60 N.Y.U. L. REV. 385 (1985); *see also* Neil B. Cohen, *Conceptualizing Proof and Calculating Probabilities: A Response to Professor Kaye*, 73 CORNELL L. REV. 78 (1987). Such a metaphor is available even when part or all of the proffered testimony is derived from information not formally quantified. To the extent that Rule 702 "sufficient reliability" is in most contexts properly conceived of as more than simple relevance and less than full sufficiency, such analyses must come into play in informing that judgment also.

[102] *See, e.g.*, Roger C. Park, Daubert *on a Tilted Playing Field*, 33 SETON HALL L. REV. 1113 (2003).

standards, making admission easier in civil cases than by the prosecution in criminal cases. Nevertheless, whatever one's ultimate position on that point, we would hope to obtain universal agreement that, to the extent prosecution proffers are held to *lower* standards of reliability than those of civil plaintiffs or criminal defendants (which seems to be the case in general),[103] something is seriously out of kilter.

FRAMING THE TASK-SPECIFIC RELIABILITY QUESTION FOR "EXPERIENCE-BASED" EXPERTISE

We have given a general form of question, which we have said applies in formulating the reliability question in every case whatsoever: "Is there good reason to believe that the product of the claim of expertise being proffered is sufficiently reliable to be considered by the jury on the question (i.e., the target empirical issue upon which the expert testimony is proffered)?" We have explored the proper filling out of that question in regard to scientific evidence, properly so called, generated by the methods of science and based on formal data. Let us now see how the content of the general reliability question changes when faced with expert claims that are not, in significant and central part, the product of science, but the product of some other source of information claimed to possess an appropriate level of reliability. To make the exercise as concrete as possible, let us do so in regard to the very facts and the very claims at issue in the *Kumho Tire* case itself.

When a tire on the vehicle being driven by Patrick Carmichael blew out, the vehicle overturned, one passenger died, and others were injured.[104] The tire that blew out was old and nearly bald, with two previous improperly repaired punctures.[105] The blowout resulted from a separation of the tread plies from the carcass of the tire.[106] It was uncontested that in a non-defective tire that had never been misused at all, the tread would not separate merely as a result of normal driving that wore the tread smooth.[107] So the blowout either

---

[103] *See generally* Risinger, *Navigating Expert Reliability*, *supra* note 3. At least in the civil cases it is generally recognized that "[r]ulings on admissibility under *Daubert* inherently require the trial court to conduct an exacting analysis of the proffered expert's methodology." McCorvey v. Baxter Healthcare Corp., 298 F.3d 1253, 1257 (11th Cir. 2002). This is not commonly done regarding prosecution proffers challenged pursuant to *Daubert/Kumho*.

[104] *Kumho Tire*, 526 U.S. at 142.

[105] *Id.* at 143.

[106] *Id.* at 144.

[107] *Id.* at 143-44.

resulted from cumulative improper use (excessive sidewall flexing from wrong inflation, damage from curb impact, etc.) or a manufacturing defect, such as improperly bonded plies in the tire carcass, that took a long time to manifest itself.[108] Only the latter circumstance would render Kumho Tire, the manufacturer, liable, so the plaintiffs had the burden of producing sufficient evidence which, if believed, would justify the conclusion that the accident more likely than not resulted from such a defect. To that end, they intended to rely at trial on the testimony of Dennis Carlson, Jr.,[109] an engineer with substantial experience in the tire manufacturing industry[110] who consulted as an expert in what he called "tire failure analysis."[111] Defendant Kumho Tire claimed that it was beyond the current state of any art to assign this failure to a manufacturing defect rather than cumulative misuse.[112] Carlson, however, claimed that he could do it, that he had done it, and that the failure resulted from a manufacturing defect.[113] In reaching this conclusion, he relied upon his visual inspection of the tire and rim,[114] his experience,[115] and certain factual propositions which he claimed were true, but which the tire company claimed were not known by him or anyone else to be true in reality.[116]

The first factual claim was that the form of abuse that most commonly resulted in tread separation was long-term underinflation, which resulted in too much tire flexing while driving, and that this was so much the most common factor that other possible sources of abuse could be ignored, at least unless there was specific evidence of them.[117]

The second factual claim was that any tire which had been subject to such underinflation would always manifest some combination of four physical symptoms which could be observed: (1) treadwear on the edges of the tread greater than in the center of the

---

[108] *Id.* at 144. Presumably, both defect and abuse might have been contributing causes. This would raise issues of comparative responsibility under the applicable state law having nothing to do with the Rule 702 issue. The Supreme Court treated the 702 issue as being properly represented by the dichotomous choice, and therefore so have we here.

[109] *Id.* at 142.

[110] *Kumho Tire*, 526 U.S. at 156.

[111] *Id.* at 142.

[112] *Id.* at 145.

[113] *Id.* at 144.

[114] *Id.* at 144, 153-54.

[115] *Id.* at 156.

[116] *Kumho Tire*, 526 U.S. at 144.

[117] *See id.* at 144.

tread; (2) wear of a groove on the tire's "bead" (the part that rests against the rim of the wheel upon which the tire is mounted); (3) sidewalls with signs of deterioration such as discoloration; and (4) marks on the flange part of the rim itself.[118]

The third factual claim was that in the absence of evidence of "significant"[119] amounts of at least two of these symptoms, a defect was the most likely cause of the tread separation and blowout.[120] (This is the part that the Court later referred to as the "two-factor test";[121] it is really an "any two of four factors test.")

While Carlson conceded that there were some signs of each of these symptoms manifested in the Carmichael tire and rim, he asserted that none of them was enough to be significant, based on his experience (or in the case of the edgewear, it was not significant because the inside and outside edges were worn in differing amounts).[122]

1. Framing the target issue

In this case, the target issue is fairly easy to frame once the facts are recounted in sufficient detail: "whether, more likely than not, the tire that failed on the plaintiff's vehicle left the Kumho Tire factory with a defect that finally manifested itself in (caused) the tire failure that resulted in the accident." Note that this question is not framed to require proof of any particular kind of defect, such as improper bonding of plies, but can be satisfied if the combined probability of all possible late-manifesting defects that could result in the failure observed was greater than the combined probabilities from non-defect-caused failure due to the combined effects of normal wear and tear and occasional misuse from improper inflation, curb trauma, etc. Note further that, unlike risk-increase causation in a toxic tort case, this question does not involve any particularly fancy legal issues, except the normal ones involved in ordinary "but for" causation, where the condition for which defendant is responsible must only have been a contributing cause, not the sole cause, of the tire's failure.

---

[118] *Id.*
[119] *Id.* at 145.
[120] *Id.* at 144.
[121] *Id.* at 157.
[122] *Kumho Tire*, 526 U.S. at 144-45.

2. Framing the claim of expertise (the characteristics of the proffered product of claimed expertise, given its context and methodology)

What was the basis of Carlson's claim that he could offer reliable information of the existence of a manufacturing defect in a tire under the factual conditions applicable to this case? Carlson's proposed testimony offers a classic example of claimed expertise which will masquerade as the product of science as far as it can and take advantage of not being the product of science whenever that is beneficial. As such, it provides an incredibly important and instructive template for examining and evaluating analogous claims that are the common grist of expert evidence in many important areas. In showing why this is so, we will have to proceed rather slowly.

First, let us deal with the issue of credentials, and how they do or do not bear on the issue of whether what is being done in a particular case is a product of science, or whether it is even significantly affected by the education and experience reflected by the credentials. Carlson was an automotive engineer. What part does science play in engineering, and under what circumstances?[123] More or less by definition, engineers are people who are educated in aspects of science applicable to the design and maintenance of certain types of artifacts or products.[124] While "scientists doing science" are interested in the frontiers of knowledge, "engineers doing engineering" (all other things being equal) prefer to deal with well established principles, because they result in fewer risks in the resulting design or solution to the engineering problem at hand. Of course, this simpleminded division of labor between scientists and engineers does not define very clear boundaries in the real world. There are plenty of people with physics degrees doing "engineering," and plenty of people with engineering degrees doing research or theory-building in ways that would count as "doing science."[125] So where did Carlson fit in?

There is no reason to believe that Carlson had ever done any pertinent actual research himself, or was familiar with any body of *formal* data bearing on determining the existence of a manufacturing defect from the examination of a failed tire. Even the science he had

---

[123] This may seem like a silly question, but since a lot of what goes on in regard to expertise masquerading as partly the product of science is dependent on answers to questions like this, it is important to consider such foundational issues directly.

[124] *See* Henry Petroski, *Reference Guide on Engineering Practice and Methods*, *in* REFERENCE MANUAL, *supra* note 75, at 577. "Science in its purest form theorizes about nature as it is found; engineering at its most basic re-forms the raw materials of nature into useful things." *Id.* at 579.

[125] *Id.* at 581-84.

learned in order to become an engineer did not seem to undergird in any definable way either his methodology or his results. No doubt the question of reconstructing the original characteristics of a tire as manufactured from the remnants of a failed tire could be the subject of research generating formal data by the standards of real science. One can imagine a regime of research consisting of running tires until failure on machines incorporating various metering devices, data from which might reveal objectively discernable indices apparent in the failed tires of various original conditions that would count as defects. The point is, Carlson did not claim to be operating based on any such formal data or research of his own or of others. Both the principles of his methodology and his own performance of it, to the extent they could be said to be based on anything even arguably called data at all, were based on his own subjective observations over the course of his experience, available only to him, and in their individual form now almost certainly only imperfectly recalled, if at all. In terms one of us has developed at length in another setting,[126] Carlson claimed accuracy for a personal subjective translational system based on experience (translational in that it translates the meaning of the characteristics of the failed tire into the characteristics of the tire as it left the factory). The claim at issue was that Carlson's experience with tires had allowed him to develop, from his subjective database, four criteria for determining accurately (by his own subjective evaluation of the "significant" presence of these criteria) whether a given failed tire had been defective when it left the factory. Thus he made both methodological claims and skill claims. That is, he claimed that his "two of four" factor test, to the extent that it directed evaluation, was a reliable method, and that to the extent that reliable outcomes depended on his particular subjective judgment as part of the method, he claimed he could make those subjective judgments accurately (a claim of "skill" in making those subjective judgments). We might say that the claim can be likened to that of a cook who has written a cookbook based on experience, with many of the recipes calling for the cook to add "a significant amount" of certain ingredients. To the extent that a claim that this results in "good cooking" is based on actually following the cookbook, we would have to have a reason to believe the cookbook (the methodology) was sound. But to the extent the proper outcome was based on a claim of skill supplementing (or even substituting for) the underspecified recipes of the cookbook, we would have to have some reason to believe that the cook actually possessed the claimed

[126] *See* Risinger, *Taxonomy*, *supra* note 37, at 522-23.

skill, and we would have to know which skill was at issue in regard to each claim. In this case, we would be properly concerned with the reasons to accept Carlson's "two of four" factor test, and the reasons to accept Carlson's judgmental skill about which levels of each factor are significant and which are not, in the combination manifested by the tire in the case before the court.

The Supreme Court was very clear that this particularized approach to evaluation was required under Rule 702. The circumstances applicable to Mr. Carlson and his methodology illustrate both these points. The proponents of Mr. Carlson's testimony tried to argue that the proper way to characterize his asserted expertise was very general, consisting of expertise in determining the existence of a defect from visual and tactile inspection, and that "a method of tire failure analysis which employs a visual/tactile inspection is a reliable method"[127] since such a general approach might often be accurate. The Court rejected this approach in no uncertain terms:

> For one thing, and contrary to respondents' suggestion, the *specific* issue before the court was not the reasonableness in general of a tire expert's use of a visual and tactile inspection to determine whether overdeflection had caused the tire's tread to separate from its steel-belted carcass. Rather, it was the reasonableness of using such an approach, *along with Carlson's particular method of analyzing the data thereby obtained*, to draw a conclusion regarding *the particular matter to which the expert testimony was directly relevant* . . . . The relevant issue was whether the expert could reliably determine the cause of *this* tire's separation.[128]

And later:

> Respondents now argue to us, as they did to the District Court, that a method of tire failure analysis that employs a visual/tactile inspection is a reliable method, and they point both to its use by other experts and to Carlson's long experience working for Michelin as sufficient indication that that is so. But no one denies that an expert might draw a conclusion from a set of observations based on extensive and specialized experience. Nor does anyone deny that, as a general matter, tire abuse may often be identified by qualified experts through visual and tactile inspection of the tire. As we said before, the question before the trial court was specific, not general. The trial court had to decide whether this particular expert had sufficient specialized knowledge to assist the

---

[127] *Kumho Tire*, 526 U.S. at 156.

[128] *Id.* at 153-54 (some emphasis added). For a fuller exposition with even more extensive citation, see Risinger, *Task at Hand*, *supra* note 61, at 773-78.

> jurors "in deciding the particular issues in the case."
>
> The particular issue in this case concerned the use of Carlson's two factor test and his related use of visual/tactile inspection to draw conclusions on the basis of what seemed to be small observational differences.[129]

Thus, according to the Supreme Court, a court's job under Rule 702 is to identify the particular claim being made in the context of the particular circumstances of the case. The real question is the reliability of the specific application of claimed expertise defined by the facts of the case. So the issue was whether Carlson's particular four-part "two factor" test, coupled with his subjective evaluation of the relative magnitude and significance of each of the four factors, had been shown to be a reliable way of determining a manufacturing defect when applied to a tire, such as the Carmichael tire, which was very old and very worn.

Note here that we have not drawn the reliability issue in such a way that an answer would apply exclusively to the Carmichael tire, by including details such as "having exactly two inexpertly repaired punctures" or "being exactly 7.3 years old" or "belonging to a person whose surname begins with 'C.'" This would be artificially narrow, even given the Supreme Court's language about "this tire." To frame the reliability question so artificially as to apply to "this tire" in so restrictive a way, and not also have it apply to other tires similarly situated in regard to the claimed expertise, would be counterproductive to the policies underlying the reliability requirement in the first place. It would raise inappropriately trivial issues concerning the existence of applicable data—for example, it is unlikely that any tests that were ever done involved exactly two improperly repaired punctures—and deprive the decision of any possible precedential meaning for future cases. The problem here is one of line drawing, both in regard to the claims of expertise and in regard to the uses to which the decision might be put in the future. To solve this "over-specificity/over-generality" dilemma, we propose the following approach: The reliability issue should be framed

> A. Narrowly enough to prevent reliability from being established only by reference to evidence from a different and non-apposite part of the claimed domain of expertise; and
>
> B. Broadly enough to have some potential issue-settling or precedential carry-over effects in other cases within the class encompassed by the question (though this carry-over application would not necessarily have to be to a broad or common class in

---

[129] *Id.* at 156-57 (citations omitted).

the real world).[130]

In addition, when approaching this framing task, a court should remember that the Supreme Court has held that the main focus should be the reliability of the expertise as applied to and under the conditions of the case before the court. This militates for the narrowest framing reasonable under the circumstances, and this is what we think our final framing of the *Kumho Tire* question accomplishes.

3. Determining the belief warrant for "experience-based" expertise ("good reason to believe")

What could count as a good reason to believe the claims that underlie Carlson's testimony? We do not want to be unrealistically demanding in regard to claims of either experience-generated methodology or skills. However, at the start there seem to be only two general approaches which might distinguish the reliable from the unreliable when such claims are made: either we trust the experience of the claimant because the claimant appears to trust it, or we look for something more. Given the human capacity for self-delusion which we have all observed, the former would not appear to have much to recommend it, and indeed the Supreme Court itself has rejected the "ipse dixit of the expert" alone as a basis for a rational belief warrant.[131] Then what "something more" might suffice?

Before approaching this question, we must assert one great guiding principle which we believe to be the most powerful lens that can be brought to bear in judging circumstances put forth as supplying that "something more." At the very least, every such candidate for the "something more" which can provide a belief warrant for experience-based methodology or skill, must be capable of passing the "astrology test"; that is, it must be something that astrologers could not plausibly assert in regard to their claims.[132] We have picked astrology because it was one of the two areas actually named in *Kumho Tire* as areas "lacking reliability" generally, and because, unlike the other area (necromancy), astrology still has a large group of believers and a community of practitioners with organs of publication and guild-like groups which can provide the form, if not the substance, of publication, general acceptance, and

---

[130] The judicial system will demand some precedential effect, and rightly so on efficiency grounds, though a precedent system does not fit well with the open-ended and dynamic nature of many empirical questions. But that is a topic for another day.

[131] *Id.* at 157 (quoting *Gen. Elec. Co. v. Joiner*, 522 U.S. 136, 146 (1997)).

[132] *See* Risinger, *Task at Hand*, *supra* note 61, at 776.

conformity to community norms of good practice which are claimed to enhance accuracy. This point has immediate relevance in that, just as individual self-belief cannot provide an adequate belief warrant for others, the mutual self-belief of a group is similarly insufficient. Thus, showing conformity with group practice, without more, is not enough. This is the main weakness of both the "general acceptance" test[133] and the "equal intellectual rigor" test[134] that are sometimes put forth as generally sufficient grounds for a belief in the reliability of particular testimony. Some claims that pass muster under those tests cannot pass the "astrology test." This was recognized by the Supreme Court itself in *Kumho Tire* when it said that such factors were not in themselves sufficient when the claim is that the discipline itself "lacks reliability."[135]

With this in mind, let us continue to examine what kinds of information can yield a proper belief warrant for a claim of experience-based methods or skills beyond self-belief. There appear to be two main sources of such information: practical success and scientific testing of claims. It should not be too surprising that we believe that actual scientific testing of claims is epistemically privileged, and trumps all when it has been done and done properly. However, like epidemiology in the case of risk-increase causation in toxic tort, such testing is expensive and difficult to do across the whole range of claimed practical areas of expertise that are proffered in legal proceedings. In the absence of high-quality testing, are there ever any circumstances that can take its place and provide adequate belief warrants for the purposes of the law?

In regard to claimed expertise at determining specific facts, at any rate,[136] there would seem to be two necessary conditions for such a belief warrant: first, that in the ordinary practice of the claimed methodology or skill, there are objectively unmistakable right and wrong results in most cases of application, and second, that there is a

---

[133] *See* Peter B. Oh, *Assessing Admissibility of Nonscientific Expert Evidence Under Federal Rule of Evidence 702*, 64 DEF. COUNS. J. 556, 565-67 (1997) (explicitly advocating *Frye* test for non-scientific evidence).

[134] This test would only require that the expert have utilized the same intellectual rigor in reaching conclusions for use in court as that used for reaching conclusions in non-forensic settings. *See* J. Brook Lathram, *The "Same Intellectual Rigor" Test Provides an Effective Method for Determining the Reliability of All Expert Testimony, Without Regard to Whether the Testimony Comprises "Scientific Knowledge" or "Technical or Other Specialized Knowledge,"* 28 U. MEM. L. REV. 1053, 1063-68 (1998).

[135] 526 U.S. at 151.

[136] Weaker warrants may suffice in regard to some subjects including "no one right answer" areas such as land valuation. *See* Risinger, *Taxonomy*, *supra* note 37; *see also supra* note 96.

generally inescapable penalty for wrong results. Under these circumstances, it is at least tenable, at any rate, to believe that humans may develop generally reliable practical methods and skills. Though the practitioners may not be able to give any useful account of the reasons for their success (being only what would have been called in an earlier time "mere empericks"), if all the law cares about is the success of the methods and skills developed in such circumstances,[137] then the judgments of such cooks (or beekeepers, or chicken sexers) may be proper candidates for admission into evidence.

Of course, clearly apparent right or wrong results, and unambiguous feedback regarding success or failure, are only *necessary* conditions for a belief warrant about experience-based methods or skills; they are not in themselves always sufficient. There are other conditions which may reinforce or undermine reliability even in the presence of such conditions, and therefore affect both the tenability of belief warrants and the question of admissibility under Rule 702. However, that is an issue for another day. What we are mainly interested in here are those areas of claimed experience-based skill or practice that do not operate under such conditions of success-or-failure feedback. What may be said of belief warrants for experience-based claims like these?

When one reflects upon it, it is surprising how many experience-based claims offered in court (including that of Carlson in *Kumho Tire*, as we shall see) fit this model. For instance, it is true of nearly all the forensic identification "sciences" based on subjective human evaluation, such as bitemark, toolmark, and handwriting identification analysis. In the normal practice of these areas, there is no clear objective index of mistaken results. The practitioners do not ordinarily have empirically unmistakable feedback about the accuracy of the majority of their skill-based decisions or applications of method. And so it was with Carlson. There was no independent way for him ever to know in fact if there was a manufacturing defect in the Carmichael tire, or any other tire, when he had finished applying his methodology and coming to his conclusion. There was no way for inaccurate conclusions to manifest themselves independently, and for Carlson to suffer for them, and learn from them. In such

---

[137] Or by extension, the accuracy of such skills acquired through training in guilds which have accumulated the results of such circumstances into teachable practical models. For a full treatment of the implications and problems of such a "guild" claim, which may often masquerade as "science," see D. Michael Risinger, Mark P. Denbeaux & Michael J. Saks, *Brave New "Post-*Daubert *World"—A Reply to Professor Moenssens*, 29 SETON HALL L. REV. 405, 441-47 (1998) [hereinafter Risinger et al., *"Post-*Daubert *World*"].

circumstances, experience itself (no matter how extensive) cannot provide a proper belief warrant for the accuracy or reliability of a process or a witness.

Does this mean that no such witness can ever be shown to be reliable in what they are claiming to be able to do? No, but in the absence of a clear accuracy feedback loop, the only thing that can supply a proper belief warrant is testing, properly designed and administered according to the normal standards of science. In addition, it is important to keep in mind that any assertion that such testing has been done must be examined carefully, to make sure that the tests themselves are not only internally valid, but actually test something close enough to the skill claimed in court so that the tests have reasonable external validity in regard to that skill. And this depends not only on the broad or narrow characterization of the skill or experience-based methodology but also on the differences between test conditions and the conditions of ordinary practice.

The latter point requires a bit of exposition. The emphasis in *Kumho Tire* is on reliability of the expertise in the circumstances of the case. New Rule 702 requires that proffered expertise be the product "of reliable methods."[138] Elsewhere, one of us and his co-authors have been at some pains to establish that there is one enormous reliability-undermining condition which applies to all expertise, but most heavily distorts expertise which is experience-based and relies on human subjective judgment.[139] We refer to so-called "observer effects," particularly those which result from conditions giving rise to expectation and suggestion from which the expert has not been insulated by appropriate masking techniques. In addition, much research indicates that the distortions resulting from such unmasked suggestion and expectancy are reinforced significantly by the kind of team identification and desire to win which are virtually inevitable in the adversary process.[140] Testing of the reliability of skills which has been done in settings without such variables (which would be the norm in the usual design of such tests) cannot establish that the skills survive in the presence of the precursors of such effects.

In any area where normal practice can be adjusted to eliminate these effects by masking or blind testing regimes, courts should

---

[138] FED. R. EVID. 702(2). This rule, which became effective on December 1, 2000, requires, *inter alia*, that in order to be admissible, expert testimony must be "the product of reliable principles and methods." *Id.*

[139] *See generally* Risinger et al., *Observer Effects*, *supra* note 18.

[140] *Id.* at 18-19, 24-27.

consider the failure to do so in determining the reliability of proffered testimony, even when some arguable relevant test data under blind conditions are available.[141] However, masking is not an option in regard to some kinds of expertise. For example, in regard to accident reconstruction, the precursors of observer effects are likely to be present as long as the experts know who hired them, and this circumstance is appropriate in determining the discount to impose on any available skills test results.

Of course, in Carlson's case, things were easy. There had been no tests of any kind of his four-variable "two factor" methodology, nor of him individually as an accurate subjective evaluator. When dealing with an experience-based expertise which has no accuracy feedback loop and no even arguably relevant test data, the admissibility question should answer itself: *ex nihilo nihil fit.* And so it did in *Kumho Tire.*

4. Determining the legal standard of certainty for the belief warrant ("reliable enough for the purposes of the law")

This question need not detain us long in regard to the kind of issue presented by Mr. Carlson. What we have said in regard to risk-increase causation is equally applicable here. Carlson's claimed expertise was proffered in a civil case. It dealt with "drawing a conclusion" or "giving an opinion," that is, translating the meaning of one set of facts equally available to expert and factfinder into another non-obvious factual proposition. It did not deal with any of the special areas of concern in the theory of expertise that involve economic or normative valuations of various kinds, which in different contexts can be the subject of special arguments about heightened or lowered standards of admissibility to discharge the law's special purposes in regard to them. It was not mere "educational" expertise, expertise used only to educate the jury to the potentially counterintuitive results of relevant research which might show that the jury's general background "major premise"[142] "social framework"[143] "jury notice"[144] information was deficient or inaccurate.

---

[141] For a discussion of the feasibility of such masked regimes in forensic practice, see *id.* at 45-50.

[142] That is, the major premises for which the specific evidence in the case provides minor premises for deductions about the ultimate facts in the individual case.

[143] *See* Laurens Walker & John Monahan, *Social Frameworks: A New Use of Social Science in Law*, 73 VA. L. REV. 559 (1987).

[144] *See* John H. Mansfield, *Jury Notice*, 74 GEO. L.J. 395 (1985). These terms ("general background," "major premise," "social framework," and "jury notice" are not independent, but essentially different labels attempting to capture the kind of

All of these cases present special questions for which a court might properly claim to exercise intelligent selection, appropriate to the kind of expertise, of both the appropriate belief warrant information and the required level of reliability for admission (the proper scope of the "discretion" and "flexibility" referred to in *Kumho Tire*). However, whenever a court is faced with a proffered expert exhibiting the same kinds of attributes as Carlson, the result should be the same as that reached by the Supreme Court in *Kumho Tire*, that is to say, exclusion. However, in spite of *Kumho Tire*, this result is not always forthcoming, as we shall see when we examine the recent record of the lower courts in regard to prosecution-proffered forensic science identification expertise.

THE IGNORING OF *KUMHO TIRE* WHEN PROSECUTION-PROFFERED EXPERTISE IS CHALLENGED

A. The Handwriting Cases

Like Mr. Carlson in *Kumho Tire*, and like other forensic identification specialties such as toolmark and bitemark identification, forensic document examiners who claim to identify handwriting by comparison of hands have no unambiguous feedback regarding right or wrong conclusions in normal practice.[145] They also usually operate in non-blind conditions with no attempt to mask out the common precursors of observer effects resulting from expectation and suggestion.[146] The only initial difference between their claims and those of Mr. Carlson is that there is a guild-like group of them who share the same beliefs and general methods of examination.[147] But, as we have already indicated, the existence of a group practice, without more, does not provide sufficient warrant to believe the claims of the group, because such an approach fails the "astrology test." There is no *a priori* way to distinguish the basis for the handwriting expert's group claims from an astrologer's group claims. As we have said, in regard to Carlson, astrology, and handwriting identification, only some regime of external testing can supply the "something else" which is required for a rational belief

---

general information jurors are allowed of necessity to bring to their task from sources outside the courtroom. *See* Risinger, *Taxonomy*, *supra* note 37, at 517 n.16 and accompanying text.

[145] For a discussion of this problem, see D. Michael Risinger & Michael J. Saks, *Science and Nonscience in the Courts:* Daubert *Meets Handwriting Identification Expertise*, 82 IOWA L. REV. 21, 64 (1996).

[146] *Id.*

[147] *See* Risinger et al., *"Post-*Daubert *World*," *supra* note 137, at 441-47.

warrant.

What kind of testing regime is required is dependent in large part on the nature of the group claims being made. There are three major variables:

> 1. How many reasonably separable subtasks are performed by the group?
>
> 2. What "experience-based" "subjective data base driven" "clinical" "subjective judgment" "skill"[148] components are involved in each subtask?
>
> 3. Under what conditions are these tasks usually performed?

As to the first variable, the literature of the group under investigation is the primary source, though it cannot be entirely dispositive. If, however (as is true in the case of handwriting),[149] that literature identifies subtasks, or subsets of practice conditions which are taken to create situations of easier and more difficult performance, these are entitled to be taken at their word. It may be that the outsider can supplement this system of subtask conditions with others not in the guild literature, but which may be reasonably likely, on other grounds, to affect results. But it would seem to border on the ludicrous to treat such an area as involving a global unitary skill which can be proven to exist by any test of any task anywhere within its bounds.[150] Unfortunately, this is exactly what most courts do, as we shall see.

As to the second variable, when an experience-based subjective judgment component is admitted to be present in normal practice (as it was in Carlson's case, and is in most "experience-based" claims of expertise, including handwriting identification), then defensibly designed tests must determine the reliability of that skill in

---

[148] This litany is not intended to suggest that each term represents something different to be dealt with in each case. It reflects terms which are almost synonymous, but which are in common usage and which capture slightly different aspects of the "experience-based expertise" phenomenon.

[149] *See* Risinger, *Task at Hand*, *supra* note 61, at 782 n.69.

[150] It is instructive to note the concurrence of Wigmore on the same point in slightly different terms:

> The capacity is *in every case a relative one, i.e.,* relative *to the topic about which the person is asked to make his statement.* The object is to be sure that the question to the witness will be answered by a person who is fitted to answer it. His fitness, then, is fitness on that point. He may be fitted to answer about countless other matters, but that does not justify accepting his views in the matter in hand.

II JOHN H. WIGMORE, TREATISE ON THE ANGLO-AMERICAN SYSTEM OF EVIDENCE IN TRIALS AT COMMON LAW § 555, at 634 (3d ed. 1940); *see also* Risinger, *Taxonomy*, *supra* note 37, at 510.

practitioners. By far the most desirable form of such a testing regime would consist of individual proficiency testing which would generate accuracy scores for each subtask for each practitioner, like the "personal equations" of nineteenth century astronomy.[151] The Australian document examiners and science-trained researchers Bryan Found and Doug Rogers, and their collaborators at the Forensic Expertise Profiling Laboratory at La Trobe University, have begun developing a testing regime like that for Australian document examiners, though the effort is still in its infancy.[152] Unfortunately, in this country, there is no such system of individual proficiency testing, at least none with known results.[153] The next best thing would be a system of group proficiency tests for each subtask. The weakness of such an approach is that, even if the group is successful, it ascribes the average performance for the group to both the group's strongest and (more troublingly) its weakest performers.[154] This is especially a problem when only what are believed to be the strongest performers are selected to take the group proficiency tests. Finally, while unacceptable performance might be determined from tests of the claimed experts alone, acceptable performance for the purposes of the law is dependent on a marginal advantage in accuracy over the jury, so tests must also be administered to appropriately selected lay groups to determine the existence of such an advantage.[155]

We have already written fairly extensively about the weaknesses of handwriting identification expertise.[156] When we started, we characterized the conclusions concerning such claims flowing from anything that could be called formal data as either mildly negative, or

---

[151] Personal equations were corrections worked out in the early nineteenth century for observational bias in astronomical observers which turned out to be fairly stable for each observer. For a description of the personal equation phenomenon, see EDWARD G. BORING, A HISTORY OF EXPERIMENTAL PSYCHOLOGY 134-35 (1929).

[152] *See* Bryan Found & Doug Rogers, *Revision and Corrective Action Package: Signature Trial 2001* (distributed on CD-ROM by the Forensic Expertise Profiling Laboratory, School of Human Biosciences, La Trobe University, Australia) (described fully in the 2003 supplement to D. Michael Risinger, *Handwriting Identification*, *in* 3 DAVID L. FAIGMAN, DAVID H. KAY, MICHAEL J. SAKS & JOSEPH SANDERS, MODERN SCIENTIFIC EVIDENCE 400-83 (2d ed. 2002), and in Jodi Sita, Bryan Found, & Douglas K. Rogers, *Forensic Handwriting Examiners' Expertise for Signature Comparison*, 47 J. FORENSIC SCI. 1117 (2002)).

[153] We say "with known results" because law enforcement laboratories may have internal proficiency tests, the existence and results of which they keep secret. This was apparently the case in regard to fingerprinting. *See infra* note 201.

[154] This is the problem with most of the extant American research on document examiner skill in handwriting identification, particularly that done by Moshe Kam and his associates. See the extensive discussion in Risinger, *supra* note 152.

[155] Risinger et al., *Exorcism*, *supra* note 38, at 731, 734-35.

[156] See writings previously cited at notes 37, 38, 53, 61, 137, 145, and 152.

nonexistent, depending on one's attitude about both the internal and external validity of the data generated by the proficiency tests sponsored by the Forensic Science Foundation.[157] Since then there has been some formal research undertaken.[158] However, when compared to the extent of research devoted to many other areas of endeavor, the amount of research devoted to handwriting identification expertise remains sparse. Nevertheless, even taking the extant research at face value, two things are clear. Most of it was undertaken under test designs which appear to have been specially tailored to make it impossible to make direct statements about individual performance.[159] In addition, most of it is directed, at best, to showing an average advantage of document examiners over lay persons in regard to what are, in the aggregate, the easiest possible subtasks. No attempt is made to determine the error rate for document examiners for the more difficult tasks, which are the tasks commonly at issue in actual criminal prosecutions. Even the studies which have been directed toward a definable subtask have all been directed toward the subtask regarded as one of the easiest subtasks in the claimed expertise's own literature—determining whether a signature was written by the person whose name is reflected, or by some other person inexperienced as a forger.[160] What have never been subjected to any tests are the two most common subtasks at issue in criminal prosecutions: attribution of authorship of block printing and attribution of the authorship of a forged signature to a person whose name is not reflected by the signature (relying only on the characteristics of the few letters in the signature, which often reflect some attempt at disguise.) And, as a follow-on to this issue, no empirical work at all has been done on the problem of determining how much "questioned" writing is necessary to perform any

---

[157] Risinger et al., *Exorcism*, *supra* note 38, at 750-51.

[158] The extant research is fully recounted and analyzed in Risinger, *supra* note 152 and supplements.

[159] This is especially true of the first three Kam studies: Moshe Kam, Joseph Wetstein & Robert Conn, *Proficiency of Professional Document Examiners in Writer Identification*, 39 J. FORENSIC SCI. 5 (1994) [hereinafter Kam I]; Moshe Kam, Gabriel Fielding & Robert Conn, *Writer Identification by Professional Document Examiners*, 42 J. FORENSIC SCI. 778 (1997) [hereinafter Kam II]; and Moshe Kam, Gabriel Fielding & Robert Conn, *The Effects of Monetary Incentives on Performance of Nonprofessionals in Document-Examination Proficiency Tests*, 43 J. FORENSIC SCI. 1000 (1998) [hereinafter Kam III]. Each of those studies has a complex design which insures that no two tests administered are exactly the same, so that group performances can be compared but comparisons of individual performances are undermined.

[160] That is true of both Moshe Kam, Kishore Gummadidala, Gabriel Fielding & Robert Conn, *Signature Authentication by Forensic Document Examiners*, 46 J. FORENSIC SCI. 884 (2001) [hereinafter Kam IV] and Sita et al., *supra* note 152.

identifications accurately—one letter? A capital "Q" but not a lower case "o"? One word? Thirty letters? There is simply no empirical evidence at all on this boundary problem, that is, on the variables that determine the threshold of reliability of the claimed skill, assuming it exists at all (when there is both plenty of questioned writing and plenty of known writing of the putative author).

As for the third variable, tests ought to attempt to simulate the conditions of actual practice as much as possible. Unfortunately, the conditions of actual practice are not always easy to determine, and the information that does exist suggests altogether less control and fewer masking protocols than would be the usual norm in the practice of science.[161] This makes it difficult to design tests which demonstrate that any skills revealed by the tests (which of necessity must be masked to a great extent) are robust enough for their accuracy to survive the confounding conditions of actual practice. In theory, tests of this question could be designed, but to date there are none. This leaves the results of such tests as there are with a giant problem of external validity on that ground. It would be nice if we could make this problem go away by introducing adequate masking or blind testing procedures into forensic practice,[162] but so far nothing of this sort has been done.

It would seem that, as a minimum, *Kumho Tire* would require any judge facing a Rule 702 reliability challenge to handwriting identification expertise to do what the Supreme Court did in *Kumho Tire* in regard to Carlson, that is, identify the particular sub-task which is at issue under the facts of the case and the attendant skill claims involved, and then to look to the empirical record to see what support there is for the claim that a document examiner can reliably perform that task, by the methods employed, given the conditions under which they were employed. So how have judges measured up to this ideal since the decision in *Kumho Tire*?

Since *Kumho Tire* there have been twenty-one available[163] federal court decisions on challenges to proffered handwriting identification expertise.[164] Eight are appellate decisions,[165] seven of which affirmed

---

[161] Risinger et al., *Observer Effects*, *supra* note 18, at 35-42.

[162] *See id.* at 45-52.

[163] We have counted as "available" all decisions in written form which have come into our hands, whether reflected on databases or counted as "reported" by local rules, or not.

[164] United States v. Crisp, 324 F.3d 261 (4th Cir. 2003); United States v. Mooney, 315 F.3d 54 (1st Cir. 2002); United States v. Kehoe, 310 F.3d 579 (8th Cir. 2002); United States v. Hernandez, No. 01-1194, 2002 WL 1335595 (10th Cir. June 19, 2002); United States v. Johnson, 30 Fed. Appx. 686 (9th Cir. 2002); United States v.

trial court global consideration and subsequent admission of such testimony, and one of which affirmed trial court admission of such testimony but restricted the expert from testifying to the conclusion of identity.[166] The district court cases generated no available opinions in those cases, and the appellate decisions all managed to find no abuse of discretion without describing the particular claim of expertise which was at stake in the case.

Of the remaining thirteen trial court cases, only two have come close to the ideal of identifying the "task at hand,"[167] and, analyzing reliability with reference to that task, they both excluded the proffered testimony completely. A third case excluded the proffered testimony without any particularized analysis,[168] and a fourth exclusion was based on the failure of the government to proffer witnesses at the *Daubert* hearing with sufficient familiarity with the empirical record to testify.[169] In the other nine district court cases, five admitted the proffered expertise only with significant limitations[170] (a testament, perhaps, to the weakness of the empirical record in regard to the reliability of handwriting identification expertise even considered globally), and the other four admitted the

---

Battle, No. 98-3246, 1999 WL 596966 (10th Cir. Aug. 6, 1999); United States v. Paul, 175 F.3d 906 (11th Cir. 1999); United States v. Hidalgo, No. CR-01-1011-PHX-FJM (D. Ariz. Nov. 6, 2002); United States v. Prime, 220 F. Supp. 2d 1203 (W.D. Wash. 2002); United State v. Lewis, Crim. Action No. 2:02-00042, 2002 WL 31055185 (S.D.W. Va. Sept. 11, 2002); United States v. Nadurath, No. 4:02-CR-32-A, 2002 WL 1000929 (N.D. Tex. May 14, 2002); United States v. Gricco, Crim. Action No. 01-90, 2002 WL 746037 (E.D. Pa. Apr. 26, 2002); United States v. Brewer, No. 01 CR 892, 2002 WL 596365 (N.D. Ill. Apr. 16, 2002); United States v. Richmond, Crim. Action No. 00-321 Section "N," 2001 WL 1117235 (E.D. La. Sept. 21, 2001); United States v. Saelee, 162 F. Supp. 2d 1097 (D. Alaska 2001); United States v. Fujii, 152 F. Supp. 2d 939 (N.D. Ill. 2000); United States v. Rutherford, 104 F. Supp. 2d 1190 (D. Neb. 2000); United States v. Santillan, No. CR-96-40169 DLJ, 1999 WL 1201765 (N.D. Cal. Dec. 3, 1999); United States v. Brown, No. CR 99-184 ABC (C.D. Cal. Dec. 1, 1999); United States v. Hines, 55 F. Supp. 2d 62 (D. Mass. 1999); United States v. Elmore, 56 M.J. 533 (N-M. Ct. Crim. App. 2001).

[165] Seven of the cases are from the courts of appeal and one is from the military court of appeal. *See Crisp*, 324 F.3d 261; *Mooney*, 315 F.3d 54; *Kehoe*, 310 F.3d 579; *Hernandez*, 2002 WL 1335595; *Johnson*, 30 Fed. Appx. 686; *Battle*, 1999 WL 596966; *Paul*, 175 F.3d 906; *Elmore*, 56 M.J. 533.

[166] *Hernandez*, 2002 WL 1335595. The court of appeals seemed puzzled at the restrictive approach, which was borrowed from *United States v. Hines*. *See Hernandez*, 2002 WL 133559, at *2.

[167] *Saelee*, 162 F. Supp. 2d 1097; *Fujii*, 152 F. Supp. 2d 939.

[168] *Brewer*, 2002 WL 596365.

[169] *Lewis*, 2002 WL 31055185. The government in this case pushed the limits of expert qualification too far even for a *Daubert* hearing.

[170] *Hidalgo*, No. CR-01-1011-PHX-FJM (D. Ariz. Nov. 6, 2002); *Rutherford*, 104 F. Supp. 2d 1190; *Santillan*, 1999 WL 1201765; *Brown*, No. CR 99-184 ABC (C.D. Cal. Dec. 1, 1999); *Hines*, 55 F. Supp. 2d 62.

testimony globally.[171]

At least four[172] (and perhaps many more, since the opinions often do not give sufficient detail to determine) of the nineteen opinions admitting the proffered testimony (with or without restrictions) involved the kind of subtask or boundary problems described above. But in none of those cases was the subtask problems identified, nor did they have an explicit effect on the outcome. Indeed, somewhat ironically, one of the four *exclusions* in this set of cases was a violation of *Kumho Tire's* requirements, invoking the precedent of two of the other exclusions without checking to see if their grounds of exclusion applied to the case before the court (they didn't).[173] So the bulk of district courts, whatever else they may be doing, are not performing very well under the requirements of *Kumho Tire*, and as a result, testimony on many subtasks of questionable reliability is being allowed in front of the jury.

This same pattern is not restricted to handwriting identification cases. An examination of all the reported opinions in criminal cases since the decision in *Kumho Tire* shows that there is only one area which has gotten substantially more attention since that decision than it did before, and that is fingerprint identification, that *ne plus ultra* of claimed perfection in the forensic identification disciplines. What comes of asking the wrong question can have no better illustration than that which comes from a consideration of the fingerprint challenges in general, and from the most famous of those cases, the case of Carlos Ivan Lllera Plaza, in particular.

B. The Fingerprint Cases

Judge Pollak's two conflicting opinions in *United States v. Llera Plaza*[174] are by now the stuff of legend. Much has been written on them already. However, both of those opinions and what has been written about them seem generally to have missed the point. Both opinions are stark violations of the approach mandated by *Kumho Tire*. One can read both opinions until the last trump is sounded,

---

[171] *Prime*, 220 F. Supp. 2d 1203; *Nadurath*, 2002 WL 1000929; *Gricco*, 2002 WL 746037; *Richmond*, 2001 WL 1117235.

[172] *Battle*, 1999 WL 596966; *Rutherford*, 104 F. Supp. 2d 1190; *Brown*, No. CR 99-184 ABC (C.D. Cal. Dec. 1, 1999); *Elmore*, 56 M.J. 533.

[173] *Brewer*, 2002 WL 596365. It might shock some of our critics in the forensic science community to realize it, but we consider exclusion in this manner as erroneous as admission.

[174] United States v. Llera Plaza, 179 F. Supp. 2d 492 (E.D. Pa. 2002) [hereinafter *Llera Plaza I*]; United States v. Llera Plaza, 188 F. Supp. 2d 549 (E.D. Pa. 2002) [hereinafter *Llera Plaza II*].

and never have an inkling about the "task at hand," as defined by the particular factual circumstances to which the claimed expertise was applied in Llera Plaza's actual case (the "target issue," in our terms). In order to understand why such formulation was critically important in this case (aside from the fact that it seems to have been mandated by the Supreme Court of the United States), we must give a quick review of fingerprint theory and practice and the issues attendant to it.

Putting aside the overblown nineteenth century language of absolute uniqueness in which the claims of fingerprint identification are usually expressed, the main claim may be reformulated in more acceptable modern terminology, thus: Human skin contains a fairly sizable extent of ridged skin on the palms and fingers of the hands and the soles and toes of the feet. The usual theoretical account for its function is that the ridges increase friction on the surfaces,[175] where increased friction for traction and gripping would be of survival benefit. Such work as has been done on the subject indicates that the pattern of ridges in any given individual is constant throughout life.[176] Setting aside the question of whether no two people share the "exact" same pattern of ridges and perceptible detail associated with them, it is clear that not every person has the same pattern of ridges as every other person across the entire extent of their ridged skin. Indeed, though surprisingly little defensible formal research has been done on the question, it seems apparent enough from anecdotal information that variation is so common that "exact matches" across the entire range of ridged skin are vanishingly rare (if they occur at all in the human population now alive). There is apparently some mechanism at work in the fetal development stage which triggers the generation of the ridges by a process that exhibits a fair amount of randomness at a fine level of organization and detail[177] (though the patterns fit general categories of pattern at a

---

[175] *Llera Plaza I*, 179 F. Supp. 2d. at 495-96 (quoting the testimony of Dr. William Babler, President of the American Dermatoglyphics Association, given in *United States v. Mitchell*, 199 F. Supp. 2d 820 (E.D. Pa. 2002)); *see* SIMON A. COLE, SUSPECT IDENTITIES: A HISTORY OF FINGERPRINTING AND CRIMINAL IDENTIFICATION 108 (2001).

[176] *Llera Plaza I*, 179 F. Supp. 2d. at 495-96 (reporting Babler's testimony); David A. Stoney, *The Scientific Basis of Expert Testimony on Fingerprint Identification*, *in* 3 DAVID L. FAIGMAN ET AL., MODERN SCIENTIFIC EVIDENCE § 27-2.2.1, at 383 (2d ed. 2002) [hereinafter Stoney, *Scientific Basis*].

[177] A start is being made at understanding how various processes, including biological processes, can generate self-organizing patterns, displaying such a combination of order and randomness by beginning with a set of relatively simple conditions and subjecting them to fairly simple algorithms which instantiate both positive and negative feedback. *See* Scott Camazine, *Patterns in Nature*, 112 NATURAL HISTORY 34, 40 (2003). *See generally* SCOTT CAMAZINE, SELF-ORGANIZATION IN

grosser level of examination).[178] However, because the ridges are generally curvilinear in complex ways, describing the amount of randomness and the likelihood of a random match is a daunting theoretical problem.[179] Even describing what constitutes a match is a problem, because, contrary to popular belief, matches are often not manifested in anything resembling perfect superimposability. The curved and deformable nature of surfaces upon which prints of ridged skin may be left, and the deformable nature of skin itself, often defeats exact superimposition, so that even with prints reflecting large extents of ridges, matching may be an exercise in complex topographical judgment in accounting for such (usually mild but perceptible) deformities preventing superimposition.[180]

So, while the formal research necessary to justify such a statement with formal data has not been done, and the empirical and theoretical work which would give a proper explanatory account of the mechanism behind the organization of ridged skin has not been done,[181] it seems uncontroversial in any serious way to say that the amount of randomness in ridge organization is such that "no two people" share the same pattern in a confusable way across the entire extent of their ridged skin. We do not, however, use the entire extent of ridged skin for identification purposes. Partly because of convenience in "rolling" such prints and partly because it is the print most likely to be left on a surface inadvertently, we use only the ridged skin on the balls of the fingers. The standard practice known to everyone who has ever been "fingerprinted" is to ink the balls of the fingers and roll them onto a card in boxes marked out for each digit. The result is a set of ten prints of known orientation comprising about one square-inch each, for a total of roughly ten square-inches of ridged skin.[182] Again, as in the case of the entire extent of ridged skin, the formal research necessary to establish random match probabilities for two sets of ten prints from different people has not been done, but it seems fair to conclude that such probability is sufficiently minuscule not to trouble the practical certainty which we seek in the law. When an unidentified body is

---

BIOLOGICAL SYSTEMS (2001).

[178] *See* COLE, *supra* note 175, at 114.

[179] *Id.* at 260; David A. Stoney, *Measurement of Fingerprint Individuality*, *in* ADVANCES IN FINGERPRINT TECHNOLOGY (Henry C. Lee & Robert E. Gaensslen eds., 2001); David A. Stoney & John I. Thornton, *A Critical Analysis of Quantitative Fingerprint Individuality Models*, 31 J. FORENSIC SCI. 1187 (1986).

[180] Stoney, *Scientific Basis*, *supra* note 176, § 27-2.2.5.

[181] Or is just beginning. *See supra* note 177.

[182] *Llera Plaza I*, 179 F. Supp. 2d at 496 (discussing testimony of David Ashbaugh).

found, and a ten print card is rolled from the fingers of the corpse, and it is found to match one on file with a law enforcement agency, doubts about the belief warrant for the identification would seem trivial.

At the other extreme, however, it is clear that there is a lower limit of certainty. If a glass found in a room where a murder had been committed had a smudge on it which showed clearly only one portion of one ridge one sixty-fourth of an inch long (a very "partial," "latent print"), neither its curvature nor any microscopic detail connected with it would allow a confident identification. Since no one knows its orientation, or which digit it came from, it would have to be compared with every short length of every ridge on every print of a candidate card, and no one knows exactly how many such short lengths of a ridge might match it in any randomly selected ten-print card.

So, in regard to the admissibility of fingerprint identification, there would seem to be two potential issues of reliability, one trivial and one extremely important. The first would challenge the admissibility of any identification derived from fingerprint comparison, on the ground that without formal research and quantified statistical modeling, its reliability could not be established. Such a challenge is puckish, quixotic, and in some ways constructive, but in others not.[183] In the end, it is doomed to failure, and not simply for the wrong reasons either. First, there has been a little empirical study that tends to indicate that, at least in regard to large clear areas of ridged skin, variability is large and coincidental matches are at least very rare.[184] Second, the extensive use of ten-print comparisons for identification of unknown persons followed by later confirmation of identity from other sources and no known record of error can be said to form a practical accuracy feedback loop unique

---

[183] Such positions can be used to paint all critics of forensic science as radical bomb-throwers and extremists deserving of small consideration. For a milder and more nuanced, but still (in our view) much too global version of a similar argument, see Edward J. Imwinkelried, *Flawed Expert Testimony: Striking the Right Balance in Admissibility Standards*, 18 CRIM. JUST. 28 (2003), which asserts that critics demand unreasonable global exclusion, and apparently argues for (to our minds) unreasonably global admission.

[184] The so-called "50k x 50k study" testified to by Donald Ziesig of Lockheed Martin Information Systems, which computer-compared each of 50,000 individual rolled loop class prints from white males with each other, was such a study. *See Llera Plaza I*, 179 F. Supp. 2d at 497. This study is referred to in S.B. Meagher, B. Budowle & D. Ziesig, *50K vs. 50K Fingerprint Comparison Test* (1999) (unpublished), in Stoney, *Scientific Basis*, *supra* note 176, § 27-2.1.2[6], at 381 n.12, and must be taken with something of a grain of salt, since it was FBI-commissioned and appears never to have been published.

among forensic identification techniques.[185] While more defensible research is to be encouraged, a global challenge to the reliability of all fingerprint identification is a non-starter.

The second potential challenge is the important one. It is based on the boundary problem described above: identification of a practical threshold of reliability for "partial prints." What standards should be applied to ensure that identifications from a small area of print found at a crime scene are sufficiently reliable for purposes of the criminal law? Here, the absence of formal data ought to be more troubling under *Daubert* and *Kumho Tire*. This is especially true because fingerprint experts either disagree on how to characterize the threshold of reliability, or more commonly, claim that such a threshold cannot be described at all.[186] This is the result of the addition to fingerprint examination over the last decades of new sources of information (often now collectively referred to under the title "ridgeology")[187] which makes old thresholds fail in some circumstances.[188]

Extensive clear prints such as ten-print cards might be quickly matchable by general pattern at the first general level of observation (sometimes called "the first level of analysis"),[189] and confirmed by

---

[185] Stoney, *Scientific Basis*, *supra* note 176, § 27-2.3.2.

[186] COLE, *supra* note 175, at 262-63; Stoney, *Scientific Basis*, *supra* note 176, § 27-2.3.1[2].

[187] "Ridgeology" as a term appears traceable to a 1983 pamphlet by David Ashbaugh, a member of the Royal Canadian Mounted Police, entitled "Ridgeology." The specific detail to which he referred, such as the presence of pores and characteristics of curvature, had been known and considered for some time (the use of pores even has its own term, "poroscopy"), but Ashbaugh's radical claim that identification was always a gestalt which could never be subject to any threshold system of points (which had been foreshadowed by a resolution of the International Association of Identification, the leading organization of fingerprint examiners) was embraced by many. *See* COLE, *supra* note 175, at 261-66.

[188] Or rather, which makes old thresholds overly conservative in the eyes of some. COLE, *supra* note 175, at 263.

[189] *Llera Plaza I*, 179 F. Supp. 2d at 496 (referring to "first level of detail"). This is all part of what is now billed as the "ACE-V" methodology, a "methodology" so lacking in methodological detail as to be, upon reflection, nearly hilarious. The A stands for "assess," that is, look at a latent print and decide whether it is too smudged or small even to try to analyze it, and whether any apparent detail ought to be ignored because it represents a "double tapped" or overlapping print. The C stands for "comparison," and that means, well, the examiner is to compare the known and the latent print, though there are apparently no fixed standards for performing such a comparison. Rather, it is based "on the training and experience of the examiner." The E stands for "evaluation," which means that the examiner decides if the two are similar enough to declare that they are a match, without reference to any particular notion of minimum points of correspondence, and V stands for "validation," which is a non-blind checking of the first examiner's work by a second examiner. This is the "scientific technique" which the government in *Llera Plaza* argued "met all four of

correspondence of individual landmarks (often called minutiae)[190] at the next level of magnification (often called the "second level of analysis").[191] These landmarks were the Galton[192] solution to the curvilinear nature of ridges, identifications of characteristics which could serve as discrete units of analysis, such as the point where one ridge divides into two (often called a "bifurcation"), or the division of two ridges followed by their closure again (a "lake"), etc.[193] The correspondence of such landmarks, the number of ridges separating them, and the relative direction and distance of their separation, are traditionally the stuff of determining the "number of points of comparison" between two prints. However, at yet higher magnification (referred to sometimes as "third level analysis"),[194] a clear print will show yet more supplementary information detail, including the width and shoreline of individual stretches of ridge, and the presence of pores separated by variable distances. Herein lies the rub. Traditional reliability thresholds typically required from seven to sixteen landmark points of comparison, with no unexplained differences.[195] Adding the third level of magnification means, according to most examiners, that fewer traditional points are necessary in a clear print because the new detail can make up for fewer landmarks in individual cases.[196] Why the new details of ridgeology cannot simply be assimilated into the "points of comparison" available to make up a conservative quantified minimum is not completely clear.[197] Given the subjective nature of

---

the *Daubert* guidelines." *Llera Plaza II*, 188 F. Supp. 2d at 560. Judge Pollak found ACE-V not to be "scientific," but appears to have taken it seriously as a "methodology." *Id.* at 561-69.

[190] Traditionally (per Sir Francis Galton) the term "minutiae" (singular, "minutia") was synonymous with "Galton points." *See* Stoney, *Scientific Basis*, *supra* note 176, § 27-2.1.2[5]; *see also* COLE, *supra* note 175, at 79-80. There may be a trend toward applying the term to the even smaller "third level" detail. *See Llera Plaza I*, 179 F. Supp. 2d at 500 (attributing similar terminology to FBI Fingerprint Unit Chief Stephen Meagher).

[191] *Llera Plaza I*, 179 F. Supp. 2d at 496 (referring to "level two detail").

[192] British biostatistician, geneticist, eugenicist and fingerprint pioneer Sir Francis Galton (1822-1911). *See* COLE, *supra* note 175, at 79-80.

[193] *Id.*

[194] *Id.*

[195] Stoney, *Scientific Basis*, *supra* note 176, § 27-2.1.2[5].

[196] *Id.*

[197] The argument seems to have two aspects. First, some landmarks, such as a "trifurcation," are so rare that their presence even without much else might be enough for confident identification. Second, the process is claimed to be a subjective gestalt process which is not rationally subject to universal thresholds made up of specified criteria. *See id.* While the latter may actually describe what examiners do, it would seem desirable to hold them to some sort of statable minimum even at

the evaluation at the boundary, the necessity, for the purposes of the law, of a mandated threshold in some form would seem most consistent with the policies of *Daubert*, *Kumho*, Rule 702, and the standard of proof beyond a reasonable doubt in the criminal law. This is especially true because such evaluations at the boundary are usually performed without any masking protocols to prevent suggestion or expectation from affecting the results, and without any line-up type foils which, in this area, could be easily supplied.[198] However, again consistent with *Kumho Tire*, such determinations of the appropriate reliability threshold should be dealt with in cases which arguably present specific examples of the boundary problem. And here is where Judge Pollak crashed. As we said previously, nowhere in either opinion does he tell us anything about the nature and extent of the latent fingerprints under examination in Llera Plaza's case. In the first opinion (*Llera Plaza I*) he prepares a hash comprised of ruminations on global reliability, the threshold problem and the lack of formal research, concluding that because of the lack of formal data, fingerprint identification globally can never support testimony by an examiner concerning actual identification, presumably even in a ten-print comparison.[199] Rather than excluding them from the courtroom completely, however, he applies the universal solvent du jour and declares that they may function as "Hines" witnesses, pointing out similarities but rendering no conclusion.[200] How a jury would be qualified by experience to evaluate such testimony as to fingerprint correspondences was left unclear.

For whatever reasons,[201] Judge Pollak reversed himself two

---

the cost of giving up the occasional accurate identification in court. (Such information could of course still be used as an investigative lead.)

[198] *See* Risinger et al., *Observer Effects*, *supra* note 18, at 43.

[199] This was based on a determination that the "AC" part of the "ACE-V" "methodology" was objective, but the "E" (for evaluation) part was too subjective to (ever) be reliable. *Llera Plaza I*, 179 F. Supp. 2d at 516.

[200] So called based on the similar decision of Judge Gertner in regard to handwriting identification testimony in *United States v. Hines*, 55 F. Supp. 2d 62 (D. Mass 1999), upon which Judge Pollak explicitly relied. *Llera Plaza I*, 179 F. Supp. 2d at 517.

[201] One could fill a book with speculations about what led to the *volte face*. The government got its toe-hold on a reconsideration by producing the results of FBI fingerprint examiner proficiency tests, tests which they had theretofore kept secret, presumably because the results (in contrast to their usual claims) were not perfect. On the government's tendency to keep empirical results secret if they do not like them, see D. Michael Risinger & Michael J. Saks, *Rationality, Research and Leviathan: Law Enforcement Sponsored Research and the Criminal Process*,___ MICH. ST. DCL L. REV. ___ (forthcoming 2003). There was also reason to believe that the proficiency tests were so easy as to have little to do with the boundary problem. At any rate, although

months later in *Llera-Plaza II*. In that opinion, he says that he had gone too far in declaring that fingerprint identification witnesses should be treated merely as "Hines" witnesses, and so he reverses himself, declaring that in general fingerprint identification witnesses should be allowed to testify as they always have, while he encourages research to continue to provide a theoretical basis for such testimony. However, once again, he never describes the factual contours of the "task at hand" in the case, and never addresses in any organized way the boundary issue regarding a reliability threshold for fingerprint identification. Toward the end of the process this apparently begins to dawn, for the last line of *Llera Plaza II* is: "At the upcoming trial, the presentation of expert fingerprint testimony by the government . . . will be subject to the court's oversight prior to presentation of such testimony before the jury, with a view to insuring that . . . fingerprints offered in evidence will be of a quality arguably susceptible of responsible analysis, comparison and evaluation."[202] Why that was not the focus of the entire *Daubert/Kumho Tire* inquiry from the beginning, we will probably never know.

If *Llera Plaza II* answered the trivial question and avoided the hard one by violating the strictures laid down by the Supreme Court in *Kumho Tire*, it has certainly been treated as disposing of *all* questions by subsequent courts facing fingerprint identification reliability challenges. This dénouement was unfortunately predictable, given how much courts want to avoid such issues and seek even inapplicable precedents to use in this way. The *Llera Plaza* debacle will probably delay appropriate judicial examination of the boundary problem of threshold reliability in regard to fingerprint identification applied to partial latent prints for a long time to come. Such are sometimes the costs of the judge's failure to frame the question before the court with proper specificity.

To give Judge Pollak his due, he at least attempted to take on a hard issue, and it is only fair to note that Judge Pollak is hardly alone in failing to frame the fingerprint reliability issue in the case before him with the specificity required by *Kumho Tire*. Since the year 2000 there have been a spate of challenges to the reliability of fingerprint identification raised in the federal courts, generating twenty-one opinions,[203] and in *not a single case* has the court described the

---

Judge Pollak was critical of them, they were good enough in the end to be a part of the basis for Judge Pollak's self-reversal. *Id.* at 565-66.

[202] *Llera Plaza II*, 188 F. Supp. 2d at 576.

[203] United States v. Crisp, 324 F.3d 261 (4th Cir. 2003); United States v. Navarro-Fletes, 49 Fed. Appx. 732 (9th Cir. 2002); United States v. Hernandez, 299 F.3d 984 (8th Cir. 2002); United States v. Ambriz-Vasquez, 34 Fed. Appx. 356 (9th Cir. 2002);

individual characteristics of the latent prints which were the subject of the challenge. It is enough to make one feel sorry for the Supreme Court, so little has *Kumho Tire* been read with care by judges.

CONCLUSION

When it comes to prosecution-proffered expertise, the approach taken by courts in handwriting and fingerprint reliability tasks is not atypical. A perusal of all the reported opinions in criminal cases since the decision in *Kumho Tire* reveals a predominance of inappropriately global examination, especially in regard to experience-based claims of expertise. The Supreme Court has only itself to blame for this state of affairs. First, while it is clear in context that references to "discretion" and "flexibility" were meant only to allow the intelligent selection of the most rationally appropriate criteria of reliability for a particular kind of expertise and its claims in relation to the particular facts of the case, they have been seized upon by the lower courts as a warrant to avoid hard tasks of framing and evaluation, at least in regard to prosecution proffers.[204] No matter

---

United States v. Turner, 285 F.3d 909 (3d Cir. 2002); United States v. Martinez-Garduno, 31 Fed. Appx. 475 (9th Cir. 2002); United States v. Williams, 29 Fed. Appx. 486 (9th Cir. 2002); United States v. Rogers, 26 Fed. Appx. 171 (4th Cir. 2001); United States v. Havvard, 260 F.3d 597 (7th Cir. 2001); United States v. Merritt, Cause No. IP01-0081-CR-01-T/F, 2002 WL 1821821 (S.D. Ind. June 26, 2002); United States v. Nadurath, No. 4:02-CR-32-A, 2002 WL 1000929 (N.D. Tex. May 14, 2002); United States v. Mitchell, 199 F. Supp. 2d 262 (E.D. Pa. 2002); United States v. Cruz-Rivera, Crim. No. 00-98-01 (CCC), 2002 WL 662128 (D.P.R. Mar. 27, 2002); United States v. Salim, 189 F. Supp. 2d 93 (D. Kan. 2002); United States v. Cline, 188 F. Supp. 2d 1287 (D. Kan. 2002); United States v. Reaux, Crim. Action No. 01-071 Section "R" (2), 2001 WL 883221 (E.D. La. July 31, 2001); United States v. Joseph, Crim. Action No. 99-238 Section "N," 2001 WL 515213 (E.D. La. May 14, 2001); United States v. Martinez-Cintron, 136 F. Supp. 2d 17 (D.P.R. 2001); United States v. Havvard, 117 F. Supp. 2d 848 (S.D. Ind. 2000); Unites States v. Cooper, 91 F. Supp. 2d 79 (D.D.C. 2000). These opinions are in addition to the two *Llera Plaza* decisions.

[204] There is also reason to believe the that district courts have held criminal defendants' proffers to a higher standard of threshold reliability than prosecution proffers. *See generally*, Risinger, *Navigating Expert Reliability*, *supra* note 3. It is in some ways easy to account for this. Absent unusual circumstances resulting in an interlocutory appeal, exclusion of prosecution proffers by the trial court is unreviewable, whereas exclusion of defense proffers, in the mind of the trial judge, can be corrected on appeal if erroneous. It is easy to see how this can incline the trial judge toward admission. *See* R. Erik Lilquist, *A Comment on the Admissibility of Forensic Evidence*, 33 SETON HALL L. REV. 1189, 1191-92 (2003). However, the Supreme Court's decision in *General Electric Co. v. Joiner*, 522 U.S. 136 (1997), mandating review of Rule 702 rulings only by an "abuse of discretion" standard, has resulted in virtual automatic affirmance by the courts of appeal, and created a situation where nobody takes responsibility for seriously evaluating the actual reliability of prosecution-proffered expertise.

how clear this may appear to the scholarly observer, however, it is not likely to change until the Supreme Court once again returns to this area to inform the lower courts that it actually meant what it said.