The Science of Gatekeeping: The Federal Judicial Center's New Reference Manual on Scientific Evidence

John M. Conley

David W. Peterson

ESSAY

THE SCIENCE OF GATEKEEPING: THE FEDERAL JUDICIAL CENTER'S NEW *REFERENCE MANUAL ON SCIENTIFIC EVIDENCE*

JOHN M. CONLEY*
DAVID W. PETERSON**

With its 1993 decision in Daubert v. Merrell Dow Pharmaceuticals, Inc., *the Supreme Court abolished the seventy-year-old* Frye *test for evaluating scientific evidence and, instead, required that courts conduct an independent inquiry into the reliability and relevance of proffered scientific testimony. To assist judges in negotiating the post-*Daubert *legal landscape, the Federal Judicial Center has recently published a* Reference Manual on Scientific Evidence. *In this essay, Conley and Peterson evaluate the* Manual *with a critical eye toward whether it is likely to lead judges to the kinds of understanding demanded by the new evidentiary regime. They praise several of the* Manual*'s reference guides for their engaging question-and-answer formats and their comprehensive overviews of individual scientific disciplines. However, they fault the* Manual *for its failure to emphasize connections between topics and the absence of an interdisciplinary introduction to the scientific method. Therefore, while the* Manual *will serve as an authoritative reference on particular topics, Conley and Peterson worry that judges may, after finding a section of the* Manual *useful in a particular case, inappropriately extend their limited knowledge of scientific principles in subsequent cases.*

---

* A.B., Harvard; J.D., Ph.D., Duke. William Rand Kenan, Jr. Professor of Law, University of North Carolina at Chapel Hill. The authors thank Sheryl Gerety for her research assistance. Special thanks are due to UNC colleague Walker Blakey for suggesting the topic of this essay.

** B.S., Wisconsin; M.S., Ph.D., Stanford. Dr. Peterson recently retired as a statistics professor from Duke University, where he taught for more than twenty years. He is currently president of PRI Associates, Inc., a statistical consulting and software firm based in Durham, N.C.

INTRODUCTION

The federal judiciary's love-hate relationship with scientific evidence dates back at least to 1908. In that year, in *Muller v. Oregon*,[1] the Supreme Court received the eponymous "Brandeis brief." Louis Brandeis, defending an Oregon statute that limited the working hours of women, submitted "a very copious collection" of authorities purporting to show the particular vulnerability of women in the turn-of-the-century workplace.[2] The Court upheld the law, commenting on Brandeis' submission with sibylline brevity: "It may not be amiss, in the present case, before examining the constitutional question, to notice the course of legislation as well as expressions of opinion from other than judicial sources."[3] Drawing on Brandeis' brief, the Court found "[t]hat woman's physical structure and the performance of maternal functions place her at a disadvantage in the struggle for subsistence."[4]

Since *Muller*, the courts' attitudes toward science in the courtroom have run the gamut from uncritical enthusiasm to dismissive Luddism, with stops at all intermediate points. Just nineteen years after *Muller*, Oliver Wendell Holmes, Jr. led the Supreme Court into the depths of evangelical credulity with his infamous opinion in *Buck v. Bell* upholding the compulsory sterilization of "imbeciles" on the basis of turn-of-the-century theories of intelligence testing.[5] A generation later, in footnote eleven of its opinion in *Brown v. Board of Education*, the Court turned to science to support its finding of constitutional fact that separate education is inherently unequal.[6] In the 1970s, as employment discrimination litigation proliferated, the Court plunged into detailed questions of scientific method, endorsing particular tests of the statistical significance of racial disparities in hiring.[7] Over the last ten years,

---

1. 208 U.S. 412 (1908). For a further discussion of the history of scientific evidence, see DAVID W. BARNES & JOHN M. CONLEY, STATISTICAL EVIDENCE IN LITIGATION: METHODOLOGY, PROCEDURE, AND PRACTICE 3-10 (1986).

2. *Muller*, 208 U.S. at 419.

3. *Id.*

4. *Id.* at 421.

5. 274 U.S. 200, 207 (1927). For a review of *Buck* and its history, see John M. Conley, *"The First Principle of Real Reform": The Role of Science in Constitutional Jurisprudence*, 65 N.C. L. REV. 935 (1987).

6. 347 U.S. 483, 494 n.11 (1954).

7. Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 308 n.14 (1977) (endorsing the proposition that employer's denial of discrimination will be "suspect" where representation of protected group in employer's workforce is more than two or three

particularly in its death penalty jurisprudence, the Court has displayed a more skeptical attitude toward scientific evidence. Sometimes this skepticism has expressed itself in the form of detailed critiques of particular research;[8] at other times, Court majorities have questioned whether broad-based scientific studies can ever be probative of individual constitutional violations.[9] Categorizing trends in the lower courts' reception of scientific evidence would be a book-length undertaking. For background purposes, suffice it to say that one can find case support for almost any side of nearly every scientific question that has ever come before the courts.[10]

Now, eighty-eight years after *Muller*, the Federal Judicial Center has intervened in an effort to bring some order to the chaotic world of forensic science. The stated purpose of the Center's newly published *Reference Manual on Scientific Evidence* is "to assist judges in managing expert evidence, primarily in cases involving issues of science or technology."[11] Such cases, the *Manual* observes, "challenge the ability of judges and juries to comprehend the issues—and the evidence—and to deal with them in informed and effective ways. As a result, they tend to complicate the litigation, increase expense and delay, and jeopardize the quality of judicial and jury decision making."[12]

The *Manual*'s assistance comes not a moment too soon. In its 1993 decision in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[13] the Supreme Court abolished the venerable *Frye* test for evaluating scientific evidence. Under *Frye*, trial courts had to determine simply whether the expert's methods were generally accepted in the relevant scientific community.[14] Under *Daubert*, however, courts must

---

standard deviations below what would be expected on basis of population data); Castenada v. Partida, 430 U.S. 482, 496 n.17 (1976) (same).

8. *See, e.g.*, Lockhart v. McCree, 476 U.S. 162, 170-73 (1976) (criticizing studies intended to show bias of guilt-phase juries in capital cases).

9. *See, e.g.*, McCleskey v. Kemp, 481 U.S. 279, 292-95 (1987) (questioning whether social science data can ever prove sentencing to be racially discriminatory in any particular case).

10. For comprehensive treatments of the history and current state of scientific evidence, see Symposium, *Scientific Evidence After the Death of* Frye*:* Daubert *and Its Implications for Toxic Tort, Pharmaceutical and Product Liability Cases,* 15 CARDOZO L. REV. 1745 (1994); *Developments in the Law—Confronting the New Challenges of Scientific Evidence*, 108 HARV. L. REV. 1481 (1995) [hereinafter *Developments*].

11. FEDERAL JUDICIAL CENTER, REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 1 (1994) [hereinafter MANUAL].

12. *Id.*

13. 113 S. Ct. 2786 (1993).

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

conduct an independent inquiry into the reliability and relevance of the methods the expert has used.[15] While there is scholarly debate over whether *Daubert* intended any material change in the day-to-day practices of trial judges,[16] there is a widely shared judicial perception that *Daubert* has made the trial courts' "gatekeeping" burden far more onerous that it used to be.[17] A common reaction is that the Supreme Court is demanding that trial judges become "amateur scientists."[18]

Although the Federal Judicial Center made the decision to create a science manual well before *Daubert*,[19] the case brought both urgency and focus to the project. Whatever its ultimate implications, *Daubert* plainly signified that federal judges had to become intelligent critics of scientific evidence—right now. *Daubert* also made clear that the most important criterion of admissibility would be whether the work of an expert followed the canons of the scientific method.[20] Judges must therefore be able to recognize valid scientific methodology in an unlimited range of contexts and to distinguish it from artifice disguised by the trappings of science.

The purpose of this essay is to assess the likelihood that the *Manual* will actually help trial judges to negotiate the post-*Daubert* legal landscape. We begin with an overview of the *Manual* and some pertinent pieces of its history. We then review the *Daubert* case, with particular attention to the understanding of "science" that it embodies and the practical burdens that it imposes on trial courts. Against this background, we next consider the intellectual skills that post-*Daubert* trial judges need to develop, as well as the fallacies they need to avoid. Finally, examining in detail four of the *Manual*'s substantive scientific sections, we ask whether the *Manual* is likely to lead judges to the kind of understandings that the new evidentiary regime demands.

---

15. *Daubert*, 113 S. Ct. at 2795.

16. *See infra* notes 94-99 and accompanying text.

17. *See, e.g.*, Daubert v. Merrell Dow Pharmaceuticals, Inc., 43 F.3d 1311, 1316 (9th Cir. 1995) (on remand from Supreme Court); *Developments, supra* note 10, at 1509-10; Amy T. Schütz, Note, *The New Gatekeepers: Judging Scientific Evidence in a Post-*Frye *World*, 72 N.C. L. REV. 1060 (1994).

18. *Daubert*, 113 S. Ct. at 2800 (Rehnquist, C.J., concurring in part and dissenting in part).

19. *See* MANUAL, *supra* note 11, at vii (discussing history of project).

20. *Daubert*, 113 S. Ct. at 2795.

I. OVERVIEW OF THE *MANUAL*

The *Manual* consists of three principal sections: a two-chapter Overview;[21] seven reference guides,[22] each dealing with a branch of forensic science that plays a prominent role in contemporary litigation; and a section on Extraordinary Procedures,[23] with one chapter on court-appointed experts and one on special masters. Although the reference guides, all written by prominent forensic scientists, comprise nearly two-thirds of the *Manual* and represent its intellectual core, it is the Overview that has generated the most controversy so far.[24] The Overview includes a "Management of Expert Evidence" chapter by Senior District Judge William W. Schwarzer, retiring Director of the Center,[25] and an "Evidentiary Framework" chapter by Professor Margaret A. Berger.[26] Schwarzer's piece is succinct and hopeful. He reviews the procedural and evidentiary rules that a judge may invoke to identify and narrow the scientific issues in a case, and suggests ways in which the reference guides can be used in this framing process.[27] While his review of the rules is accurate and helpful, his comments about the utility of the reference guides seem to rest on the dubious assumption that the average trial judge will be able to wade into the coverage of a particular technical topic and emerge with a workable understanding of the critical operating principles.[28]

Berger's review of the law pertaining to scientific evidence—particularly *Daubert*—provoked a vigorous prepublication attack by the organized plaintiffs' bar. During the *Manual*'s preparation, the Center sent copies of each chapter to reviewers with expertise in the relevant field.[29] Barry Nace, past president of the Association of Trial Lawyers of America, the country's most prominent plaintiffs' lawyers' group, attacked Berger's chapter in a fif-

---

21. MANUAL, *supra* note 11, at 7-117.
22. *Id.* at 119-523.
23. *Id.* at 525-622.
24. *See infra* notes 29-39 and accompanying text.
25. William W. Schwarzer, *Management of Expert Evidence*, *in* MANUAL, *supra* note 11, at 7-35.
26. Margaret A. Berger, *Evidentiary Framework*, *in* MANUAL, *supra* note 11, at 37-117.
27. Schwarzer, *supra* note 25, at 17-19.
28. For example, his discussion of how the *Manual* might be used to simplify a dispute over DNA evidence, *id.* at 18, seems to assume a vast amount of technical competence on the judge's part.
29. MANUAL, *supra* note 11, at 623.

ty-seven-page letter to Chief Justice Rehnquist.[30] Nace called the *Manual* " 'poorly conceived, researched and organized—and so demonstrably wrong in so many places—as to be beyond repair.' "[31] He accused Berger of taking " 'a very unfair, twisted and defense-oriented view of *Daubert*.' "[32] Nace and other critics argued that Berger had materially overstated the discretion that *Daubert* gives trial judges to exclude scientific evidence.[33] *Daubert* clearly instructs trial judges to assess the reliability of an expert's *methods* before admitting scientific evidence.[34] But, the critics argued, Berger treated *Daubert* as inviting judges to exclude evidence on the independent ground that the expert's *conclusions* were contrary to generally held scientific positions and thus unreliable.[35]

As this Essay will demonstrate, *Daubert* leaves unanswered a number of questions, including the relationship between Federal Rule of Evidence 702, which focuses on the reliability of the expert's methods, and Rule 703, which deals with the acceptability of the expert's sources.[36] There is also a legitimate question whether Federal Rule of Evidence 403, which invites courts to balance the probative value of any evidence against the potential for prejudice, gives trial judges what is effectively a blank check in the scientific evidence context.[37] In the published version of her chapter, Berger does nothing more than note these questions and fairly summarize the arguments on all sides.[38] Nonetheless, leaders of the plaintiffs' bar have persisted in their criticism. Arthur Bryant, Executive Director of Trial Lawyers for Public Justice, for example, accused the Center of " 'completely ignor[ing]' " earlier criticisms and characterized Berger's chapter as " 'a very defense-oriented and improper view of *Daubert*.' "[39] Such comments bear little relation to what Berger actually says. Indeed, so intemperate and disproportionate are these

30. Mark Curriden, *Plaintiffs' Lawyers Rap Evidence Manual*, A.B.A. J. 20, 21 (Mar. 1995). Nace represented the *Daubert* plaintiffs in the trial court.

31. *Id.* (quoting unpublished Nace letters).

32. *Id.* at 20.

33. *Id.* at 21-22 (quoting criticisms by Nace and Trial Lawyers for Public Justice Executive Director Author Bryant, as well as response of Federal Judicial Center staff coordinator Joe S. Cecil).

34. *Daubert*, 113 S. Ct. at 2795-97.

35. Curriden, *supra* note 30, at 22.

36. *See infra* notes 109-19 and accompanying text.

37. *See infra* note 119.

38. *See* Berger, *supra* note 26, at 103-17.

39. Curriden, *supra* note 30, at 22 (quoting Bryant); *see also* James L. Dam, *Scientific Evidence Explained*, LAW. WKLY. USA, Dec. 19, 1994, at 1, 14 (reviewing same debate).

criticisms that we suspect that the very idea of motivating trial judges to look more critically at scientific evidence is what offends Nace, Bryant, and their constituents.

The seven reference guides deal with the topics of epidemiology, toxicology, survey research, DNA evidence, general statistics, multiple regression analysis, and the estimation of economic losses, in that order. The choice of topics is reasonable. Although some have commented on the absence of computer science,[40] the rationale for excluding this topic seems to have been that it arises most commonly in intellectual property cases, which are usually dealt with by the patent specialists of the United States Court of Appeals for the Federal Circuit.[41] While one could readily counter with a litany of copyright and trade secret cases in which "secular" judges have staggered through complex computer science issues,[42] no single volume can deal with everything, and no one can contend that the subjects selected are not both important and difficult.

The order of presentation is hard to understand, however. The chapter on basic statistical theory by David H. Kaye and David A. Freedman[43] covers statistical concepts such as sampling and significance testing that underlie each of the reference guides. The statistics chapter is thus introductory to all of the others, and it should be read first by any user of the *Manual* who is not already conversant with statistical theory. For this reason alone, the Kaye and Freedman chapter, not the epidemiology chapter, should have come first.

Putting the general statistics chapter first would also have helped to solve a related problem: insufficient emphasis on the intellectual continuity among the various reference guides. Subject to the caveat expressed in the preceding paragraph, each of the guides is self-sufficient and can, as a practical matter, be read in isolation. On one level, this is a virtue, since one could hardly expect judges to read the entire 400-plus pages the first time that they deal with any single topic. However, this self-sufficiency may mask the deeper truth that the guides consist in large part of variations on a relatively limited number of major themes. Having grasped, for example, the major

---

40. *See* Junda Woo, *New Guide for Judges Tries to Clarify Scientific Issues*, WALL ST. J., Dec. 14, 1994, at B1.

41. *See id.*

42. For sheer complexity, it may be hard to top the numerous opinions in the protracted Lotus-Borland software copyright litigation, culminating in Lotus Dev. Corp. v. Borland Int'l, Inc., 49 F.3d 807 (1st Cir.) *cert. granted*, 116 S. Ct. 39 (1995).

43. David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, *in* MANUAL, *supra* note 11, at 331-414.

points of the basic statistics chapter, the reader is already equipped to deal with most of the issues in the epidemiology and survey chapters, as well as some of the most important questions that arise in relation to toxicology, DNA profiling, and the estimation of economic loss. To be fair, the connections among the chapters are noted from time to time; there are, for instance, numerous cross-references.[44] What is lacking is an emphatic and repeated statement that science is not a conglomeration of discrete "studies" but rather a coherent approach to analyzing the world. As discussed in Part III of this essay, *Daubert* all but requires this understanding of science; as argued in Parts IV and V, this is also the approach most likely to enhance the quality of the trial courts' day-to-day processing of scientific evidence.

The *Manual* concludes with a chapter on "Court-Appointed Experts" by Joe S. Cecil and Thomas E. Willging, both members of the Center's research staff,[45] and a chapter on "Special Masters" by Professor Margaret G. Farrell.[46] The court-appointed expert chapter contains a comprehensive review of the important legal and practical considerations and the results of a survey of the entire federal district court bench concerning the frequency and desirability of appointing experts. Two aspects of the survey results stand out. First, judges rarely appoint their own experts (eighty percent of the survey respondents never had appointed an expert),[47] but they find them very helpful when they do.[48] Second, the trial judges almost never appoint an expert who functions as a confidential advisor to the court rather than as a conventional expert witness.[49] If *Daubert* indeed requires trial judges to make more informed judgments about scientific evidence, then court-appointed expert witnesses and advisors are likely to become more prominent players in the judicial process.

Farrell's chapter on special masters is a useful mix of legal doctrine and practical commentary. Farrell is truly comprehensive,

---

44. *See, e.g., id.* at 367 n.108 (noting cross-reference from statistics to regression chapter); *id.* at 479 nn.3-4 (providing cross-references from economic loss chapter to, respectively, multiple regression and survey chapters). However, despite the fundamental status of the statistics guide, it is disappointing that its 187 citation-laden footnotes contain only a handful of cross-references to the other guides.

45. Joe S. Cecil & Thomas E. Willging, *Court-Appointed Experts*, *in* MANUAL, *supra* note 11, at 525-73.

46. Margaret G. Farrell, *Special Masters*, *in* MANUAL, *supra* note 11, at 575-622.

47. Cecil & Willging, *supra* note 45, at 535.

48. *Id.* at 537.

49. *Id.* at 534. Cecil and Willging contend that while the authority to appoint such an advisor does not derive explicitly from Rule 706, a court's inherent power to do so is "virtually undisputed." *Id.*

dealing with everything from the historical roots of Federal Rule of Civil Procedure 53[50] to the details of how much and by whom a master should be paid.[51] Her overriding message is, appropriately enough, that special masters should be reserved only for the most extraordinary circumstances, and that the mere presence of complex scientific issues is not enough to justify this exotic departure from ordinary practice.[52]

II. THE *DAUBERT* DECISION

A. The Holding

The essential holding of *Daubert* is that Federal Rule of Evidence 702, which establishes the standard for the admissibility of scientific evidence, supersedes rather than incorporates the *Frye* test.[53] The *Frye* test, propounded in 1923, required that an expert's methods be generally accepted in the relevant scientific community: "[T]he thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs."[54] Rule 702, by contrast, permits "a witness qualified as an expert by knowledge, skill, experience, training or education" to offer an opinion "if scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue."[55] *Frye* was followed almost universally by the federal courts prior to the adoption of the Federal Rules of Evidence in 1975.[56] Since that time its continuing applicability has been a subject of debate among both courts[57] and commentators.[58]

The factual issue in *Daubert* was whether the anti-nausea drug Bendectin causes birth defects, as the plaintiffs claimed.[59] In the district court, the defendant moved for summary judgment on the basis of expert testimony "that no epidemiological study ever

50. Farrell, *supra* note 46, at 579.
51. *Id.* at 611-14.
52. *Id.* at 621-22.
53. Daubert v. Merrell Dow Pharmaceuticals, Inc., 113 S. Ct. 2786, 2793 (1993).
54. *Id.* (quoting Frye v. United States, 293 F. 1013, 1014 (D.C. Cir. 1923)).
55. FED. R. EVID. 702.
56. *See* David L. Faigman et al., *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, 1808 (1994).
57. *See Daubert*, 113 S. Ct. at 2792-93, 2793 n.3.
58. *See id.* at 2793 n.4 (listing many of the participants in the debate and noting the coinage of the term "*Frye*-ologist" to describe them).
59. *Id.* at 2791.

performed has concluded that the use of Bendectin by pregnant women has a statistically significant association to birth defects in these women's children."[60] The plaintiffs countered with affidavits from eight experts with "impressive credentials" who concluded "that Bendectin *can* cause birth defects."[61] Their conclusion was based on

> "in vitro" (test tube) and "in vivo" (live) animal studies that found a link between Bendectin and malformations; pharmacological studies of the chemical structure of Bendectin that purported to show similarities between the structure of the drug and that of other substances known to cause birth defects; and the "reanalysis" of previously published epidemiological (human statistical) studies.[62]

The district court granted the defendant's motion for summary judgment, holding that "expert opinion which is not based on epidemiological evidence is not admissible to establish causation" because it cannot be characterized as meeting the *Frye* standard of general acceptance in the relevant scientific community.[63] The plaintiffs' experts' recalculations of data in previously published studies that had found no causal link between the drug and birth defects did not meet the general acceptance standard, in part because they had not been published or subjected to peer review.[64] The Ninth Circuit affirmed.[65]

In considering whether Rule 702 had superseded the *Frye* test, the Supreme Court looked first to the text of Rule 702 and its drafting history. It found the absence of any reference to *Frye* or a "general acceptance" standard to be persuasive.[66] Moreover, the Court emphasized, the *Frye* test was at odds with the "liberal thrust" and "permissive backdrop" of the rules.[67]

To replace the mechanical *Frye* test, the Court found in the text of Rule 702 a two-part obligation: "[T]he trial judge must ensure that any and all scientific testimony or evidence admitted is not only

---

60. 727 F. Supp. 570, 575 (S.D. Cal. 1989) (granting summary judgment), *aff'd*, 951 F.2d 1128 (9th Cir. 1991), *vacated and remanded*, 113 S. Ct. 2786 (1993), *on remand*, 43 F.3d 1311 (9th Cir. 1995).

61. *Daubert*, 113 S. Ct. at 2791 (emphasis added).

62. *Id.* at 2791-92.

63. 727 F. Supp. at 575. Contrary to the Supreme Court's approach, the district grounded its general acceptance test not in Rule 702, but in Rule 703. *See id.* at 572; *infra* notes 109-19 and accompanying text.

64. 727 F. Supp. at 575-76.

65. 951 F.2d 1128 (9th Cir. 1991).

66. 113 S. Ct. at 2794.

67. *Id.*

relevant, but reliable."[68] The reliability requirement emerges from Rule 702's reference to "scientific . . . knowledge": "The adjective 'scientific' implies a grounding in the methods and procedures of science. Similarly, the word 'knowledge' connotes more than subjective belief or unsupported speculation."[69] The relevance criterion derives from the Rule's requirement that the scientific testimony will "assist the trier of fact" in order to be admissible.[70]

The Court devoted most of its attention to the standards for assessing reliability. Initially, it defined science in epistemological terms. That is, " 'science is not an encyclopedic body of knowledge about the universe' " but rather " 'a *process* for proposing and refining theoretical explanations about the world that are subject to further testing and refinement.' "[71] In sum, "in order to qualify as 'scientific knowledge,' an inference or assertion must be derived by the scientific method."[72]

Before admitting scientific evidence, trial judges must therefore make a preliminary assessment of whether the proffered testimony is scientifically valid, in the sense of being the product of the scientific method.[73] (Trial judges will no doubt be heartened that the Supreme Court is "confident that federal judges possess the capacity to undertake this review.")[74] In so doing, they are directed to consider at least four factors.

The first question to ask in deciding whether a theory or technique is scientific and thus reliable is "whether it can be (and has been) tested."[75] Quoting without elaboration from Karl Popper and other philosophers of science, the Court enjoined the trial bench to seek such hallmarks of science as " 'falsifiability, or refutability, or testability.' "[76]

A second factor is whether the work relied on has been subjected to the peer-review publication process. The presence of peer-reviewed publication should be viewed as a positive, "in part

---

68. *Id.* at 2795.
69. *Id.*
70. *Id.*
71. *Id.* (quoting Brief for American Association for the Advancement of Science and the National Academy of Sciences as *amici curiae* at 7-8, Daubert v. Merrell Dow Pharmaceuticals, Inc., 113 S. Ct. 2786 (1993) (No. 92-102)).
72. *Id.*
73. *Id.* at 2796.
74. *Id.*
75. *Id.*
76. *Id.* at 2797 (quoting KARL POPPER, CONJECTURES AND REFUTATIONS: THE GROWTH OF SCIENTIFIC KNOWLEDGE 37 (5th ed. 1989)).

because it increases the likelihood that substantive flaws in methodology will be detected."[77] The lack of publication should not be viewed as fatal, however; some good work may be too new, too narrow, or too innovative to have survived the peer-review process.[78] (And some may even be done by law professors, who normally publish in student-edited journals.)

Third, "the court ordinarily should consider the known or potential rate of error . . . and the existence and maintenance of standards controlling the technique's operation."[79] This cryptic directive is accompanied by nothing more than two case citations.[80] The Court gave no further clue to what it had in mind in broaching this complex and multifaceted topic. When dealing, for example, with an established laboratory procedure, it may be possible to estimate the "rate of error" fairly directly, by comparing the handling of control samples to that of experimental samples.[81] By contrast, in an epidemiological study of a possible relationship between an environmental hazard and a disease, "rate of error" might refer to the statistical probability that the observed relationship could occur as a matter of chance.[82]

Finally, *Frye* reappears, if only through the back door. Although not required, "widespread acceptance can be an important factor in ruling particular evidence admissible";[83] conversely, the lack of peer support for a particular technique "may properly be viewed with skepticism."[84] Presumably, general acceptance will be proved as it had been under *Frye*, by evidence of professional organizations that

---

77. *Id.*

78. *Id.*

79. *Id.*

80. The Court cites United States v. Smith, 869 F.2d 348, 353-54 (7th Cir. 1989) (surveying studies of the error rate of spectrographic voice identification technique) and United States v. Williams, 583 F.2d 1194, 1198 (2d Cir. 1978) (noting professional organization's standard governing spectrographic analysis), *cert. denied*, 439 U.S. 1117 (1979).

81. *See, e.g.*, Judith A. McKenna, Joe S. Cecil & Pamela Coukos, *Reference Guide on Forensic DNA Evidence*, *in* MANUAL, *supra* note 11, at 273, 292-93 (discussing DNA laboratory error rates).

82. *See, e.g.*, Linda A. Bailey et al., *Reference Guide on Epidemiology*, *in* MANUAL, *supra* note 11, at 121, 151-56 (discussing possible attribution of results of epidemiological study to "random error"); *cf.* Daubert v. Merrell Dow Pharmaceutical, Inc., 43 F.3d 1311, 1317 n.4 (observing difficulty of determining "rate of error" of plaintiffs' experts' reassessment of prior research). *See generally Developments, supra* note 10, at 1545-46 (reviewing cases dealing with rates of error).

83. *Daubert*, 113 S. Ct. at 2797.

84. *Id.*

set standards for using the method, published references to its successful use, and admissibility of similar testimony in other cases.[85]

The relevancy branch of the Rule 702 inquiry is more straightforward. The scientific evidence, even if reliable, must "fit" in the sense of providing helpful information on a fact in issue.[86] The Court noted that a scientifically valid study of the phases of the moon might be relevant to the factual question of whether a certain night was dark. That same valid study would have no relevance, however, if the disputed question was whether a particular individual was likely to have behaved irrationally on a particular night.[87]

In the concluding paragraph of its Rule 702 discussion, the Court made two highly significant points. First, it emphasized that the Rule 702 inquiry is "a flexible one" whose "overarching subject is the scientific validity—and thus the evidentiary relevance and reliability—of the principles that underlie a proposed submission."[88] Second, trial courts are directed to focus "solely on principles and methodology, not on the conclusions that they generate."[89] This implies, as the organized plaintiffs' bar has argued, that as long as an expert uses the scientific method, a trial court has no authority to exclude his or her testimony, no matter how absurd the results may seem. But as we shall see in the next section, and as Margaret Berger has correctly observed in the *Manual*,[90] this pronouncement is not as absolute as it appears.

B. Unanswered Questions

1. How Will the Lower Courts Apply the *Daubert* Standard?

Daubert left a number of important questions unanswered. The first is the practical one of how the lower courts are to apply such indeterminate criteria as "falsifiability" and "rate of error."[91] Given its pervasive ambiguities, *Daubert* had the potential to skew trial court results in either direction; scientific evidence might well have become

85. For a review of the various approaches to admissibility employed in cases decided under *Frye*, see *Developments*, *supra* note 10, at 1494-97.

86. *Daubert*, 113 S. Ct. at 2795-96.

87. *Id.* at 2796.

88. *Id.* at 2797.

89. *Id.*

90. Berger, *supra* note 26, at 104-06, 113-17 (assessing propriety of using Rules 703 and 403 to exclude evidence that seems to satisfy Rule 702).

91. *Daubert*, 113 S. Ct. at 2797.

significantly harder or easier to admit than under *Frye*.[92] But a preliminary sampling of decisions published in August 1994 suggested that *Daubert* was having little impact on case outcomes. Its author, Thomas J. Mack, argued that trial judges were simply reaching pre-*Daubert* results and then dressing them up in post-*Daubert* rationales.[93]

Outcomes aside, it is difficult to determine at this point whether the lower courts are doing more work, either qualitatively or quantitatively, in response to *Daubert*. Two of the most widely cited post-*Daubert* cases suggest opposite conclusions. In *In re Paoli Railroad Yard PCB Litigation*, the Third Circuit implied that courts that had already moved away from *Frye* in the direction of a more discretionary reliability test could adapt to *Daubert* with little difficulty.[94] The Ninth Circuit's opinion on remand in *Daubert* itself provides a striking contrast.[95] Judge Kozinski's discussion of the *Daubert* standard drips with undisguised sarcasm. Under the heading "Brave New World,"[96] he characterized the *Daubert* burden as follows:

> [T]hough we are largely untrained in science and certainly no match for the witnesses whose testimony we are reviewing, it is our responsibility to determine whether those experts' proposed testimony amounts to "scientific knowledge," constitutes "good science," and was "derived by the scientific method."[97]

Addressing the reliability question, the Ninth Circuit observed that the Supreme Court's list of criteria was nonexclusive, and added a new and "very significant" criterion of its own: whether the expert testifies on the basis of research that was conducted independent of the litigation.[98] All of the *Daubert* plaintiffs' experts failed this test, having developed their opinions for the purpose of testifying.[99] If

---

92. *See generally* Berger, *supra* note 26, at 45-47 (assessing impact of *Daubert*).

93. Thomas J. Mack, *Scientific Testimony After* Daubert*: Some Early Returns from Lower Courts*, TRIAL, Aug. 1994, at 23, 24.

94. 35 F.3d 717, 742 (3d Cir. 1994) (emphasizing consistency between *Daubert* and the Third Circuit's earlier decision in United States v. Dowering, 753 F.2d 1224 (3d Cir. 1985)).

95. Daubert v. Merrell Dow Pharmaceuticals, Inc., 43 F.3d 1311 (9th Cir. 1995).

96. *Id.* at 1315.

97. *Id.* at 1316.

98. *Id.* at 1317.

99. *Id.* Berger points out that under the *Frye* "general acceptance" standard, such techniques as voiceprint analysis, handwriting analysis, and paraffin testing were readily admitted even though they had been developed purely for forensic purposes and their validity had never been established empirically. Berger, *supra* note 26, at 75-76. She

this is to be a dispositive factor in the future, then outcomes will change, and probably in a dramatic way. The exclusion of expert testimony will tip the legal balance against those plaintiffs unfortunate enough to be injured by something that has not already been the subject of basic research.

The Ninth Circuit also engaged in an extended analysis of the "fit" requirement, which the Supreme Court had disposed of in a single paragraph. The Ninth Circuit expressed some qualms about granting summary judgment through the simple expedient of rejecting the plaintiff's expert affidavits on the newly articulated reliability grounds: The plaintiffs had, after all, submitted those affidavits when they reasonably believed that *Frye* was the law.[100] The court found the answer to its dilemma in California tort law, which "requires plaintiffs to show not merely that Bendectin increased the likelihood of injury, but that it more likely than not caused *their* injuries."[101] The best that could be said, however, for those plaintiffs' affidavits that conceivably might have survived the reliability analysis was that they tended to show " 'that Bendectin could *possibly* have caused plaintiffs' injuries.' "[102] Thus, even if the plaintiffs' experts' *methods* were deemed sufficiently reliable, their *results* were not relevant to the issue specified by the controlling legal standard. Because the plaintiffs bore the burden of proof on causation, the exclusion of their expert evidence led inevitably to summary judgment against them.[103]

---

questions whether comparable developments are likely under *Daubert*, "with its emphasis on testing." *Id.* at 75.

100. *Daubert*, 43 F.3d at 1319-20.

101. *Id.* at 1320.

102. *Id.* at 1322 (quoting the district court opinion, 727 F. Supp. 570, 576 (S.D. Cal. 1989)).

103. *Id.* The court suggested that epidemiological evidence can support a prima facie case only if it demonstrates that the suspect substance can increase an exposed person's *relative risk* of an associated disease by a factor of more than two. Relative risk is the ratio of the incidence of the associated disease in the exposed population to its incidence in the unexposed population. For example, a ratio of 10% to 5% would indicate a relative risk of two. *Id.* at 1320-21; *see* Bailey et al., *supra* note 82, at 168-69; *infra* notes 221-26 and accompanying text. According to the court, a relative risk greater than two is equivalent to a legal showing of causation by a preponderance of the evidence. By contrast, a relative risk of less than two, while possibly suggestive, "actually tends to disprove legal causation, as it shows that Bendectin does not double the likelihood of birth defects." *Daubert*, 43 F.3d at 1321 (footnote omitted).

The logic of the court's analysis is as follows: A relative risk of 1.0 means that the suspect agent has no influence on the incidence of disease (the rates of incidence are the same in the exposed and unexposed populations). A relative risk of 2.0 means that incidence doubles with exposure. At a 2.0 level, the agent can be assumed to be responsible for as many cases of the disease as all the background (non-exposure) causes

On one level, the Ninth Circuit's analysis is perfectly straightforward. It found that the plaintiffs' studies, even if methodologically reliable, simply did not "fit" the applicable legal standard. At best, they were sound studies of the wrong question. On a deeper level, however, the court's analysis raised what may be the most difficult question left unresolved by *Daubert*: the extent to which a trial court can evaluate an expert's conclusions in ruling on admissibility.

As noted above, the plaintiffs' bar has attacked the *Manual* for even suggesting that a court might examine the conclusions of a study in ruling on its admissibility.[104] However, how can a court evaluate the "fit" between expert testimony and the issues in the case unless it can examine the substance of that testimony? Looking at whether an expert's work has been peer-reviewed or has a low rate of error sheds little light on whether what the expert has to say will assist the trier of fact in deciding a material issue. At a minimum, trial courts need to be able to evaluate conclusions in the limited sense of determining whether the expert's methods yield answers to questions that are properly within the purview of the trier of fact. Whether the court, when deciding on admissibility, is entitled to make a preliminary value judgment about the soundness of those answers is a much more controversial question.

Yet this is exactly what the Ninth Circuit did on remand in *Daubert*, albeit by an indirect route. Recall that the Ninth Circuit found that the plaintiffs' studies were irrelevant because they shed light only on the question of whether Bendectin "could *possibly* have caused plaintiffs' injuries," when the dispositive question was whether Bendectin *did* cause those injuries.[105] Had the plaintiffs' studies been addressed to the former, immaterial question, their exclusion on relevancy grounds would have been noncontroversial. But they were not, of course; their purpose was to show that Bendectin did cause the plaintiffs' injuries.[106] What the Ninth Circuit actually did was to delve into the results of the studies and conclude that, on their

---

combined. Thus, a relative risk greater than 2.0 suggests that the suspect agent is responsible for more than 50% of all cases of the disease; alternatively, it implies a 50% likelihood that an exposed individual's disease was caused by the agent. *See* Bailey et al., *supra* note 82, at 168-69. Nonetheless, epidemiologists are careful to point out that "the cause of an individual's disease . . . is beyond the domain of the science of epidemiology." *Id.* at 167 (footnote omitted).

104. *See supra* notes 29-39 and accompanying text.

105. *See supra* notes 101-02 and accompanying text.

106. *Daubert*, 43 F.3d at 1314 ("It is by such means that plaintiffs here seek to establish that Bendectin is responsible for their injuries.").

scientific merits, they could not support an opinion on the dispositive legal question.[107]

There is nothing in the text of *Daubert* to justify the approach that the Ninth Circuit took. *Daubert* says, rather, that in ruling on admissibility under Rule 702, the trial judge should: (1) identify the questions that the expert is raising, (2) decide whether the expert has used reliable methods to study those questions, and (3) if so, decide whether the questions are relevant to material issues in the case.[108] There is no mandate for considering whether the experts' conclusions are right or wrong, sound or unsound, when deciding on admissibility. Contrary to the complaints of the plaintiffs' bar, nothing in Margaret Berger's chapter in the *Manual* suggests otherwise.

2. What Is Left of Rule 703?

The Ninth Circuit's application of the *Daubert* standard on remand, however misguided it may have been, does suggest a further question: If it is inappropriate for a court to evaluate an expert's conclusions when making an admissibility decision under Rule 702, then when, if ever, can it do so? Obviously, the court must look at the conclusions if it happens to be sitting as the trier of fact. It would seem equally obvious that the court should do so in deciding whether the expert's testimony is sufficient to withstand a motion for summary judgment. The Supreme Court suggested in dictum at the end of the *Daubert* opinion that there are likely to be cases in which the plaintiff's expert testimony, while admissible, is not sufficient to defeat a defense motion for summary judgment or a directed verdict.[109] One might reasonably ask, however, how a trial court that had already found the evidence both scientifically reliable and legally relevant could then, without invading the province of the jury, find the same evidence legally insufficient.[110] With reliability and relevance already determined, what grounds for criticism would be left except credibility, which is always left to the trier of fact? Although Berger notes that the "often blurred" distinction between the admissibility and the sufficiency of scientific evidence "is clearly

---

107. *See supra* note 103. The court might plausibly have rejected the plaintiffs' studies under the criterion of reliability because of their failure to deal adequately with relative risk; instead, it addressed relative risk under the heading of fit. 43 F.3d at 1320-21.

108. Daubert v. Merrell Dow Pharmaceuticals, Inc., 113 S. Ct. 2786, 2795 (1993).

109. *Id.* at 2798.

110. *Cf. Developments*, *supra* note 10, at 1550-51 (discussing cases seeking to preserve the distinction between the admissibility and summary judgment standards).

reaffirmed in *Daubert*," her discussion in the *Manual* offers no assistance in resolving this logical dilemma.[111]

Berger is more forthcoming in discussing the backdoor route to the consideration of expert conclusions that so troubled the plaintiffs' bar—Rule 703.[112] The *Daubert* Court mentioned Rule 703 only in passing, observing that it "provides that expert opinions based on otherwise inadmissible hearsay are to be admitted only if the facts or data are 'of a type reasonably relied upon by the experts in the particular field in forming opinions or inferences upon the subject.' "[113] As a practical matter, it might be argued, very little is left of Rule 703: Does a finding of scientific validity under Rule 702 not necessarily imply that the expert relied on sources that would be deemed reasonable by those in the field?

Berger properly identifies, however, several theories that support Rule 703's continued vitality. One possibility, Berger suggests, drawing on a concurring opinion in *Christophersen v. Allied-Signal Corporation*,[114] is that Rule 703 supplements the Rule 702 reliability inquiry, but only when the expert is relying on hearsay sources.[115] Thus, as long as the expert is relying on admissible sources, the Rule 702 reliability test is wholly dispositive. But if some of the expert's sources are hearsay, they must pass the further test of Rule 703's "reasonably relied upon" standard.[116]

Rule 703 may have a broader impact, however. According to the majority opinion in *Christophersen*, as quoted by Berger, " 'district judges may reject opinions founded on critical facts that are plainly untrustworthy, principally because such an opinion cannot be helpful to the jury."[117] Does this interpretation survive *Daubert*? *Christophersen*'s term "trustworthy" seems to have been superseded by

---

111. Berger, *supra* note 26, at 52-53.

112. Rule 703, "Bases of Opinion Testimony by Experts" states:

> The facts or data in the particular case upon which an expert bases an opinion or inference may be those perceived by or made known to the expert at or before the hearing. If of a type reasonably relied upon by experts in the particular field in forming opinions or inferences upon the subject, the facts or data need not be admissible in evidence.

FED. R. EVID. 703.

113. 113 S. Ct. at 2798 (quoting FED. R. EVID. 703). The district court had relied on Rule 703 as the basis for its general acceptance standard. *See supra* note 63.

114. 939 F.2d 1106, 1118 (5th Cir. 1991) (en banc) (Clark, C.J., concurring), *cert. denied*, 112 S. Ct. 1280 (1992).

115. Berger, *supra* note 26, at 105-06.

116. *Id.*

117. *Id.* (quoting *Christophersen*, 939 F.2d at 1114).

Daubert's "reliability" analysis, while *Daubert*'s construction of helpfulness (that is, "fit") does not seem to invite an open-ended inquiry into the expert's sources.

Berger (the plaintiffs' bar's institutional outcry notwithstanding) does not pass judgment on these issues. She does, however, conclude her discussion of Rule 703 by listing the following questions as asked but unanswered by the *Daubert* opinion:

> May a court rely on Rule 703 to exclude an expert's opinion that reaches a conclusion that is inconsistent with a scientific consensus or that lacks scientific foundation? Does such a reading constitute a backdoor resurrection of the *Frye* "general acceptance" test, which was rejected by the Court as incompatible with the Federal Rules of Evidence? Should a court use a sufficiency analysis rather than an admissibility analysis when an expert uses appropriate methodology and relies on data that experts reasonably rely upon but nevertheless reaches an opinion at odds with the scientific community?[118]

She is correct that these questions are at least latent in *Daubert*, and there is no doubt that they are unanswered. Their very existence will tempt trial court judges who are looking for a way to exclude ostensibly reliable scientific evidence that they simply do not like.[119]

3. Is *Daubert*'s Definition of Science Adequate?

As noted above, *Daubert* defines science in terms of its methods.[120] Science, in the Court's view, is the process of generating and testing hypotheses; those hypotheses must be explicitly, if not quantitatively, falsifiable.[121] While this definition of science may strike the legal reader as noncontroversial, it is by no means universally accepted in the contemporary academic community.

---

118. *Id.* at 111-12 (footnotes omitted).

119. The Ninth Circuit might well have relied on Rule 703 in its *Daubert*-remand opinion. The plaintiffs' epidemiologists "reanalyzed studies done by other scientists," 43 F.3d at 1311 (9th Cir. 1995); thus, they relied on nominally hearsay sources in developing their own opinions. The Ninth Circuit could have concluded that those underlying studies were not of the sort reasonably relied on by experts in the field because they did not reflect a relative risk greater than two.

Alternatively, a court might exclude a suspect study under Rule 403, which permits the exclusion of otherwise relevant evidence "if its probative volume is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury." *Daubert*, 113 S. Ct. at 2798; *see also* Ayers v. Robinson, 887 F. Supp. 1049 (N.D. Ill. 1995) (excluding expert testimony on value of life under both Rule 703 and Rule 403).

120. *See supra* notes 68-73 and accompanying text.

121. *Daubert*, 113 S. Ct. at 2796-97.

The idea that science is a dispassionate search for the truth conducted according to rigorously applied, value-neutral rules is usually called "positivism."[122] Most of the people whom the courts and the public would identify as "scientists" are indeed positivists, and conduct their research according to positivist ideals.[123]

But many who practice what they believe to be science reject positivism as unrealistic, undesirable, or both. So-called postmodernists argue that, especially in the social sciences, it is impossible to choose research topics, to set and apply criteria of validity, and especially to evaluate the work of others without being influenced by politics and ideology.[124] Since so much of science is inevitably interpretive, they argue, we should own up to what we are doing and abandon positivism as a deceptive myth.[125]

People who see themselves as inhabiting the real world may dismiss this critique as merely another ephemeral academic trend—what happens when smart people with huge egos have too much time on their hands. But even if postmodernism is largely written off as an ivory-tower reverie, there is a core point that cannot be avoided: Many things that the law, the political system, and the public routinely accept as "science" cannot fit within a rigorously positivist definition of the sort propounded by the *Daubert* Court.

Consider that commonplace scientific miracle, the visit to the doctor's office that results in the successful treatment of a disease. In most instances, the physician's primary tool will be differential diagnosis.[126] This involves identifying the range of possible causes of the patient's condition, comparing and contrasting their respective symptoms, and then examining and testing the patient in an effort to rule out all but the most likely diagnosis.[127] Is this science? An immediate reaction is, Of course it is. Medical doctors have years of scientific training, and they agree that this is the right way to proceed.

---

122. *See* John M. Conley, *The Social Science of Ideology and the Ideology of Social Science*, 72 N.C. L. REV. 1249, 1250, 1254-55 (1994).

123. *See* Margaret G. Farrell, Daubert v. Merrell Dow Pharmaceuticals, Inc., 15 CARDOZO L. REV. 2183, 2189-93 (1994).

124. *See* Conley, *supra* note 122, at 1250-51; JOHN K. SMITH, AFTER THE DEMISE OF EMPIRICISM: THE PROBLEM OF JUDGING SOCIAL AND EDUCATIONAL INQUIRY 9-10 (1993). For a readable general introduction to postmodernism, see ERNEST GELLNER, POSTMODERNISM, REASON AND RELIGION (1992).

125. *See* Conley, *supra* note 122, at 1251-52.

126. *See* Ellen Relkin, *Some Implications of* Daubert *and Its Potential for Misuse: Misapplication to Environmental Tort Cases and Abuse of Rule 706(a) Court-Appointed Experts*, 15 CARDOZO L. REV. 2255, 2258-59 (1994).

127. *See id.* at 2258 n.12.

Moreover, even a cursory glance at the last hundred years of medical history indicates that differential diagnosis works astonishingly well.

But is differential diagnosis science in the narrower, positivist sense specified by *Daubert*?[128] Can one test the judgment of a physician in sorting out unquantifiable probabilities? How would one calculate the rate of error for a large sample of differential diagnoses when each one is rendered in a unique context? Yet the Supreme Court could not have meant to exclude medical testimony based on differential diagnosis.[129]

To take one more example, what about evolutionary biology? The Supreme Court thinks that this is science, at least when contrasted with creationism.[130] Darwin's hypothesis of natural selection cannot be tested, except against the remote evidence of the fossil record. No one has ever seen natural selection at work at the level of individual organisms. Moreover, from a present sensory perspective, the whole concept is absurd. What would have induced fish to grow limbs that would ultimately help them walk on land, when at the time they grew the appendages they were much less efficient than fins for moving through water?[131] One could readily conclude that creationism provides a more parsimonious explanation for these complexities.

Despite these ironies, however, most of us—including those of us who are judges—readily accept differential diagnosis and evolutionary biology as "science" in every sense of the word. We do so not because they satisfy a positivist checklist, but because common sense tells us that they are trustworthy explanations of natural phenomena. In the medical example, we observe that large numbers of people go to the doctor and, after being subjected to differential diagnosis, get better. In the case of evolutionary biology, either we defer to the authority of the experts or, unless impelled in other directions by religious conviction, think about the problem for a while and conclude that natural selection really is the best available explanation of why

---

128. *See id.* at 2259.

129. *Cf.* Carroll v. Litton Sys., 1995 U.S. App. LEXIS 2015, at *6-7 & n.6 (4th Cir. Feb. 1, 1995) (distinguishing, in toxic tort case, between admissible testimony of physicians that plaintiffs' ailments were consistent with diagnosis of exposure to suspect agent, and inadmissible testimony of same physicians that was based on unsubstantiated theory of "environmental half-life" of agent).

130. *See* Edwards v. Aguillard, 482 U.S. 578 (1987) (invalidating Louisiana law requiring school teachers who teach evolutionary theory to discuss creationism also).

131. For a clever biological resolution of this particular conundrum, see Carl Zimmer, *Coming Onto the Land*, DISCOVER, June 1995, at 118.

Velociraptor (the villain of *Jurassic Park*) had that nasty middle finger.

In the wake of *Daubert*, the pertinent question is how the Supreme Court intended the lower courts to deal with evidence that meets the everyday, common-sense criteria for science but falls outside the strict positivist definition. One clue may lie in a footnote in which the Court observed that, although Rule 702 applies not only to science but also to " 'technical or other specialized knowledge,' "[132] the holding of the case was limited to "scientific" evidence.[133] Thus, the answer may be that any expert evidence which lies outside the bounds of positivist science is to be evaluated under the pre-*Daubert* standards for nonscientific expert testimony, which one leading commentator has characterized as "laissez-faire."[134]

This interpretation would lead to perverse results, however. Arguably scientific research that seemed likely to fail the *Daubert* test could then be repackaged as mere "technical or other specialized knowledge" and admitted under the more lenient non-science standard. A more plausible, if textually unsupportable, interpretation is that the Court intended *Daubert* to be the exclusive standard for anything that one might reasonably characterize as science. Under this view, the "technical or other specialized knowledge" category was intended not for bad science or almost-science, but for other kinds of expert testimony, such as the valuation opinions of appraisers.[135]

Although the latter view is our preferred solution, it leaves no safe harbor for evidence that is widely viewed as scientific, is accepted as sound, but cannot meet the *Daubert* criteria. This appears to be a dilemma that the lower courts will have to resolve on their own, without help from either the text of *Daubert* or the *Manual*.[136]

---

132. *Daubert*, 113 S. Ct. at 2795 n.8 (quoting FED. R. EVID. 702).

133. *Id.*

134. Edward J. Imwinkelried, *The Next Step After* Daubert: *Developing A Similarly Epistemological Approach to Ensuring the Reliability of Nonscientific Expert Testimony*, 15 CARDOZO L. REV. 2271, 2280 (1994).

135. For example, in Ayers v. Robinson, 887 F. Supp. 1049, 1064 (N.D. Ill. 1995), the court rejected the testimony of an expert who valued a life by means of a "willingness-to-pay" approach under the *Daubert* criteria for scientific reliability. By contrast, another recent case held that a quasi-scientific opinion concerning metal fatigue should be evaluated under Rule 701's lay opinion standard; it found the testimony inadmissible on the record before it. Asplundh Mfg. Div. v. Benton Harbor Eng'g, 57 F.3d 1190, 1204-05 (3d Cir. 1995).

136. Berger deals indirectly with this issue in her discussion of *Daubert*'s application to the social sciences. *See* Berger, *supra* note 26, at 84-87 (reviewing debate between David

III. WHAT DO JUDGES NEED TO KNOW ABOUT SCIENCE?

Whatever its ambiguities, *Daubert* does make one thing clear. For judges, the scientific method must become the intellectual equivalent of pornography: They must be able to know it when they see it.[137] *Daubert* does not require, however, that judges be familiar with the substance of any particular branch of science. On the contrary, at least at the gatekeeping stage, *Daubert* all but forbids comparing an expert's results to the accepted wisdom in the field.[138] Thus, *Daubert*'s mandate is not to learn a little bit about many areas of scientific investigation, but rather to learn a fair amount about science itself.

The authors have between them many years of experience teaching science to judges.[139] We can state with a conviction borne of that experience that judges, notwithstanding their general enthusiasm and diligence, tend to be highly resistant to the sort of learning that *Daubert* demands. Time and time again, we have been told that studying methodology is too abstract, mere theory; judges are practical people who want to get down to concrete examples and talk about whether particular "studies" should be admitted into evidence.[140] On our more sanguine days, we try to believe that the judges are telling us that they are willing to learn methodology, but want to do so in an inductive way. But they are not: Our nearly

---

McCord and David L. Faigman over proper evidentiary treatment of "soft" sciences).

137. *See* Jacobellis v. Ohio, 378 U.S. 184, 197 (1964) (Stewart, J., concurring).

138. *See* Daubert v. Merrell Dow Pharmaceuticals, Inc., 113 S. Ct. 2786, 2797 (1993) ("The focus of course must be solely on principles and methodology, not on the conclusions they generate.").

139. One of the authors (Conley) teaches the social science component of the University of Virginia's Graduate (LL.M.) Program for Judges and is faculty director for Duke Law School's annual Judging Science summer program. The other (Peterson) lectures on statistics at the Duke program and has testified in numerous cases as an expert in the fields of statistics and computer science. Both authors frequently lecture on statistics to continuing judicial education audiences.

140. The ongoing work of John Monahan and Laurens Walker is the most significant attempt to impose some theoretical coherence on the admissibility of individual studies. Their proposal, in oversimplified terms, is that in some circumstances, courts might elevate particular social science studies to the level of precedent, which would permit future courts to cite and rely on them as they do cases. *See, e.g.*, Laurens Walker & John Monahan, *Social Facts: Scientific Methodology as Legal Precedent*, 76 CAL. L. REV. 877 (1988). They endeavor to reconcile their theory with the *Daubert* standards in John Monahan & Laurens Walker, *Judicial Use of Social Science Research After* Daubert, 2 SHEPARD'S EXPERT & SCI. EVIDENCE Q. 327 (1994).

uniform experience with hundreds of judges at every level is that they think methodology is something for academics to worry about.

This zeal for getting to the evidentiary bottom line—a kind of judicial Occam's razor—works in tandem with other problems to make judges a uniquely difficult audience for science education. First, because science aptitude plays no part in judicial selection, judges range from closet Einsteins to proud Luddites.[141] Moreover, judges are rarely told that their ways of doing things are wrong (except, occasionally, by their appellate supervisors). Instead, lawyers and witnesses feel compelled to adapt to the judges' idiosyncracies; what the judge is not interested in becomes unimportant. This state of affairs makes it extremely difficult to persuade judges to put aside the modes of analysis with which they are comfortable and to approach problems from new and sometimes forbidding perspectives. And finally, we teachers lack any effective "stick": No judge has ever been impeached for failing a statistics course.

The result of all this, in our experience, is that judicial education is often counterproductive. Instead of a few principles of general applicability, what judges take away from courses, seminars, and handbooks are fragments of information about particular kinds of cases. Having mastered a few details of, say, a toxic tort case, a judge will be tempted to transfer that learning to every case involving epidemiological proof. But unless the judge understands the principles that underlie the details, the result can be the misapplication of rules that had no bearing on the situation in the first place.

Two contrasting examples will illustrate the problem: the Fourth Circuit's unsuccessful foray into statistical analysis in *Bazemore v. Friday*,[142] and District Judge Jack B. Weinstein's admirable scientific work in the enormously complex *Agent Orange* litigation.[143] In the *Bazemore* case, the plaintiffs alleged racial discrimination against black employees of the North Carolina Agriculture Extension Service. Numerous practices of the employer were challenged, and after digesting a large amount of statistical evidence, the trial court ruled

---

141. The remainder of this section provides contrasting examples of judicial scientific aptitude. For a critical review of various courts' efforts to apply basic scientific principles, see *Developments*, *supra* note 10, at 1536-46.

142. 751 F.2d 662 (4th Cir. 1984), *aff'd in part, vacated in part, and remanded*, 478 U.S. 385 (1986) (per curiam), *on remand*, 848 F.2d 476 (4th Cir. 1988).

143. There are numerous opinions in this case. The best illustration of Judge Weinstein's scientific dexterity may be *In re* "Agent Orange" Prod. Liab. Litig., MDL No. 381, 611 F. Supp. 1223 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987).

for the defendants on all counts.[144] On appeal, the Fourth Circuit affirmed.[145] Rather than simply adopting the district court's findings, however, the court of appeals undertook its own statistical analysis of the extent to which the performance ratings awarded to black employees were similar to those given to white employees.[146] The requisite data were already in the record, and the plaintiffs had somehow failed to analyze them as part of their case in chief at trial. Understandably, the Fourth Circuit was curious about the issue.

On its own initiative, the court compared the proportion of black employees receiving any of the top three of four rating levels with the proportion of white employees receiving such ratings, and concluded that the two proportions were not significantly different.[147] It repeated the analysis for each of the six geographic districts into which the Extension Service is organized, and found that for no district were the racial proportions substantially different.[148] However, these analyses contain three significant errors that, in combination, led the court materially astray.

First, built into the court's analysis is the presumption that since there were four levels in the rating system, 25% of all employees would be found at each level. That this presumption is false is obvious from the data themselves: In fact, only 20.1% of employees fell into the lowest category.[149] The court's erroneous presumption caused it to understate the extent of the underrepresentation of black employees at the upper three levels. Second, while it was important to examine each of the six districts in turn for evidence of disparity, the court failed to consider the cumulative effect of the disparities across all districts.[150] Thus, while no single district may have exhibited a significant racial disparity, the fact that most or all of them showed at least some disparity to the disadvantage of black employees was significant.[151] And finally, by not examining the

144. 478 U.S. at 386.
145. *Id.*
146. 751 F.2d at 672-74.
147. *Id.* at 673.
148. *Id.* at 673-74.
149. *See id.* at 673 (providing raw data).
150. *See id.*
151. For a discussion of statistical significance, see *infra* note 159. The point here is the simple one that a series of seemingly minor discrepancies can produce a compelling pattern. Five coin flips may not reveal that a coin is weighted, but 500 probably will. In the instant case, a rank sum analysis of the racial pattern of performance ratings within each of the five districts having both black and non-black employees revealed imbalances of 1.69, 1.10, 2.25, 2.48 and 0.85 standard deviations, all to the disadvantage of black

breakdown by race within each of the top three rating levels (even though this degree of detail was exhibited in its tables),[152] the court failed to notice that white employees were consistently rated more highly than black employees. The overall effect of these miscues was to make a racial disparity well in excess of three standard deviations appear to the court to be only about one standard deviation.[153] The court thus converted statistics that actually favored the plaintiffs into dispositive defense evidence.[154]

Perhaps these errors occurred because the only statistical formulas the court (and its clerks) knew were those for comparing two proportions, and it simply did the best it could in applying that knowledge. Had the court possessed a surer grasp of the principles of statistical inference, it would have realized that somehow they could be applied to the comparison by race of the entire pattern of performance ratings. Then, even if it did not itself know the appropriate formulas, the court could have sought professional help from the parties or an expert appointed by the trial court.

Contrast the work of the Fourth Circuit in *Bazemore* with that of Judge Weinstein in the *Agent Orange* litigation. In one of his numerous opinions in the case, Judge Weinstein dealt with the claims of plaintiffs who had opted out of the *Agent Orange* class actions.[155]

---

employees. The overall cumulation of these imbalances comes to 3.76 standard deviations, larger than any one of the five individual imbalances, and well in excess of all commonly used thresholds for defining statistical significance. For a discussion of standard deviation, see *infra* note 153.

152. *Bazemore*, 751 F.2d at 673 (reproducing court's tables, formulas, and calculations).

153. "The standard deviation can be thought of as a measure of the typical or expected variation of the numbers in a group from their average." BARNES & CONLEY, *supra* note 1, at 129. According to the Supreme Court, when the representation of a protected group in a work force is more than two or three standard deviations below what might be expected in the absence of discrimination, discrimination may be inferred, unless some other explanation is offered. *See* Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 311 & n.17 (1977).

154. The court of appeals' misinterpretation of the performance ratings raised an ethical dilemma for the defense experts on remand. Although the court of appeals' decision was reversed in part by the Supreme Court, its findings regarding the fairness of the performance rating system survived review. Should the defense remain silent and take advantage of an erroneous statistical conclusion which was now the law of the case? *See* Bazemore v. Friday, 848 F.2d 476 (4th Cir. 1988) (remanding to district court for further proceedings following Supreme Court disposition). The case was settled before further district court proceedings began. This and other issues in the case are discussed more fully in WALTER B. CONNOLLY & DAVID W. PETERSON, THE USE OF STATISTICS IN EQUAL EMPLOYMENT OPPORTUNITY LITIGATION, app. F (1980 & Supp. 1995).

155. *In re* "Agent Orange" Prod. Liab. Litig., MDL No. 381, 611 F. Supp. 1223 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 187 (2d Cir. 1987).

They, like the class plaintiffs, alleged that they suffered from various health problems as a result of being exposed to the herbicide Agent Orange while serving in the Vietnam war.[156] In four succinct pages, Judge Weinstein reviewed the then-current state of the relevant epidemiological evidence, concluding that it would not support a finding of a causal relationship between Agent Orange exposure and the plaintiffs' health problems.[157]

In reviewing the available studies, Judge Weinstein's opinion focused over and over again on just a couple of essential concepts. First, he correctly defined the objective of the epidemiological research as the detection of abnormally high incidences of health problems among the population that had been exposed to Agent Orange.[158] He recognized that an inference of causation might be supported by evidence of statistically significant differences in disease rates between otherwise comparable exposed and unexposed populations.[159] Finding no basis in the available epidemiological studies for an inference of causation and rejecting the plaintiffs' non-epidemiological evidence, he entered summary judgment for the defendants.[160]

Reasonable minds might differ concerning some of the specifics of Judge Weinstein's analysis. For example, did he put undue stress on statistical significance as a Rubicon dividing reliable from unreliable epidemiological studies?[161] Should he have considered the potential biasing effect of the government sponsorship of most of

156. 611 F. Supp. at 1228.

157. *Id.* at 1231-34.

158. *Id.* at 1231.

159. *See id.* at 1231-34. A finding is said to be statistically significant if chance, acting alone, probably would not have caused it. In many situations, the probability of the chance occurrence of a relationship or difference (in this case, the difference between the disease rates in the exposed and unexposed populations) of a particular magnitude can be calculated and expressed in terms of a p-value. *See* BARNES & CONLEY, *supra* note 1, at 32-34. In many research contexts, a relationship or difference will be considered statistically significant if $p<.05$; that is, if there is less than a one-in-twenty probability that a comparable finding would have emerged as a matter of chance alone. In such a case, the chance explanation (the null hypothesis) will be rejected, and other explanatory hypotheses (for example, that Agent Orange exposure did cause a particular disease) may be entertained. *See id.*

160. 611 F. Supp. at 1230-31, 1234, 1260, 1264.

161. *See e.g., id.* at 1233 (rejecting study of Australian veterans for failure to find statistically significant increases in death rates following exposure). For a discussion of over-reliance on statistical significance, see *Developments*, *supra* note 10, at 1544.

the studies he cited?[162] Was he too quick to disregard the clinical judgment of the plaintiffs' experts?[163]

But irrespective of one's view on these questions, Judge Weinstein's approach is exactly what *Daubert* demands. He did not confound the adversary system or run the risk of scientific disaster by introducing his own evidence or redoing the experts' analyses. Rather, using transcendent principles of quantitative analysis, he assessed a vast body of complex evidence in a succinct, coherent, and clearly defensible manner. Although he framed his analysis in terms of the pre-*Daubert* understanding of the expert testimony rules, it could readily be recast under the *Daubert* headings of reliability and fit. We turn next to an assessment of whether the *Manual* is likely to assist judges in duplicating Judge Weinstein's accomplishment in other contexts.

IV. DOES THE *MANUAL* DO ITS JOB?

We conclude this essay by evaluating the *Manual*'s success in achieving the objectives we identified in Part III: giving judges a working familiarity with the scientific method and some of science's most important operating principles. To do so, we examine four reference guides that deal with two major topics. First, we consider together the *Reference Guide on Statistics* by Kaye and Freedman and that on *Multiple Regression* by Daniel L. Rubinfeld, and we then discuss the reference guides on the closely related topics of *Epidemiology* and *Toxicology* by, respectively, Linda A. Bailey, Leon Gordis, and Michael Green; and Bernard D. Goldstein and Mary Sue Henifin.

A. Statistics and Regression

Because statistical concepts underlie each of the other seven reference guides, the *Reference Guide on Statistics* is the logical place to collect those concepts for exposition and explanation. To a large degree this is the chapter in which the essence of the methods applied in all the others should reside. He who masters this chapter, one would think, has grasped scientific fundamentals having wide

---

162. *See* 611 F. Supp. at 1232-34 (discussing studies sponsored by the United States and Australian governments).

163. *See id.* at 1235-39. The scientific status of clinical diagnosis is discussed *supra* at notes 126-29 and accompanying text.

application; she who understands these materials should be able to handle her responsibilities under *Daubert.*

The *Reference Guide on Statistics* opens with a discussion of the breadth of the field of statistics,[164] and then poses and responds to a series of questions concerning data collection, presentation, and interpretation.[165] There is an appendix providing technical details on the standard error, normal curve, and p-value,[166] and also a glossary containing explanations of technical terms.[167]

The question-and-answer format is engaging, and is a fresh departure from the traditional axiomatic treatment of this material. Such questions as "Are the measurements recorded correctly?"[168] and "Can the results be generalized?"[169] can be taken directly from the paragraph subheadings and used in the courtroom. The question-and-answer format extends over a wide range of topics, highlighting many aspects of a statistical study that might make it unreliable. What is more difficult to extract from the text and carry into the courtroom are the truly fundamental ideas—the scientific essence—that the post-*Daubert* court needs. It is one thing to ask the right question, quite another to be able to evaluate its answer.

The most fundamental of these ideas may be the relationship between a research design and the statistical properties of the resulting data. Most quantitative research designs are based on one of two ideals: the ideal of a designed or controlled experiment,[170] and the ideal of a random sample.[171] Inherent in each of the ideal designs is a mechanism for the calculation of probabilities. These probabilities, in turn, give definition to such terms as p-value,

---

164. Kaye & Freedman, *supra* note 43, at 335-39.

165. *Id.* at 341-87.

166. *Id.* at 389-93. Standard error is closely related to standard deviation, which is discussed *supra* at note 153. The normal curve is the familiar bell curve. *See* Kaye & Freedman, *supra* note 43, at 401. P-values are discussed *supra* at note 159.

167. Kaye & Freedman, *supra* note 43, at 395-412.

168. *Id.* at 342-43.

169. *Id.* at 349-50.

170. "The true experiment is defined by the random assignment of subjects either to an experimental (or treatment) group or to one or more comparison (or control) groups that do not receive the treatment under investigation." JOHN MONAHAN & LAURENS WALKER, SOCIAL SCIENCE IN LAW 59 (2d ed. 1990); *see* Kaye & Freedman, *supra* note 43, at 397. Consider, for example, a test of a new drug among 500 patients; on the basis of a coin flip or some comparably random procedure, 250 patients get the drug (the experimental group), while the other 250 get a placebo (the control group).

171. A sample is a selected subset of a population; a random sample is one that is drawn on the basis of chance (rolling dice, for example). *See* Bailey et al., *supra* note 82, at 176 (epidemiology chapter).

statistical significance, standard deviation, confidence interval,[172] and expected value.[173] The further a research design strays from the relevant ideal, the less practical meaning we can attach to any of those terms.

Thus, a firm grasp of these two ideals is necessary for an informed evaluation of almost every application discussed in the *Manual*. On this critical point, however, the *Reference Guide on Statistics* fails to deliver. True, we are told that "[c]ontrolled experiments are, far and away, the best vehicle for establishing a causal relationship,"[174] but never do we get to see why, and in particular, never do we get to see how the p-value depends both on the experimental design and its outcome.[175] While a detailed example of a random sample is given in the appendix,[176] its elemental features are swamped by the plethora of possible outcomes, causing the authors to retreat quickly to formulas, and leaving the reader with little understanding of the link between a confidence interval and the sampling circumstances that drive it. This is especially disappointing, because one of the authors, David Freedman, is the lead author of a text that provides an outstanding introduction to each of these two fundamental ideals.[177]

Aside from this strategic omission, there are some tactical miscues in the *Reference Guide on Statistics*. The issue of confidence intervals in relation to p-values is not well-treated. It is suggested that there is some tension as to which is more appropriate or more

---

172. When a quantity in a population is being estimated on the basis of a sample drawn from that population, the estimate is sometimes expressed as a range, called a confidence interval. Suppose, for example, that we wished to estimate the average height for the American adult male population on the basis of a random sample. Suppose that this results in a 95% confidence interval of 69-75 inches. This means that we can be 95% certain that the true population average lies within that range. *See* Kaye & Freedman, *supra* note 43, at 396 (statistics chapter); Bailey et al., *supra* note 82, at 173 (epidemiology chapter).

173. This is the average value that we would expect some particular quantity, or variable, to have on the basis of chance alone. If the variable were "percentage of heads" in a series of coin tosses, the expected value would be 50%. If over the course of 100 tosses the percentage were, say 20%, we would begin to suspect a loaded coin. We could assess the statistical significance of this discrepancy between observed and expected values by calculating a p-value: the probability that chance alone would result in a discrepancy of such magnitude. *See* BARNES & CONLEY, *supra* note 1, at 26-27.

174. Kaye & Freedman, *supra* note 43, at 347.

175. When the assumptions underlying the controlled experiment are violated, the p-value's practical utility is diminished.

176. Kaye & Freedman, *supra* note 43, at 389-91.

177. DAVID FREEDMAN ET AL., STATISTICS 3-7, 252-81, 328-34 (2d ed. 1991).

informative,[178] when in fact they are complementary. Where possible, both should be computed because they provide different information.[179] Elsewhere, the authors misstate an aspect of the relationship between standard deviations and p-values.[180] The glossary is useful for obtaining a rough sense of the meanings of the terms listed there, but several of the definitions are fuzzy or incomplete, and they are written in a style likely to intimidate the novice.[181]

The *Reference Guide on Multiple Regression* by Daniel Rubinfeld may be thought of as an extension of the statistics reference guide, in that regression is a particular type of statistical model.[182] Once again, the presentation is in question-and-answer format. The general topics addressed deal with the identification of the question to be addressed,[183] the choice of a statistical model,[184] the practical sig-

178. Kaye & Freedman, *supra* note 43, at 384 & n.164.

179. For example, suppose a 95% confidence interval on the gender coefficient in a regression analysis extends from -$100 to +$900. To infer the exact p-value from this information, one needs to know the number of degrees of freedom associated with the regression and to have access either to a sophisticated calculator or else to an appropriate probability table. Conversely, given the information that the gender coefficient estimate of $400 in a regression has a p-value of 0.07, one cannot ascertain the 95% confidence interval without knowledge of the number of degrees of freedom involved, and without either a complex calculation or else an appropriate probability table. Much of the practical appeal of the confidence interval is that it is expressed largely in natural units (in this case dollars), likely to be familiar to triers of fact. The appeal of the p-value is that its emphasis is on the weight to be given the evidence (of gender disparity in this example). For a discussion of regression analysis, see *infra* notes 183-93 and accompanying text.

180. *Id.* at 380 n.145. Specifically, note 145 perpetuates the myth that a three-standard-deviation disparity corresponds to a (two-tailed) p-value of about .01, when in fact the p-value associated with such a disparity is .0027. *Id.* The authors state the proposition correctly in the appendix. *Id.* at 391; *see* FREEDMAN ET AL., *supra* note 177, at A.86.

181. Kaye & Freedman, *supra* note 43, at 395-412. The Fisher's Exact Test, for example, is tied to the comparison of sample proportions, when instead it should be tied to an experiment involving fixed marginal totals. *Id.* at 399. The discussion of the p-value presumes that only large positive values of the test statistic may be significant, when in many experiments it is small or even large negative values that are significant. *Id.* at 402. The discussion of z- and t-statistics is turgid at best, and does not distinguish between the two in the case of small samples. *Id.* at 410-12.

182. Daniel L. Rubinfeld, *Reference on Multiple Regression, in* MANUAL, *supra* note 11, at 415-69. "A regression model attempts to combine the values of certain variables (the independent variables) to obtain expected values for another variable (the dependent variable)." Kaye & Freedman, *supra* note 43, at 404-05 (statistics chapter). Thus, where the issue is sex discrimination in salary at a defendant firm, a regression model might be used to predict salary (the dependent variable) on the basis of such independent variables as age, education, and seniority. If actual salaries for women fell substantially below predicted salaries, the regression model might be probative evidence for the plaintiffs.

183. Rubinfeld, *supra* note 182, at 423.

184. *Id.* at 423-28.

nificance of the regression results,[185] the robustness of the results,[186] the qualifications of the expert,[187] and the presentation of regression evidence.[188] An appendix on the more mathematical aspects of regression[189] and a glossary of terms[190] follow the main body of the *Guide*.

Regression is a rich and broad topic, and Rubinfeld has touched authoritatively on a remarkable number of issues within the confines of his chapter. Indeed, so extensive is this coverage that almost any issue that may arise in connection with a regression is at least mentioned. Thus, aside from its value as a source of explanation of regression concepts, the *Guide* can be used to confirm that an issue raised in litigation about a regression may have a material rather than incidental impact on its interpretation.

The *Regression Guide* makes no attempt to convey a sense of how regression works; or of how a set of fact circumstances could ever lead logically to a set of estimates for regression coefficients,[191] or to a forecast accompanied by a confidence interval.[192] However, it would be almost impossible to supply such an understanding to a reader who did not have a firm grasp on the connection between a random sample and a confidence interval, and as already noted, such a grasp is not likely to be gleaned from other parts of the *Manual*.[193] Thus, there is something of a gap in the discussion of regression, and though Rubinfeld's treatment is coherent and wide-ranging, it does not quite reach the foundation of the subject. Nevertheless, a careful reading of this *Guide* gives one a good sense of the kinds of things regression can do, and the kinds of things that may go wrong.

---

185. *Id.* at 429-32.
186. *Id.* at 432-38.
187. *Id.* at 439.
188. *Id.* at 441-43.
189. *Id.* at 445-62.
190. *Id.* at 463-68.
191. The regression coefficient "indicates the average change in the dependent variable associated with a one unit change in [a particular] independent variable." BARNES & CONLEY, *supra* note 1, at 408. Thus, in a regression model of salary, the regression coefficient for the independent variable "experience" would tell us how large an increase in salary was associated with the addition of one year's experience. *See supra* note 182.
192. In regression analysis, "the results can be used to predict a value of the dependent variable [salary, for example] that will occur given knowledge of the values of all the independent variables [age, education, and experience, for example]." BARNES & CONLEY, *supra* note 1, at 449. This predicted value can be expressed as a range, or confidence interval, within which it is likely to occur. *See supra* note 172.
193. *See supra* notes 170-77 and accompanying text.

If there is a weakness to the *Reference Guide on Regression*, it lies not so much within this *Guide* as in the absence of comparable neighboring guides. The inclusion in the *Manual* of a chapter devoted only to regression may suggest that it is the preferred framework for analysis of problems involving two or more variables. In fact, there are other approaches to the analysis of such problems,[194] and the choice of an approach should always be dictated by the fact situation. Regression imposes on a fact situation a framework to which that situation may not conform. For a variety of reasons, that incompatibility may materially distort the regression results.[195] While regression is a fascinating and widely used method of analysis, its importance as evidence in a particular legal proceeding must rest on its fit to the circumstances at hand, not its general popularity. The legal profession's tendency to appeal to case law authority for the use of a particular formula or method of analysis must, in the interest of better adjudication, give way to an appeal to a more general principle: that a statistical analysis be crafted to fit the fact circumstances, wherever that may lead.[196] Of two analyses, that which better fits the fact circumstances is the more reliable and thus should be the more credible, even though it may be based on more obscure formulas or models.[197] Once again, the post-*Daubert* court must be alert to the possibility that good science may not lead to regression

194. One approach is based on forming cohorts of individuals or objects that are similar with respect to certain qualities (such as education and experience), and then examining the correlation within each cohort between other qualities, such as salary and gender. The results of these individual examinations are then aggregated to produce an overall pattern of association between, for example, gender and salary. This approach is sometimes called cohort analysis, and the statistical tests it employs are variations on a Mantel-Haenszel test. For a detailed description of cohort analysis and some of its variations, see CONNOLLY & PETERSON, *supra* note 154, §§ 9.03[1], 9.03[2], and 9.03[3][b].

195. For a discussion of ways in which regression may rest on inapt assumptions, see *id.* app. E.

196. The Fourth Circuit's analysis of the data in Bazemore v. Friday, 751 F.2d 662 (4th Cir. 1984), discussed *supra* at notes 144-54 and accompanying text, illustrates what can happen when the statistical technique does not fit the facts. Had the case not settled, the district court presumably would have been saddled with the wrong analysis as the law of the case. Such problems of fit suggest a potential difficulty with Monahan and Walker's proposal to give precedential value to some social science studies. *See supra* note 140.

197. There is no simple, objective rule for judging quality of fit to the facts. For example, one cannot generally choose between two competing regression models based solely on their R squares. (R square describes the proportion of the variation in the dependent variable that is accounted for by all the independent variables included in a regression model. Thus, a model of salary with an R square of .35 would be accounting for 35% of the variation in salary, leaving the remaining variation unexplained. *See* BARNES & CONLEY, *supra* note 1, at 443-44.).

models, and so the court should not automatically presume that regression, because it has been used in other cases, is the most probative statistical framework.

B. Epidemiology and Toxicology

Each of these two guides begins with a concise definition of its subject discipline. Bailey, Gordis, and Green define epidemiology as "the field of public health that studies the *incidence*, distribution, and etiology of disease in human populations";[198] its purpose "is to better understand disease causation and to prevent disease in groups of individuals."[199] Toxicology, according to Goldstein and Henifin, is "the science of poisons," defined as " 'the study of the adverse effects of chemical agents on biological systems.' "[200]

Early on, each guide establishes the relationship between its subject discipline and the other. From the perspective of epidemiology, toxicology is applicable once a relationship has been observed between exposure to a suspect agent and an adverse human health effect. Controlled animal studies by toxicologists "often provide useful information about pathological mechanisms,"[201] while *in vitro* or test-tube studies can assess the effect of a toxic substance at the cell or tissue level, subject always to concerns about "whether one can generalize the findings from the tissues in laboratories to whole human beings."[202]

The *Toxicology Guide* states the complementarity of the two disciplines in more forceful terms: "These sciences often go hand-in-hand in assessing the risks of chemical exposure without artificial distinctions being drawn between the two fields."[203] It goes on to observe and lament the fact that "while courts generally rule epidemiological expert opinion admissible, admissibility of toxicological expert opinion has been more controversial because of uncertainties regarding extrapolation from animal and *in vitro* data to humans."[204] The Ninth Circuit's remand opinion in *Daubert* il-

198. Bailey et al., *supra* note 82, at 125.

199. *Id.*

200. Bernard D. Goldstein & Mary Sue Henifen, *Reference Guide on Toxicology*, *in* MANUAL, *supra* note 11, at 185 (quoting LOUIS J. CASARETT & JOHN DOULL, CASARETT AND DOULL'S TOXICOLOGY: THE BASIC SCIENCE OF POISONS 3 (Mary O. Amdor et al. eds., 4th ed. 1991)).

201. Bailey et al., *supra* note 82 at 130.

202. *Id.*

203. Goldstein & Henifen, *supra* note 200, at 194.

204. *Id.*

lustrates the point: It separated the two disciplines, attributed too much explanatory power to epidemiology, and too little to toxicology.[205] Goldstein and Henifin note that for a variety of reasons, including expense and ethical constraints, epidemiological studies of suspected toxic agents are few and far between.[206] The existing toxicological database on the same compounds, while still limited, is far larger, and new toxicological studies are cheaper and easier to carry out.[207] In addition, epidemiological and toxicological studies tend to be strikingly consistent.[208] Facts such as these prompt the inference—never explicitly advanced by the authors—that the courts should eschew their pejorative outlook on toxicology.

These introductory points, all well taken, ironically prompt our single major criticism of the two chapters: They would better have been written as a single guide on the broader topic of how scientists study the causes of disease. In simple terms, the two chapters argue convincingly that epidemiology and toxicology are two sides of the same coin. Epidemiology suggests that something is going on and toxicology shows how it could be happening. Courts, on the other hand, have tended to treat epidemiology as "real science" while dismissing toxicology as speculation.[209] By seemingly reaffirming the "artificial" division between the fields, the *Manual* may be unwittingly re-enforcing this bias.

This philosophical problem aside, each of the chapters is individually effective, with a few notable exceptions. The *Epidemiology Guide* begins with a useful introductory discussion of causation versus association, general versus specific causation, and sample size.[210] It is, in our view, the most effective discussion of these pervasive issues in the entire *Manual.* Fortunately, this is the first chapter. Nonetheless, one wonders again why there is no introductory chapter that deals with these and other overarching concepts.

---

205. *See* Daubert v. Merrell Dow Pharmaceuticals, Inc., 43 F.3d 1311, 1314, 1316-22 (9th Cir. 1995). The court implied that while animal and chemical studies are largely irrelevant to questions of specific, individual causation, epidemiological studies that meet certain statistical criteria can make a prima facie case of causation. *See supra* notes 101-03 and accompanying text. Epidemiology itself abjures the question of individual causation. *See* Bailey et al., *supra* note 82, at 167.

206. Goldstein & Henifen, *supra* note 200, at 194.

207. *Id.* at 194-95.

208. *Id.* at 195.

209. *See supra* note 205.

210. Bailey et al., *supra* note 82, at 125-28.

The authors next lay out three questions that typically frame the use of epidemiology in legal disputes:

1. Were the research methods trustworthy?
2. If so, is exposure to the agent associated with disease?
3. If the agent is associated with disease, is it a causal relationship?[211]

While these three questions indeed comprise a logical approach to the evaluation of epidemiological evidence, the authors do not address the question of how they square with the analysis required by *Daubert*. For example, are all three questions relevant to admissibility as defined by *Daubert*, or is the third directed at a study's conclusion, and thus not a part of the *Daubert* admissibility analysis?[212]

At this point, the authors move into the details of their subject, beginning with an excellent discussion of research design.[213] This, in our experience, is a very difficult topic for judges. They tend to interpret prudent conventions (for example, the requirement that an experiment test the effect of only one variable at a time) as flaws and to place impossible demands on researchers. When asked, Socratically, to design a study of a particular topic, judges regularly propose the impossible. Bailey, Gordis, and Green do as good a job as can be done by setting out with clarity what different research methods can and cannot accomplish.[214]

The succeeding discussion of sample selection and size issues is not nearly as successful.[215] It is too technical in style and phrasing.[216] The discussion of "power curves" is all too likely to confirm the popular conception that sampling is voodoo.[217] The discussion fails to make the important point that one can often draw meaningful conclusions from samples that may seem "small" from a common-sense perspective. There is also no effective explanation of how tests of statistical significance factor sample size into their calculations: As a general proposition, the smaller the sample size,

---

211. *Id.* at 128.

212. In its *Daubert* remand opinion, the Ninth Circuit dealt with all three under the Rule 702 admissibility rubric. *See supra* notes 101-03 and accompanying text.

213. Bailey et al., *supra* note 82, at 129-38.

214. This is another of those recurrent issues whose successful treatment here argues for an introductory chapter on the methods of science.

215. Bailey et al., *supra* note 82, at 138-43.

216. *See, e.g., id.* at 140-41 (discussing calculation of appropriate sample sizes).

217. *Id.* at 141-43. Power curves depict the probability that at a particular sample size a researcher will be able to detect an increased risk of disease of a particular magnitude.

the bigger the observed difference or relationship will have to be to meet a given threshold of statistical significance.

There follows a brief but lucid treatment of data collection problems.[218] The critique of scientific evidence, whether by judges or opposing experts, often jumps right to the level of results and interpretation. But, as Bailey, Gordis, and Green convincingly argue, the intelligent critic needs to look carefully at how the data to be relied on were collected. Moreover, the pertinent questions are largely matters of common sense that can be dealt with by nonspecialist judges.[219]

Next, under the heading of "Association Between Exposure and the Disease,"[220] the authors take on the vitally important topics of relative risk, odds ratio, and attributable risk. These are three related measures of the extent to which exposure to a suspected toxic agent increases a person's risk of a particular disease beyond the background level observed in the unexposed population. One or more of these measures is discussed in almost every toxic tort case; indeed, they are nearly talismanic.[221] The discussion of relative risk is superb—clearly written with simple but fully adequate examples.[222] The discussion of odds ratio is far less successful. The authors talk around the question of what the odds ratio is, but jump directly into the algebra without offering a clear prose definition.[223] The reaction of many judicial readers is likely to be frustration. Attributable risk (or APR, for attributable proportion of the risk) is described as "[p]erhaps the most useful measurement of risk."[224] If so, why does it come last, and in particular why does it come after a discussion of odds ratio that is likely to have confused the reader? The prose definition of APR does not connect well to the algebraic expression,

---

218. *Id.* at 143-46.

219. *See, e.g., id.* at 145 ("For example, a researcher may be interested in whether fetal malformation is caused by a mother's exposure to virus during pregnancy . . . . Mothers of the malformed infants may tend to recall inconsequential fevers or runny noses during pregnancy that readily would be forgotten by a mother who had a normal infant.")

220. *Id.* at 147-56.

221. Once again, the Ninth Circuit's remand opinion in *Daubert* is illustrative. *See supra* note 103; *see also* Bailey et al., *supra* note 82, at 168-69 (discussing relative risk cases).

222. Bailey et al., *supra* note 82, at 147-49.

223. *Id.* at 149. The odds ratio compares the odds of having a disease when exposed to the suspect agent to the odds of having the disease without exposure.

224. *Id.*

although a graph near the end of the section may clarify the issue for some readers.[225]

The discussion of the role that statistical significance and confidence intervals play in interpreting epidemiological data is the only truly disappointing part of the epidemiology *Reference Guide*.[226] Unlike most of the rest of the chapter, this section is unduly complex. There is no discussion of the mechanism that links significance testing with research design.[227] The important practical question of the relationship, if any, between statistical significance and the legal burden of proof is relegated to a lengthy textual footnote, which is dense to the point of impenetrability.[228] The fact that the confidence interval is the product of a calculation is not adequately explained.[229] This goes without saying among those with even rudimentary statistical knowledge, but it is our observation that judges and other statistical neophytes routinely treat confidence intervals as if they were merely assumed. Moreover, here as in the *Statistics Guide*, p-values and confidence intervals are treated as competing alternatives, when in fact they are complementary means of assessing whether an observed difference or relationship should be taken seriously or written off to chance.[230]

The succeeding, critically important section on inferring causation from association is, fortunately, particularly well done.[231] Bailey, Gordis, and Green effectively present the concept of confounds—factors other than the suspected agent that may actually be causing the disease under investigation.[232] They give an equally clear explanation of Koch's postulates, a set of long-accepted criteria to guide researchers in deciding whether to infer causation from an apparent association between the disease and the exposure of interest.[233]

The final section in the *Epidemiology Guide* deals with the very difficult question of the role that epidemiological evidence can play

---

225. *Id.* at 149-50. APR separates the proportion of the disease in the exposed population that can be associated with the suspect agent from the proportion that must be attributed to factors that affect both exposed and unexposed people.

226. *Id.* at 151-56.

227. For a discussion of this relationship, see *supra* notes 170-77 and accompanying text.

228. Bailey et al., *supra* note 82, at 153 n.80. Compare the discussion of this important concept in *Developments, supra* note 10, at 1548-56.

229. Bailey et al., *supra* note 82, at 154-55.

230. *Id.* at 153-55; *see supra* notes 178-79 and accompanying text.

231. Bailey et al., *supra* note 82, at 157-66.

232. *Id.* at 158-60.

233. *Id.* at 161-64.

in establishing individual causation.[234] This is, of course, the ultimate question in most cases involving epidemiological evidence. The bottom line, the authors candidly admit, is that "the cause of an individual's disease. . . . is beyond the domain of the science of epidemiology."[235] Nonetheless, they contend, epidemiological evidence alone can sometimes be sufficient to satisfy a plaintiff's burden of proof in a toxic tort case.[236] Relating the epidemiological concept of relative risk to the civil burden of proof, they explain why "[a] relative risk greater than 2.0 would permit an inference that an individual plaintiff's disease was more likely than not caused by the implicated agent."[237] This relationship, the authors observe, has been appropriately invoked in a number of cases; their brief but lucid discussion should ensure that others get it right as well.[238]

The *Epidemiology Guide* concludes with a glossary of terms.[239] Unlike its counterpart in the statistics chapter, this glossary is succinct and clear. Whereas the statistics glossary looks like the compromise work product of a committee of squabbling scientists, this one appears to have been written with the reader in mind. Without sacrificing accuracy, this glossary should enable judges to look things up and come away informed rather than frustrated.

The *Toxicology Guide* is at once the shortest and the clearest chapter in the entire *Manual.* Goldstein and Henifin begin with a clear definition of toxicology and an excellent discussion of the different kinds of research that toxicologists do.[240] They turn immediately to the aspect of toxicology that has troubled courts most: extrapolation.[241] Specifically, can one properly extrapolate from laboratory animals to humans, and from the very high doses of suspected poisons that are typically administered in laboratory research to the much lower doses that usually characterize human exposures to the same agents in the real world? Their response is both candid and coherent; appropriately, they stress that the most

234. *Id.* at 167-70.
235. *Id.* at 167.
236. *Id.* at 168-69. Their discussion properly deals with these issues under the heading of evidentiary sufficiency, not admissibility, in contrast to the approach taken by the Ninth Circuit in the *Daubert* remand, Daubert v. Merrell Dow Pharmaceuticals, Inc., 43 F.3d 1311, 1316-22 (9th Cir. 1995); *see supra* notes 101-03 and accompanying text.
237. Bailey et al., *supra* note 82, at 168-69.
238. *Id.* at 168-70.
239. *Id.* at 171-78.
240. Goldstein & Henifen, *supra* note 200, at 185-95.
241. *Id.* at 191-92.

reliable results come about when epidemiology and toxicology work in concert.[242]

Like their epidemiological colleagues, Goldstein and Henifin must also deal with the question of when a toxicologist can conclude that an agent caused a particular person's disease.[243] The general problem is the same as in epidemiology, although the criteria that guide the judgment differ in their specifics.[244] This *Guide* also concludes with a straightforward glossary that should be useful even to judges working with toxicological evidence for the first time.[245]

To summarize, the epidemiology and toxicology guides should prove extremely useful. With the few exceptions noted, they offer clear prose descriptions of the most difficult and important concepts in the two fields. Their most significant contribution is to explain the relationship between epidemiology and toxicology. Contrary to what the Ninth Circuit seemed to believe in considering *Daubert* on remand,[246] that relationship is complementary rather than hierarchical. Judges who read only the introductory pages of these two guides should, at a minimum, be able to improve on the Ninth Circuit's understanding.

C. *Conclusion*

In the *Reference Guides on Statistics*, *Multiple Regression*, *Epidemiology*, and *Toxicology*, as in the *Manual* as a whole, much is accomplished, but some significant opportunities are lost. Most of the recurrent significant issues in forensic science are addressed. Despite occasional errors and lapses into impenetrable jargon, the treatment is usually accurate and coherent. Readers are given a comprehensive overview of each of the included disciplines and, importantly, are provided with templates for identifying common forms of misapplication and overreaching.

Our principal criticism of the *Manual* has less to do with what it is than with what it fails to be. Judges and lawyers who encounter particular topics in particular cases will be able to find useful and generally authoritative references. What they may not get is a sense

---

242. *Id.* at 192, 194-95.

243. *Id.* at 205-08.

244. *Id.* The toxicological criteria include the fact and manner of exposure to the toxic compound, how the human metabolic system interacts with the compound, and the temporal relationship between exposure and the onset of disease. *Id.*

245. *Id.* at 213-17.

246. *See supra* notes 101-03 and accompanying text.

of the broader principles that unite the *Manual*'s constituent disciplines. Because the *Manual* lacks an interdisciplinary introduction to the scientific method, and because individual guides often fail to make important connections among topics,[247] scientifically inexperienced readers may be tempted to see science as a collection of discrete problems and solutions rather than as a network of interrelated concerns and approaches.

In response to this criticism, one might well ask, "So what?" If the *Manual* is at least improving the odds that judges will find accurate answers to individual questions, is it not then an unqualified good thing?

The problem, as we see it, has two dimensions. First, whatever the aspirations of its authors might have been, the *Manual* will inevitably become Talmudic. Judges will use it and cite it, incorporating its preferences into the case law, and lawyers will strive to tailor their scientific evidence to fit those preferences. Second, judges are lawyers trained in the common law: When they find an authority squarely on point, they follow it; when they do not, they analogize to the closest alternative. This mode of reasoning has worked tolerably well with cases for almost a thousand years, but it is often ill-suited to science. What is "on point" in the common-sense world of the common law is not necessarily what is on point in the scientific world; likewise, what are analogous categories to the lawyer may be apples and oranges to the scientist.

The danger, then, is that judges who learn a little bit about a number of ostensibly isolated topics—but who fail to appreciate the principles that tie those topics together—will try to extend their knowledge in inappropriate ways. And because they are judges, such extensions will become precedent. Should all this come to pass, the *Manual*, in a consummate irony, will have succeeded in exacerbating the very problem it was designed to correct.

We hope that we exaggerate the danger. The *Manual* began as a bold idea and its production was an undertaking of heroic proportions. We endorse its goals and appreciate the varied and significant accomplishments of its authors. Only time will tell if those accomplishments will be subverted by its subtle shortcomings.

---

247. See, for example, our discussion of the failure to connect research design with significance testing, *supra* notes 170-77 and accompanying text.