OF CHERRIES, FUDGE, AND ONIONS: SCIENCE AND ITS COURTROOM PERVERSION

DAVID W. PETERSON* AND JOHN M. CONLEY**

I

INTRODUCTION

The Supreme Court's decisions in *Daubert v. Merrell Dow Pharmaceuticals, Inc.,*[1] *General Electric Co. v. Joiner,*[2] and *Kumho Tire Co., Ltd. v. Carmichael*[3] (the "*Daubert* trilogy") redefine the duties of trial courts as arbiters of good scientific evidence.[4] Whereas under the venerable *Frye* test judges needed only to hear that the evidence was generally accepted in the relevant scientific community,[5] they are now thrust into an active "gatekeeping" role.[6] Above all, the *Daubert* trilogy requires courts to recognize the scientific method.[7] The *legal* test for reliability of scientific evidence is whether it conforms to the scientific method.[8] Courts must scrutinize purportedly scientific evidence for specific indicia of the scientific method; where these indicia are present, the evidence will

This article is also available at http://www.law.duke.edu/journals/64LCPPeterson.

* Ph.D., Senior Vice President, Peopleclick.

** William Rand Kenan Jr. Professor of Law, University of North Carolina, Chapel Hill.

The authors thank John A. Conley for research assistance on this article.

1. 509 U.S. 579 (1993).

2. 522 U.S. 136 (1997).

3. 526 U.S. 137 (1999).

4. The specific effect of the *Daubert* trilogy has been to clarify the standard for admitting expert testimony under Rule 702 of the Federal Rules of Evidence. In response to the trilogy, Rules 701 to 703 have been amended, effective December 1, 2000. *See* FED. R. EVID. 701-03. Under the new Rule, expert testimony on matters of scientific, technical, or other specialized knowledge will now be admissible "(1) if the testimony is based on sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case." FED. R. EVID. 702. The drafters of the new rule clearly believed that it was consistent with the overall approach of the *Daubert* trilogy, although they chose not to codify any of the specific *Daubert* reliability factors. *See* FED. R. EVID. 702 advisory committee's note. According to one member of the Advisory Committee, "the amendments accurately codify and clarify the [*Daubert* trilogy] tests for admission of scientific and technical evidence." Hugh B. Kaplan, *Panel of Scholars Looks at Daubert-Inspired Changes to Federal, Uniform Evidence Rules*, 69 U.S.L.W. 2285, 2286 (Nov. 14, 2000) (quoting Kenneth S. Broun). This article proceeds on the assumption that the amended rules of evidence will not change expert practice in any way that is material to our analysis.

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

6. *See Daubert*, 509 U.S. at 592.

7. *See id.* at 592-93.

8. *See id.*

be deemed reliable.[9] In other words, at least with respect to reliability, the law should accept what science should accept.[10]

Although its nominal focus is on admissibility alone, the *Daubert* trilogy also has profound implications for causation, in particular the relationship between scientific and legal causation. The purpose of the epistemological system that we call the scientific method is to provide rules for deciding when evidence counts—when it can be relied on to support a deduction about truth or an inference about causation.[11] The ultimate standards of truth and causation employed by science and law are very different, though both standards are couched in terms of probabilities. Science focuses on the specific probability of the chance occurrence of a particular result, while the law speaks vaguely, asking whether proffered accounts of truth and causation are "more probable than not" in a civil case, or are convincing "beyond a reasonable doubt" in a criminal case.

Despite these differences, the *Daubert* trilogy has linked the two standards inextricably, and the scientific method is the bridge. The very existence of the expert testimony controversy reflects the fact that, in many cases, a finding of scientific causation will be highly material to the question of legal causation. Indeed, in *Daubert* itself, a finding of legal causation could be based only on a claim of scientific causation.[12] In response to the question when purported evidence of scientific causation is sufficiently reliable to be admitted in support of legal causation, the Court answered: when it conforms to the scientific method.[13] Thus, if a claim of scientific causation appears to be based on scientific methodology, then it can be translated into evidence of legal causation.

The *Daubert* cases assume that trial judges will be able to discern the scientific method with reasonable accuracy. Taking their duties seriously, trial judges in *Daubert* jurisdictions have added a new term of art to the legal lexicon—the "*Daubert* hearing," a voir dire examination of an expert to investigate whether his or her methods were properly scientific.[14] The resulting case law suggests that these hearings are far from perfunctory, with much attempted separation of wheat from chaff.[15]

---

9. *See Daubert*, 509 U.S. at 592-94. Under *Kumho*, non-scientific expert testimony must meet a more general test of reliability. *See* Kumho Tire Co., Ltd. v. Carmichael, 526 U.S. 137, 149-50 (1999).

10. This is very different from saying that the law should accept what science *does* accept. That was the *Frye* standard. In some cases, people who call themselves scientists, for example, forensic "scientists," accept what they should not, whereas in other cases, science has not yet had time to accept things that it should and ultimately will. *See* 1 DAVID L. FAIGMAN, ET AL., MODERN SCIENTIFIC EVIDENCE 8-10 (1997).

11. *See generally* David Goodstein, *How Science Works, in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 67 (Federal Judicial Ctr. ed., 2d ed. 2000) (a physicist, writing for judges, compares the epistemological systems of law and science).

12. *See Daubert*, 509 U.S. at 582-84.

13. *See id.* at 592-94.

14. *See* William W. Schwarzer & Joe S. Cecil, *Management of Expert Evidence, in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE, *supra* note 11, at 53-54.

15. This article will deal primarily with statistical evidence. For examples of post-*Daubert* courts giving intense scrutiny to statistical expert testimony in disparate contexts, see *Wessman v. Gittens*, 160

The thesis of this article, however, is that *Daubert*'s focus on the scientific method, however rigorously applied, invites certain classes of abuses. Specifically, by moving courts from the *Frye* test's concrete focus on general acceptance to a more abstract inquiry, the new standard may have opened the door to evidentiary wolves in sheeps' clothing: claims that have the external manifestations of science, but in fact do violence to core tenets of the scientific method. Whereas *Frye* posed a fairly simple empirical question—do other experts say that this "science" is generally acceptable?[16]—*Daubert* asks whether the evidence has the attributes of science.[17] These attributes are supposed to be scrutinized in a sophisticated way.[18] But, as Joseph Sanders put it during the discussion at the editorial conference for this symposium, a direct if unintended effect of *Daubert* is that evidence that *looks* more scientific will more probably be deemed admissible.[19] There are instances in which evidence can be made to look more scientific by a process that in fact and substance makes it utterly unscientific. Indeed, the very process that creates the deceptively scientific evidence may conceal the deception from everyone but its orchestrator.

Part II of this article begins with a description of the main tenets of the scientific method, illustrated with reference to the ideal scientific experiment. Our example is a study conducted in the 1950s to determine the effectiveness of the Salk vaccine in inhibiting polio in children. Central to this discussion is a type of probability called a p-value, a measure of the extent to which a body of empirical evidence is consistent with a particular theory. Part III describes in more general terms the progression of a scientific inquiry, likening it to passage from the outer layers of an onion toward its center. Part IV identifies some of the compromises necessary when applying scientific methods.

With these constructs established, Part V discusses some of the ways in which the scientific method can be purposefully misapplied in civil litigation.[20] Such misapplications undermine the reliability of the method and ultimately distort scientific and legal judgments about causation. We describe how, through a process of "fudging" an investigation and "cherrypicking" results, one can produce scientific-looking evidence for presentation in court that is, in fact, not scientific at all. More troubling is that, by controlling the information and assignments given to various experts, a party can ensure that the testifying expert is unaware that the testimony he or she gives is unreliable. Most disturbing

---

F.3d 790, 802-05 (1st Cir. 1998) (school desegregation); *City of Tuscaloosa v. Hacros Chem., Inc.*, 158 F.3d 548, 562-67 (11th Cir. 1998) (price-fixing and bid-rigging); *Coward v. ADT Sec. Sys.*, 140 F.3d 271, 274-75 (D.C. Cir. 1998) (employment discrimination); *In re Executive Telecard Ltd. Sec. Litig.*, 979 F. Supp. 1079 (S.D.N.Y. 1997) (securities damages).

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

17. *See Daubert*, 509 U.S. at 592-94.

18. *Cf. id*. at 600-01 (Rehnquist, C.J., concurring in part and dissenting in part; questioning the ability of federal judges to function as "amateur scientists").

19. *See infra* note 112.

20. The Federal Rules of Evidence, as interpreted in the *Daubert* trilogy, apply to both civil and criminal cases. We limit our analysis to civil contexts, however, because that is where our experience and expertise lie.

of all, there does not appear to be any mechanism within the current rules of civil discovery ensuring that such abuses come to the attention of either the opposing party or the court.[21] Part VI proposes reforms to those rules.

II
A SCIENTIFIC IDEAL: THE DESIGNED EXPERIMENT

A. The Salk Vaccine Trials

As of the early 1950s, the poliomyelitis virus was a scourge of America's children; hundreds of thousands were afflicted, and many were disabled for life.[22] Among several vaccines under development, the one produced by Dr. Jonas Salk showed considerable promise.[23] The United States Public Health Service decided to conduct a very large-scale experiment to determine its effectiveness.[24] Ultimately, about two million school children were involved in the tests, though only a fraction of them actually received the vaccine.[25]

In designing the experiment, the Public Health Service attempted to take account of a variety of special considerations.[26] The first was the nature of the disease itself.[27] Polio is a hygiene-related disease: children who live in unhygienic environments are more likely to contract polio than those who live in cleaner environments.[28] Perversely, it is the latter children who are most severely affected when they do contract the disease;[29] the former are likely to be exposed early to the polio virus, so they tend to suffer only mildly and to develop an immunity to further harm.[30] Polio is also contagious, so that if one second-grader is infected, there is an increased chance that his or her classmates will be infected.[31] Furthermore, polio is epidemic, so the incidence is much greater in some years than in others.[32]

Second, there were ethical and technical considerations distinct from the nature of the disease. Clearly, one could not ethically require that any particular child be vaccinated with the experimental vaccine without permission from the child's parents or guardian.[33] But it is possible that children whose parents would grant permission would differ in some material and systematic ways from

---

21. Once again, criminal cases are beyond the scope of this article. In general, expert discovery is considerably narrower in criminal cases. *See* FED. R. CRIM. P. 16(a)(1)(E) & (b)(1)(C) (granting limited right to obtain written summary of expert testimony to be presented).

22. *See* DAVID FREEDMAN ET AL., STATISTICS 3 (2d ed. 1991).

23. *See id.*

24. *See id.*

25. *See id.* at 4.

26. *See id.*

27. *See id.*

28. *See id.*

29. *See id.*

30. *See id.*

31. *See id.* at 5.

32. *See id.* at 4.

33. *See id.*

children whose parents would withhold permission.[34] For example, it might be that well-educated and relatively affluent parents would tend to grant permission, while less well-educated and less affluent parents would not.[35] A potential result is that relatively many children living in hygienic circumstances would be permitted to take part in the study, and relatively few of those living in less hygienic conditions.[36] This imbalance could seriously skew the results of the study.[37]

There is also the problem that the behavior of the child or the parents might be influenced by the fact that the child had been vaccinated.[38] A child thus protected need not be quite as cautious in avoiding possible exposure to polio, and therefore might tend to engage in riskier behavior than his non-vaccinated neighbor. This too could seriously distort the study results. Moreover, since polio comes in both mild and severe forms, it is not always clear whether a child has contracted the virus.[39] As a result, a clinician examining a child who had been vaccinated might be less inclined to diagnose polio for that child than for her unvaccinated neighbor.[40]

Sorting through this web of considerations, the Public Health Service decided upon the following course of action. First, it selected schools across the nation where the incidence of polio was relatively high.[41] It then sought the permission of parents of first, second, and third graders for their children to be vaccinated as part of the study.[42] Half of the participating children were selected *at random* and injected with the Salk vaccine.[43] The other half of the subjects were injected with a placebo, a saline solution designed to have no medical effect whatsoever.[44] The children and guardians did not know whether a child received the vaccine or the placebo.[45] All of the children—both those who received the Salk vaccine and those who received the placebo—were monitored over the ensuing months by clinicians who also were not told which of the children had received which treatment.[46]

An experiment of this sort is termed randomized and double-blind.[47] It is randomized because the choice of whom to give the real vaccine is made by the toss of a coin or some other equally detached chance process.[48] This process vir-

---

34. *See id.*
35. *See id.*
36. *See id.*
37. *See id.*
38. *See id.* at 5.
39. *See id.*
40. *See id.*
41. *See id.* at 4.
42. *See id.* at 4-5.
43. *See id.* at 5.
44. *See id.*
45. *See id.*
46. *See id.*
47. *See id.*
48. *See id.*

tually guarantees that there will be no systematic difference between the group of children given the real vaccine and those who are given the false vaccine.[49] It is blind in the first instance because the children and their parents are unaware of whether they have actually been vaccinated.[50] Consequently, there is virtually no chance that the vaccinated children, as a group, will behave any differently from those who were given the false vaccine.[51] It is blind in the second instance because the people evaluating the health of the children do not know which of their subjects received the Salk vaccine and which received the placebo.[52] As a result, it is virtually certain that the same methods and standards for diagnosis will be used for the vaccinated group as for the placebo group.

In all, about 200,000 students received the Salk vaccine and about 200,000 received the placebo in this phase of the experiment.[53] The incidence of polio among the vaccinated group was approximately twenty-eight cases per 100,000, while that among the placebo group was about seventy-one cases per 100,000.[54] Given the randomized, double-blind construction of this experiment, there are only two possible explanations for these results.[55] The first is that the Salk vaccine really differed from the placebo in its effect on polio and that the difference was in the direction of reducing polio.[56] The second is that the Salk vaccine was no different from the placebo in its effect on polio, and that the observed reduction in the incidence in polio was due solely to the manner in which children were assigned to the treatment groups.[57] In other words, the children who contracted polio were destined to get it regardless of their treatment and the fact that most of them were placed in the group given the placebo was due purely to the luck of the coin toss.

B. The p-Value

Part of the genius of this experimental design is that one can calculate the probability that, if indeed the Salk vaccine were identical to the placebo in its effect on the polio virus, the coin toss mechanism would result in so many of the children predestined to contract polio being assigned to the placebo group.[58] That probability is about one in a billion.[59] Thus, it seems safe to say that the difference in the incidence of polio between the Salk group and the placebo group cannot reasonably be attributed to the random assignment of children to

49. *See id.*
50. *See id.*
51. *See id.*
52. *See id.*
53. *See id.* at 6.
54. *See id.*
55. *See id.*
56. *See id.*
57. *See id.*
58. *See id.*
59. *See id.* at 7.

treatment and placebo groups. The only remaining possibility is that the difference was caused by the greater effectiveness of the Salk vaccine.[60]

The above probability is called a p-value; it is a standard component of a report of the results of a scientific experiment. Assuming the truth of what is called the "null hypothesis," a p-value is the probability that evidence would arise that contradicts the null hypothesis at least as strongly as the evidence at hand contradicts it.[61] In the present case, the null hypothesis has two components: (1) that certain children were predestined to contract polio regardless of treatment, and the vaccine had no more effect than the placebo; and (2) that each child was assigned to either the Salk group or the placebo group based on the outcome of a random process. The evidence here is that the incidence of polio in the former group was twenty-eight per 100,000 and, in the latter, seventy-one per 100,000. The p-value is the probability, assuming no difference in efficacy between the placebo and the vaccine, that the random assignment process alone would produce a disparity in polio rates as great or greater than that which actually was observed.

A small p-value indicates that the observed data are inconsistent with the null hypothesis—the smaller the p-value, the greater the inconsistency. Hence, a small p-value is evidence that the null hypothesis is not true. If the experiment has been ideally constructed, rejection of this hypothesis will leave but one alternative. In the Salk case, the minuscule p-value (one in a billion)[62] would lead the researchers to reject the null hypothesis that the coin-toss assignment process—that is, chance alone—accounted for the disparity in polio rates between the two groups. Given the structure of the experiment, there is only one other plausible explanation: that the Salk vaccine was effective.

This is the logic of the ideal scientific experiment. The design features of random assignment and double-blinding virtually rule out the possibility of systematic differences between the experimental and control groups other than exposure to the suspected causal agent. When a very low p-value indicates that chance is too unlikely an explanation for an observed disparity in outcomes, the only alternative is to conclude that the agent's causal effect is real.

The problem is that few questions, scientific or otherwise, can be settled with the elegant finality of the Salk trials. This is particularly true in legal contexts, where the evidence is almost never so neat. In many instances, logistical or ethical barriers preclude a true experiment. Researchers are relegated to uncontrolled observational studies or after-the-fact data analysis. In all such cases, the focus on the suspected causal agent can never be as sharp as in the ideal, well-designed experiment.

---

60. This is because, given the random division of participants into the vaccine and placebo groups, there is no basis to suspect any systematic differences between the groups other than exposure or nonexposure to the vaccine.

61. *See* David H. Kaye & David A. Freedman, *Reference Guide on Statistics, in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE, *supra* note 14, at 83, 122.

62. *See supra* note 37 and accompanying text.

These situations have proved to be particularly troubling for courts under the *Daubert* regime. When the only scientific evidence of causation falls short of the Salk trials' gold standard, as it does in most cases, does it comport sufficiently with the scientific method to be admissible? That is, is it sufficiently reliable to be translated into legal causation? This is precisely the question that troubled the courts that heard the *Daubert* case itself as they wrestled with the issue of whether Bendectin could be reliably shown to cause birth defects.[63]

Our focus is somewhat different, however. In cases like *Daubert*, it is apparent to the court that the scientific evidence of causation falls short of the ideal, and the question becomes what to do with the second-best. In other cases, though, the evidence may appear far neater than it really is, in ways that are often invisible to the court and even to the adversary. Null hypotheses are assumed, ostensibly robust p-values are calculated, and causal agents—for example, discriminating employers or conspiring competitors—are scrutinized. The problem here is not what to do with inferior evidence, but rather how to discover that it is inferior. We turn next to an analysis of several such situations.

III

THE PROBLEM OF DISCERNING SCIENTIFIC PROGRESS: PEELING AN ONION

The first problematic category involves sorting out pieces of evidence that appear to be equally scientific, but in fact present quite different approximations of the truth in which a court is interested. Consider the historical record of a scientific inquiry that has not been driven by litigation. This record will be characterized by multiple analyses that seek the underlying truth in varying ways. Each analysis may, standing alone, be reasonable and thoroughly scientific. In hindsight, it will be easy to delineate progress and to identify false starts and blind alleys. But take away the historical perspective and consider all the different analyses simultaneously, without knowing in advance which will ultimately prove to be correct.[64] A vigorous advocate presents each analysis and attacks the others with equal vigor. The evaluation process is now infinitely harder. This is precisely the dilemma in which courts often find themselves as they try to create order out of competing hypotheses.

Consider the analogy of an onion. A scientific inquiry can be likened to peeling away the layers of an onion. At the center of the onion lies the truth being sought. Depending on where we are in the peeling process, we can examine and comprehend only some outer layers of the onion. Our knowledge of what lies further inside is merely inferential. With more investigation, we can

63. *See* Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579 (1993).

64. To take a famous example, if one reads the history of Michael Ventris's deciphering of the Linear B tablets from Homeric Greece, it is easy to see why his approach worked and those who came before him failed. *See* JOHN CHADWICK, THE DECIPHERMENT OF LINEAR B (2d ed. 1992). But imagine having to evaluate all the different approaches at once, without knowing in advance which one was right. This is the situation in which courts often find themselves.

peel away layers and progress toward the center. Realistically, however, there is little hope of ever clearly laying bare the very center. Our inferences about what lies there are based on what can be learned from some intermediate layer, and the closer that layer is to the center, the more reliable the inferences.

A scientific inquiry progresses toward the center of its onion by successive refinement. These refinements may involve greater measurement precision, as in the use of atomic clocks instead of stop watches, the use of micrometers instead of foot rulers, or the use of a nationally drawn sample instead of a few clinical cases. They may involve increased controls, such as accounting for differences in age, college, and major when comparing the salaries of men and women in the civilian labor force. They also may involve successively more elaborate and realistic theories, as when scientists try to explain phenomena relating the mass, energy, and motion of subatomic particles.[65]

The questions facing a judge often go well beyond whether a proffered piece of evidence is nominally scientific. Its attributes may be undeniably scientific, but what layer did it come from? Is that layer closer to the center of the onion than other competing evidence? Even if it is closer, is it still so far from the ultimate truth being sought that it is too unreliable or too marginally relevant to be admissible? And even if admissible, does it deserve any significant weight?[66]

Consider, for example, a situation in which an employer is accused of gender-based salary discrimination. Suppose that this employer's jobs do not carry fixed pay rates, and that among people holding the same job title, people are paid at various rates that may or may not be justified by differences in their seniority, proficiency, or dependability. Suppose further that the numbers of employees involved, their average pay rates, and their departmental and job title affiliations are as shown in Table 1.

---

65. *See* Goodstein, *supra* note 11, at 72-74.

66. For a sophisticated analysis of the admissibility versus the weight of complex statistical analysis, see *Bazemore v. Friday*, 478 U.S. 385, 400 (1986) (citation omitted):

> While the omission of variables from a regression analysis may render the analysis less probative than it otherwise might be, it can hardly be said, absent some other infirmity, that an analysis which accounts for the major factors 'must be considered unacceptable as evidence of discrimination.' Normally, failure to include variables will affect the analysis' probativeness, not its admissibility.

It is an interesting conceptual question whether evidence that passes the *Daubert* test of reliability and relevance can be held to be so insignificant that its proponent can still lose on summary judgment. *See* John M. Conley & David W. Peterson, *The Science of Gatekeeping: The Federal Judicial Center's New Reference Manual on Scientific Evidence*, 74 N.C. L. REV. 1183, 1199 & n.10 (1996).

TABLE 1:
WEEKLY PAY DATA

		Dept A		Dept B		Combined	
		Avg $	# Emps	Avg $	# Emps	Avg $	# Emps
Job 1	Male	360	5	432	5	396	10
	Female	360	10	432	5	384	15
Job 2	Male	540	5	648	5	594	10
	Female	540	20	648	10	576	30
Combined	Male	450	10	540	10	495	20
	Female	480	30	576	15	512	45

From the data in Table 1, we can make the following comparisons between the average compensation of men and women:

a) Overall, women are paid more than men.

b) Men in Job 1 are paid more than women in Job 1; Men in Job 2 are paid more than women in Job 2.

c) Women in Dept. A are paid more than men in Dept. A; Women in Dept. B are paid more than men in Dept. B.

d) Within each job and department, men are paid the same as women.

Analysis 1: The plaintiff begins with generalized allegations that there are gender disparities in compensation, and that these disparities are caused by employer discrimination. These allegations frame the issue at the center of the onion: whether the employer really is discriminating against women. In an initial response, the employer might point out that, on average, its forty-five female employees are paid $512 per week, which is more than the average pay ($495 per week) of its twenty male employees. This statistic is "scientific" in a primitive way: Its arithmetic is reliable, and averages are a staple of statistical investigation. Nor is it entirely irrelevant, because it says something about how the employer treats women. It does not strike very close to the heart of the onion, however, because there is little assurance that it compares females to similarly situated males. Indeed, it blurs job title distinctions, departmental af-

filiation distinctions, and whatever distinctions may exist with respect to seniority, proficiency, and dependability.

Analysis 2: The plaintiff might then point out that the ten males holding Job 1 are paid more, on average, than the fifteen females holding Job 1 ($396 versus $384) and that, likewise, the ten males holding Job 2 are paid more on average than the thirty females holding Job 2 ($594 versus $576). Clearly, this comes closer to comparing female workers with similarly situated males than does the comparison first offered by the employer. We have progressed toward the center of the onion in a way that requires no special scientific sophistication; it is the sort of judgment that courts routinely make in employment discrimination cases.[67] Given the enhanced nexus between the plaintiff's method of comparison and the ultimate issue, one must discard the employer's comparison in favor of the plaintiff's. It would be folly to credit both equally, and worse to accept that of the employer over that of the plaintiff.

Analysis 3: The employer might then parry the plaintiff's thrust with the observation that the thirty females in Department A are paid more, on average, than the ten males in that department ($480 versus $450) and, likewise, that the fifteen females in Department B are paid more on average than the ten males ($576 versus $540). This too is closer to the center of the onion than is the employer's first comparison, but it is not obviously closer than the plaintiff's comparison—it is merely different. Both plaintiff and defendant have now improved on the employer's first comparison, but neither of these new analyses is clearly more probative than the other. They lie at two separate points on the same layer of the onion.

Analysis 4: To repel the plaintiff's initiative, the employer must plunge deeper into the onion. It can do so by showing that within Department A, women in both Job 1 ($360) and Job 2 ($540) are paid exactly the same, on average, as are men, and that the same is true for each of the two jobs in Department B. This comparison is a refinement of that offered by the plaintiff in Analysis 3, because it is closer to the question of whether the employer pays similarly situated women and men the same. Most would probably conclude that, with this analysis, the defendant has trumped the plaintiff's analysis by striking closest to the ultimate issue with favorable results.[68]

---

67. For examples of courts making this sort of judgment, see *Ingram v. NWS, Inc.*, No. 92 C 8339, 1997 U.S. Dist. LEXIS 17190, at *27-30 (N.D. Ill. Oct. 21, 1997) (concluding that plaintiff's statistics were not based on sound methodology, because the statistician did not necessarily isolate age from related but non-discriminatory factors, such as seniority, as a deciding factor in employment decisions, and because the statistician did not exclude from the data pool those employees who did not meet objective minimum requirements); *Contractors Ass'n. of Eastern Pa., Inc. v. City of Philadelphia*, 893 F. Supp. 419, 427-31 (E.D. Pa. 1995) (finding that the plaintiffs' statistician's "disparity index" was not sufficient to prove that discrimination took place because, while it indicated disparities in minority hiring, it did not rule out "several neutral explanations for [the] statistical disparities," and his study was "methodologically flawed" in part because he did not make sure that the contractors in his data pool were all "qualified, willing, and able" to perform the available jobs).

68. This depends on the theory of discrimination being pursued. If, as we assume here, the allegation is that, among employees performing the same tasks, the employer pays women less than it does men at least in part because of their gender, then the defendant has moved closer to the center of the

This series of relatively simple analyses could vex a court on multiple levels. In light of the subsequent analyses, Analysis 1 is deceptively simplistic. But in some sense it is scientific, so do its drawbacks render it inadmissible? As illustrated above, Analysis 4 gets us significantly closer to the center of the onion than any of the other three; does its superiority make its competitors inadmissible? Keep in mind the consequences of admitting a flawed analysis. In a bench trial, these consequences are minimal, since the too-generous gatekeeper, now the fact-finder, can simply ignore that analysis in favor of a superior competitor. But if a jury ultimately credits an analysis that the gatekeeper-judge found marginally admissible though unimpressive, the judge's only recourse is to upset the verdict on the grounds that the analysis was (1) a reliable exemplar of the scientific method; (2) relevant to the issue at the core of the onion; but (3) for some other reason not good enough evidence on which to base a verdict.

As difficult as it might be for a court to make legal sense of the apparent scientific progression of the above analyses, an even greater problem is introduced by the consideration of analyses that have *not* been done or put before the court. The above analyses are, in reality, a sample of a potentially infinite set of analyses of the discrimination question. Recall that Analysis 1 began to look meaningless only in light of Analyses 2 and 3, and that they, in turn, seemed inadequate only in the reflected light of Analysis 4. But are there other possible analyses (5, 6, 7 . . . n) that would similarly reveal the inadequacies of Analysis 4? A litigant might never have thought of such analyses; she might have thought of them but left them undone for tactical reasons; or she might have conducted them but buried the results. Do *Daubert* gatekeepers need to worry about these contingencies, or should the contingencies even concern the legal system at large? We return to these issues at the end of the article.

IV

THE PROBLEM OF NO DESIGN

A related problem is that the data that become the subject of analysis in litigation—such as the analyses in the preceding section—are seldom the product of any research design. The ideally designed experiment is rarely achieved outside of a laboratory setting or a formal inquiry like that involving the Salk vaccine. In litigation, one must accept what is available. This messy reality greatly complicates the ability to recognize the scientific method and its appropriate application.

Consider, for example, a situation in which it is alleged that an employer, forced by an economic downturn to lay off workers, terminated older employees when choosing among otherwise similar individuals. The raw empirical evi-

---

onion than the plaintiff. If, however, the allegation is that the employer disadvantages women by assigning them to less well-paid "slots"—job or department combinations—than men, then the defendant's "progress" into the onion is nothing more than a sideslip along its surface. *See*, *e.g.*, Valentino v. U.S. Postal Serv., 674 F.2d 56, 69, 71 n.23 (D.C. Cir. 1982) (discussing statistical implications of discrimination in slotting employees into job groups).

dence consists of a list of employees as of the day before the force reduction, indicating their names, job titles, hire dates, birth dates, departmental affiliations, and an indication of whether that person was laid off the next day. These data were not generated pursuant to an experimental design; they are simply a more or less complete record of what happened. In using such data to determine whether the employer took account of age to the disadvantage of older employees, one usually attempts to identify all of the groups of employees who were similarly situated just prior to the layoff. "Similarly situated" is defined in terms of the information available about the individual employees that would be reasonable and appropriate to consider when making termination decisions—perhaps job title, department, and seniority, but not age. The employer might never have constructed such groupings, but our analysis will *assume* that it should have. Our objective is to determine what alternatives were reasonably available to the employer at the time it made its selections. If the employer consistently chose older employees for termination relative to the reasonably available alternatives, the inference is raised that it made impermissible use of age in making those selections.

More specifically, the analysis presumes that, within a group of similarly situated employees, all were or should have been considered equally acceptable alternative candidates for termination. If there are ten people in the group and one of them was in fact terminated, the presumption is that there were nine equally plausible alternatives to that choice.

With these assumptions in place, we can now determine whether the age pattern of the employees who were actually terminated deviated substantially from what might have been expected on the basis of chance alone. If older employees were terminated at a higher rate than their younger counterparts, but the discrepancy could be reasonably attributed to chance, one should not infer discrimination. Making this judgment requires the calculation of a p-value.

To calculate the p-value for the pattern of actual terminations, one assumes that the employer selected people for termination completely at random from each group of similarly situated employees. Thus, in the group of ten employees, one of whom was terminated, each member would have had one chance in ten of being terminated. Had there been two people terminated from that group, each member would be presumed to have had two chances in ten of being selected; had there been no terminations, all members would be considered to have been immune from selection. Building up from this premise and working group-by-group, one calculates an overall p-value for the actual pattern of terminations. This p-value measures the extent to which the actual pattern is consistent with the null hypothesis that selections were made randomly within each group of similarly situated employees.

In reality, employers rarely make termination decisions based on random selection; there is nearly always some marginal consideration that will break a tie between otherwise similar candidates for layoff. The consideration may be as subjective as a good work attitude or a pleasing personality, or as objective as

regularity of attendance—or, invidiously, age—but it is almost never the case that the final choice comes down to the toss of a coin. In contrast, the coin-toss mechanism was explicitly built into the Salk vaccine trials; it was a critical part of the experimental design, and the feature that permitted the calculation of a p-value.

In analyzing the employer's layoff record, we start with a hypothesis that is highly implausible, and then check to see whether the employer's record contradicts it. If a small p-value results, we conclude that the employer did not make selections at random from the groups of similarly situated employees, and that, therefore, some other process must have been used. Note that, unlike the Salk vaccine situation, many possible alternative processes remain. Perhaps the employer selected employees based on height; perhaps it was their personalities; perhaps it was their ages after all. The statistical analysis rules out only the implausible hypothesis that the employer, working with groups structured in the manner described, selected people for termination randomly from these groups.

In this application, we adopt the form of the designed experiment without its substance. Why does one have any interest in determining whether the age pattern of terminations is consistent with a hypothesis known at the outset to be highly implausible? First, in practice, there is a need to distinguish minor age disparities between those terminated and those retained from disparities that are so large as to suggest that the employer may have used age as a factor in terminating people. Second, this adaptation of the designed experiment produces *results* that are plausible, even though the underlying assumptions are not.

Suppose, for example, that calculations based on the implausible model yield the following results: (1) 24.3 people over age forty would have been terminated if the ages of people terminated had been perfectly representative of the ages of all those eligible for termination; and (2) between twenty and twenty-eight people over age forty would have been terminated if employees had been randomly selected from each peer group.[69] If the employer in fact terminated twenty-seven people over age forty, then there is no statistical indication here that older people were systematically singled out for termination. That is, even though the perfect parity figure of 24.3 may be erroneous because of errors made in forming the groups of "similarly situated" employees, the model provides for reasonable departures from parity, departures considered to be innocuous. The possibility of age discrimination becomes plausible only if the employer terminated a number of people over age forty that is significantly different from the model's perfect parity figure or, in this case, outside the range of twenty to twenty-eight.[70]

---

69. These results mean that the termination patterns for which the p-values are not too small (for example, not less than 0.05) are those for which the number of terminated employees over age 40 is in the range of 20 to 28.

70. It may seem odd that, having pointed out the implausibility of the statistical model, we cite as its redeeming feature the fact that it does not predict very exactly the number of people of each age who will be terminated. A more logical approach to this situation would begin by quantifying the de-

For all its imperfections, this approach leads to a qualitatively reasonable result, even though it is not a product of a designed experiment. There is a range of values within which the employer's record may fall without raising an inference of discrimination, because chance alone is a sufficient explanation. A record outside that range, however, demands a non-chance explanation. Because the "design" falls far short of the Salk model, we are not drawn inescapably to the suspect variable—age—but we may now at least entertain the possibility that age was a factor in the selection process.

Making sense of this chain of reasoning may be daunting to a judge charged with deciding if it adds up to "science." At first, that judge might be cautious, because the data seem so unsystematic, so far removed from the Salk model of collection. But the judge is then told that there is a way to construct an experiment of sorts, a decent approximation of the Salk model, directed at the question of whether the employer engaged in age discrimination. The very next piece of information, though, is that this quasi-experiment will test a hypothesis that is counterintuitive and almost certainly counterfactual. Why bother? Why is the law interested in a test of the self-evidently false proposition that the employer terminated people at random?

The answer is that, given the available information, this appears to be the only way to begin to shed scientific light on whether employer discrimination caused the observed pattern of terminations. This analysis will not prove that discrimination caused the pattern, but it may effectively rule out chance as a cause. In this sense, the analysis is quite modest. But by opening or closing the door to consideration of the universe of non-chance factors that includes age discrimination, it is highly relevant, indeed essential, to deciding the ultimate legal issue. Therefore, a judge should conclude that the analysis is scientifically reliable and relevant, but only with respect to a narrow, yet vital, question. As the preceding discussion should suggest, reaching this conclusion will test the subtlety and patience of even the most diligent gatekeeper.

V

PROBLEMS WITH HIDDEN EVIDENCE

A. Cherrypicking: When a p-Value Is Not a p-Value

Another problem involves what we think of as deceptive p-values: probabilities that have the formal attributes of p-values, but in fact misstate the significance of the results they appear to support. In a recent article written for an economics audience, Judge Richard Posner illustrated the problem by reference to what he termed "witness shopping":

---

gree to which the model itself is flawed, and then tracing the effects of these flaws. But there is no practical way of doing this in many instances, while the method suggested in the text is both practical and, within limits, plausible.

> Suppose that the lawyer for the plaintiff hired the first economist whom he interviewed and the lawyer for the defendant hired the 20th economist whom she interviewed. The inference is that the defendant's economic case is weaker than the plaintiff's. The parallel is to conducting 20 statistical tests of a hypothesis and reporting, as significant at the 5 percent level, the only one that supported the hypothesis.[71]

To take a more detailed look at Posner's problem, suppose that a drug manufacturer advertises that its new pain medication, R2-5, is more effective than aspirin in providing quick relief from minor aches and pains. Such a claim should, of course, be backed up by a scientific experiment proving that fact. One could design such an experiment along the lines of the Salk vaccine experiment by randomly dividing people into two groups, one that is treated with R2-5 and the other that is treated with aspirin. When the subjects have been evaluated and the results tallied, a p-value can be calculated based on the random assignment process.

Based more on tradition than rational principle, users of statistical reasoning generally regard p-values of less than .05 (or five percent) as "statistically significant," an indication that the observed data are sufficiently inconsistent with the hypothesis of chance occurrence as to require a search for some alternative explanation. Assume that, in fact, R2-5 is neither more nor less effective than aspirin. There is still a five percent probability that the experiment designed to compare its effectiveness with that of aspirin will produce a p-value less than .05. This means that even if R2-5's effects are identical to those of aspirin, there is a one-in-twenty chance that the experiment will indicate otherwise.

Suppose now that the drug manufacturer were to commission two independent studies of the effectiveness of R2-5 to be done by two different research organizations. There is a ninety-five percent chance that the first organization's study will show no significant difference between R2-5 and aspirin. Likewise there is a ninety-five percent chance that the second organization's study will show no difference. But there is only a ninety-five percent of ninety-five percent chance—that is, a 90.25% chance—that *both* organizations' studies will show no significant difference. Thus, there is a 100% - 90.25% = 9.75% chance that at least one of the two studies will indicate that R2-5 is more effective than aspirin, even though the two drugs are equally effective. If the studies are independent and neither of the two research organizations is aware of the work of the other, there is an approximately one-in-ten chance that at least one of them will conclude that R2-5 is more effective than aspirin. The drug manufacturer might then cite such a study in support of its advertising claim, purposefully omitting mention of any study reaching a contrary conclusion.

The odds that this scheme will succeed can be vastly improved by hiring ten independent organizations to do ten independent studies. Under these conditions, chances rise to about four in ten that at least one will indicate that R2-5 is significantly more effective than aspirin, even though there is no difference at

71. Richard A. Posner, *The Law and Economics of the Economic Expert Witness,* 13 J. ECON. PERSP. 91, 98 (1999).

all.[72] If the drug manufacturer bases its advertising campaign on just the one report among ten that was most favorable to it, the campaign overstates its case.[73]

The p-value associated with that one report does not take account of the other nine studies, or the fact that the one study selected for publication was selected precisely because it was the most favorable. The organization producing this particular study is blameless, assuming it has no knowledge of the studies of its sister institutions. It might have conducted an entirely proper and objective study, and there is no reason it should not be willing to stand by its results, even to the point of providing sworn testimony. Each of the other nine research organizations might be surprised to learn that R2-5 has been shown to be more effective than aspirin, but each will realize that it apparently does not know all the results of all tests that were performed. Upon learning of the contrary result, each organization will probably conclude that, if it had done its studies differently or with more subjects, it too would have been able to detect at least some differences, however slight, in the effects of R2-5 and aspirin. In sum, the only party who may fully appreciate the knavery of this exercise may be its orchestrator. To all other parties, it will appear that R2-5 is a scientifically confirmed improvement on aspirin.

Apply this same type of thinking to an employer who has employees in each of the following nine job categories: (1) managers and administrators, (2) professionals, (3) salespersons, (4) technicians, (5) clerical workers, (6) craft workers, (7) operatives, (8) laborers, and (9) service workers.[74] Suppose that, within each category, the employer monitors the promotion rates of men versus women and minorities versus non-minorities and expresses the differences in promotion rates as p-values. Thus, there are nine p-values reflecting promotion rate differences by gender, and nine others reflecting promotion rate differences by race. Recall that a non-discriminating employer will produce a p-value less than or equal to five percent about five percent of the time. There are many chances here for at least one of the eighteen p-values to be less than five percent and, hence, to be considered statistically significant. Indeed, if the eighteen p-values are independent of one another—which may or may not be true, depending on how closely employee race and gender are correlated—there is only a 39.7% probability that all eighteen p-values will be greater than five percent. Therefore, it is more likely than not that a non-discriminating employer will have at least one job category in which there appears to be a statistically significant disparity in promotion rates by race or gender.[75]

---

72. The probability that all ten organizations' studies will be insignificant at the five percent level is 0.95 raised to the tenth power, or 0.599. Hence, the probability that at least one of the organizations will produce a study demonstrating that R2-5 is more effective than aspirin is 0.401, or about four chances in ten.

73. This is true regardless of whether R2-5 is more effective than aspirin.

74. These job categories are commonly used by the U.S. Equal Employment Opportunity Commission in auditing the employment practices of employers.

75. Since, by definition, this sort of cherrypicking is difficult to detect, reported instances in the case law are rare. Nonetheless, the example in the text is generally analogous to the age discrimination case of *Ingram v. NWS, Inc.*, No. 92 C 8339, 1997 U.S. Dist. LEXIS 17190 (N.D. Ill. Oct. 21, 1997),

Suppose now that a federal agency focuses solely on the most egregious of these eighteen disparities and brings to a court's attention the fact that, among, say, craft workers, women received statistically significantly less than their proportionate share of promotions last year. The p-value cited by the agency for this group is not the true p-value, because it does not take into account the fact that the job category was selected precisely because it has the largest disparity. A proper characterization of the evidence that this employer discriminated would take into account not only the particular p-value cited by the agency, but also all of the others it reviewed.

The above examples involve picking from among a collection of more or less independent comparisons that one that best comports with one's views. We call this behavior "cherrypicking." It involves surveying several bodies of data and choosing to credit only those which produce the "right" answer for no reason other than the fact that they do so.

B. Fudging the Analysis: Another Path Around the Onion

A variation on "cherrypicking" is called "fudging the analysis."[76] "Fudging the analysis," however, involves applying a variety of different analyses to the same, fixed body of data. The result of each analysis is deemed credible based on the extent to which it is favorable to the party's views. In this process, little or no thought is given to progressing toward the center of the onion.

Consider the case of an employer who laid off twenty employees in a force reduction. The plaintiff, age fifty-four at the time, compiles statistics showing that a disproportionate number of employees older than fifty-one were among those laid off. Defendant employer counters with the observation that the layoff rate of people over the age of forty-two is identical to that of people under age forty-two. If the plaintiff has run his calculations using age cutoffs of forty, forty-one, forty-two, and so forth, and then reported only the result most favorable to him, then his report distorts the evidence, unless he discloses the fact that other cutoffs were examined and found to be inferior, or unless he adjusts his reported p-value to take full account of this fact. The same principle applies as well to the defendant's calculations. [77]

---

where the court criticized an expert statistician for the exclusion of certain job group data that would have tended to undercut his party's position. *See id.* at *32. The court required modifications to the expert's analysis before it would allow his testimony. *See id.* at *40. In another age discrimination case, *Adams v. Indiana Bell Telephone Co.*, 2 F. Supp. 2d 1077, 1100-03 (S.D. Ind. 1998), the court made similar criticisms of a statistician, noting as well that his data selection had been influenced by counsel. *See also* Garrett v. Kenmore Mercy Hosp., 1998 U.S. Dist. LEXIS 2132, at *17-23 (W.D.N.Y. 1998) (deeming statistical study inadmissible in part because expert selected data in reliance on client's direction); Contractors Ass'n of E. Pa., Inc. v. City of Philadelphia, 893 F. Supp. 419, 431-33 (E.D. Pa. 1995) (same).

76. The term "fudging" in our experience usually is applied to instances in which an analyst alters empirical data with the purpose of causing them to produce a desired result when viewed through the lens of a particular method of analysis. Here, we use the term to denote the selection of a method for analysis that is driven by consideration of the result it produces.

77. As in the case of cherrypicking, fudging is rarely condemned in the reported cases because, by definition, it is hard to detect. A few courts, however, have displayed sensitivity to the general prob-

As a less transparent example, consider the employer alleged to pay its female employees less than its male employees. A common way of exploring such allegations, after doing some simple average pay comparisons, is through the use of regression analysis. Regressions constitute a large class of models in most circumstances of this type, so the analyst has great latitude in selecting the particular one or ones on which to rely. An analyst can, in principle, run hundreds or even thousands of different models using the same underlying data in search of that one model most flattering to his client.[78]

Virtually all regression models are computed using general purpose, commercially available software. Each computation generates a standard set of p-values that describe various aspects of the application of that model to the data at hand, but none of these p-values depends in any way on or takes any account of the other regression models that may have been run and examined by this analyst using the same data. In effect, the p-values produced by standard regression software are all based on the presumption that this is the first time the analyst has examined these data using any model. Consequently, when the analyst runs ten different regression models on the same database and then relies on the one that produces the most helpful p-value, that p-value does not reliably perform its function of reflecting the likelihood of the chance occurrence of comparable results.[79]

The first of the above two examples (involving the self-serving choice of an age cut-off) is so transparent that it is likely that both the litigants and the trier of fact will be cognizant of the inappropriateness of both p-values. As a result, the trier of fact will discount them both and deal with the case on other grounds. The second of the two examples, however, is not so easily shrugged off. The clever attorney can hire one analyst to run dozens or perhaps even hundreds of regressions in search of the one with the most acceptable p-values, and then hire a fresh analyst for the sole purpose of "discovering" that one regression.[80] If all goes well, the new analyst, unaware of the previous exploratory

---

lem. The hypothetical in the text resembles the facts of *Adams v. Indiana Bell Telephone Co*., 2 F. Supp. 2d 1077, 1100-03 (S.D. Ind. 1998), where a statistician both selected data and shaped his analysis of alleged age discrimination in a way that struck the court as result-driven. *See also* Munoz v. Orr, 200 F.3d 291, 301 (5th Cir. 2000) (holding the statistician's evidence inadmissible because he "began his analysis with the assumption that [defendant's] promotion system discriminated against Hispanic males," and then selectively excluded variables and modes of analysis); Allard v. Indiana Bell Tel. Co., 1 F. Supp. 2d 898, 905-07 (S.D. Ind. 1998) (companion case to *Adams*; criticizing the *Adams* expert for allowing the client to influence the framing of research questions and analysis).

78. This problem was identified a generation ago. *See* Michael O. Finkelstein, *Regression Models in Administrative Proceedings*, 86 HARV. L. REV. 1442, 1449 n.27 (1973). Finkelstein described the problem as "seldom discussed," a term that is equally apt today. He proposed a set of four "protocols" to address a number of latent problems with regression models of economic data. To our knowledge, however, his protocols have never been cited by a federal court. *See infra* notes 102-103 and accompanying text.

79. For a discussion of the application of a series of regression models to the same data, see EDWARD E. LEAMER, SPECIFICATION SEARCHES: AD HOC INFERENCE WITH NONEXPERIMENTAL DATA 87-120 (1978).

80. The new analyst might be instructed as follows: "Dr. Expert, we think that starting pay properly depends on level of education and years of prior experience, and not gender. We have prepared

work and given a narrowly tailored task, produces the desired regression, or one much like it, on the first try. He can then testify truthfully that, to him, the p-values for that regression are exactly what they purport to be.

C. Finding the Hidden Analyses: The Inadequacy of Current Discovery Rules

The foregoing "cherrypicking" and "fudging" examples are variations of what Posner calls "witness shopping."[81] If the scheme is properly managed, no single analyst has necessarily done anything wrong or, more to the point, unscientific. Indeed, one of the examples of cherrypicking, the R2-5 analysis, appears to meet the Salk vaccine gold standard. Likewise, each of the other examples reports an ostensibly respectable p-value. Thus, from the *Daubert* perspective, there seems to be no basis to exclude any of them.

Yet when the full story is known, each analysis taken alone gives a distorted answer to the ultimate question of legal interest: respectively, whether R2-5 is more effective than aspirin, or whether the employer discriminated. These dilemmas are more difficult to sort out than those discussed in Parts II-IV of the paper. They share with those earlier problems the requirement of a sophisticated understanding of the scientific method and the meaning of p-values, but add the problem of hidden evidence. This is because, under current law, it is far from certain that either an adversary or the court has any means to discover the sorts of deceptions we have described.[82]

As Posner's complaint implies, there is no definitive requirement that a party disclose the identities of all the experts it has consulted.[83] The basic expert discovery rule, Federal Rule of Civil Procedure ("Rule") 26(b)(4), as amended in 1993, creates two categories of experts. In the first category, "any person who has been identified as an expert whose opinions may be presented at trial"

---

this database containing that sort of information. Could you please have a look at this and let us know what you think? We very much need your help, and have been awed by your prior work in this area . . . ." *See Allard*, 2 F. Supp. 2d at 1101 (stating that the expert "explained his procedures by stating that he used information provided by the plaintiffs' counsel about the characteristics of the workforce he studied").

81. *See* Posner, *supra* note 71.

82. The analysis that follows in the text demonstrates the lack of any discovery right that could be used to uncover such deceptions. Some lawyers who have commented on this paper, however, have suggested that there often may be practical opportunities for detection. For example, in at least some cases, lawyers for the opposing parties may agree to full disclosure of all experts, testifying and non-testifying. *See* FED. R. CIV. P. 26(f) (requiring parties to meet to discuss discovery matters). Trial judges also have considerable authority to customize discovery procedures in individual cases under FED. R. CIV. P. 16(b), (c) & (e). Local rules sometimes underscore the court's authority, *see, e.g.,* E.D.N.C. R. 23.07(c) (involving the court in discovery planning), but these rules usually simply restate the rights and duties created by the Federal Rules of Civil Procedure. *See, e.g.*, D.S.C. R. 26(b)(4)(A) (requiring disclosure of testifying experts); E.D. Va. R. 26(D) (incorporating the mandatory disclosure obligations of Federal Rules). Our own experience, confirmed by informal discussions with lawyers throughout the country, is that discovery of non-testifying experts is very rarely agreed to, and even more rarely ordered. The cases in which it does occur are almost always large and closely managed ones.

83. *See* Stephen D. Easton, *Ammunition for the Shoot-Out With the Hired Gun's Hired Gun: A Proposal for Full Expert Witness Disclosure*, 32 ARIZ. ST. L. J. 465 (2000).

can be deposed as a matter of right.[84] Depending on when a lawyer is called upon to decide which experts may be presented at trial, it is possible to shield the "wrong answer" experts from deposition. If, for example, the demand for depositions of testifying experts is made after multiple analyses have been completed, the responding lawyer will already have taken those who produced unhelpful results out of the testifying category and will not need to produce them.

In the second category is any "expert who has been retained or specially employed by another party in anticipation of litigation or preparation for trial."[85] An expert in this category can be subject to interrogatories or a deposition, but "only as provided in Rule 35(b) [which deals with court-ordered physical and mental examinations] or upon a showing of exceptional circumstances under which it is impracticable for the party seeking discovery to obtain facts or opinions on the same subject by other means."[86] Note that the latter provision defines "exceptional circumstances" in terms of the impracticability of obtaining other opinions *on the same subject*;[87] it makes no reference to the adversary's need to examine alternative studies. With the possible exception of the R2-5 example, in each of the hypothetical cases described in this section, it is entirely practicable for the opposing party to obtain opinions *on the same subject.*[88] Thus, Rule 26(b)(4) does not provide access to hidden and obviously non-testifying experts—nor even a duty to disclose their identities.[89]

A related question is whether the deposition of the testifying expert—the one who found the "right" p-value—would inevitably lead to the discovery of the hidden analyses. This discovery would not occur, because the client or lawyer hires separate analysts until the desired result is reached, and then walls the analysts off from each other.

Another possible source of discovery of the hidden analyses is the mandatory disclosure provisions that were added to Rule 26(a) in 1993.[90] At least

---

84. FED. R. CIV. P. 26(b)(4)(A).

85. *Id.* 26(b)(4)(B). The apparent purpose of this special retention language is to distinguish an expert witness from a fact witness, such as a treating physician who also happens to be an expert. *See Id.* 26(b)(4) advisory committee's notes; *cf.* Patel v. Gayes, 984 F.2d 214, 218 (7th Cir. 1993) (holding that a physician is a mere fact witness if his opinions are based solely on treatment, but becomes an expert if opinions are also based on other sources); Shapardon v. West Beach Estates, 172 F.R.D. 415, 417 (D. Haw. 1997) (same).

86. FED. R. CIV. P. 26(b)(4)(B).

87. *See*, *e.g.*, Queen's Univ. at Kingston v. Kinedyne Corp., 161 F.R.D. 443, 447-48 (D. Kan. 1995); Santos v. Rando Mach. Corp., 151 F.R.D. 19, 21-22 (D.R.I. 1993).

88. Assume, for example, false advertising litigation over the claim that R2-5 was more effective than aspirin. It might be quite impracticable for the opposing party—the FTC, perhaps—to conduct the studies necessary to obtain opinions on the subject. Here, conceivably, a court might allow discovery of the conclusions reached by the non-testifying experts. In each of the other cases, however, there is no reason why an opposing party could not obtain opinions on the questions addressed in the "hidden" studies.

89. *See Queen's Univ.*, 161 F.R.D. at 447; Ager v. Jane Stormont Hosp., 622 F.2d 496, 500-01 (10th Cir. 1980).

90. As a preliminary matter, note that about half the district courts opted out of the mandatory disclosure rule, but the right to opt out has been abrogated by the 2000 amendment to the Federal Rules.

ninety days before trial,[91] each party must disclose "the identity of any person who may be used at trial to present evidence under Rules 702, 703, or 705 of the Federal Rules of Evidence [the expert rules]."[92] In the case of an expert who is retained or specially employed, or an employee of a party who regularly gives expert testimony, the disclosure must also include a written report that contains: (1) a complete statement of opinions; (2) the reasons for such opinions; (3) "the data or other information considered by the witness in forming the opinions"; (4) the exhibits the witness will use; and (5) the witness's qualifications and compensation.[93]

Even this apparently all-encompassing disclosure, however, is unlikely to capture the hidden analyses we have described.[94] The party need only disclose those experts who are expected to testify. Unless the court overrides the ninety-day rule and orders the mandatory disclosures very early in the case, the "wrong" analyses will have been identified and buried well before Rule 26(a)(2) ever comes into play. Moreover, the requirement that the expert's report include a complete statement of opinions, reasons, and underlying information is also likely to be unavailing, since the expert who survives to testify will have no knowledge of those who were discarded along the way.[95]

---

91. The 90-day rule may be modified by order of court or stipulation. *See* FED. R. CIV. P. 26(a)(2)(c). In addition, expert evidence intended solely to contradict or rebut the other side's evidence need not be disclosed until 30 days after the opposition's disclosure. *See id.*

92. *Id.* 26(a)(2)(A).

93. *Id.* 26(a)(2)(B); *see also* Schwarzer & Cecil, *supra* note 14, at 49-50 (reviewing mandatory expert disclosure).

94. THE MANUAL FOR COMPLEX LITIGATION, THIRD (1995) ("MCL"), advocates that courts order even more extensive mandatory expert discovery early in employment discrimination cases, including an exchange of statistical databases. *See id.* § 33.53. However, nothing recommended in the MCL addresses the specific problems we have identified. *Cf. id.* § 21.48 (discussing the limited nature of discovery available against "consulting" but non-testifying experts).

95. Logically, the work of discarded experts would fit within the work product doctrine. Originating in the Supreme Court's 1947 decision in *Hickman v. Taylor,* 329 U.S. 495 (1947), and codified since 1970 in Rule 26(b)(3), the doctrine confers presumptive immunity on "documents and tangible things . . . [otherwise discoverable] prepared in anticipation of litigation or for trial by or for another party [or a party's lawyer]." FED. R. CIV. P. 26(b)(3). This immunity can be overcome only upon a showing "that the party seeking discovery has a substantial need of the materials in the preparation of the party's case and that the party is unable without undue hardship to obtain the substantial equivalent of the materials by other means." *Id.* If the court orders disclosure, it must "protect against disclosure of the mental impressions, conclusions, opinions, or legal theories of an attorney or other representative." *Id.* There have been a number of inconsistent decisions on the question of whether *attorney* work product that is shown to an expert is discoverable. *See* Easton, *supra* note 83, at 528-36. However, with respect to an expert's own work, it is clear from the preamble of Rule 26(b)(3) ("Subject to the provisions of subdivision (b)(4) . . ."), the 1970 Advisory Committee notes to Rule 26(b)(3) (characterizing as "ill-considered" the cases that sought to bring experts within the work product doctrine), and the case law that Rule 26(b)(4) is the exclusive vehicle for discovery. *See e.g.,* Hewlett-Packard Co. v. Bausch & Lomb, Inc., 116 F.R.D. 533, 536-37 (N.D. Cal. 1987); Beverage Mktg. Corp. v. Ogilvy & Mather Direct Response, Inc., 563 F. Supp. 1013, 1014 (S.D.N.Y. 1983).

VI

CONCLUSION: A PROPOSAL FOR REFORM

Three points should now be clear. First, the subtleties of the scientific method in general, and statistical inference in particular, can make discernment of the scientific method difficult, even in cases of full disclosure of all available analyses. Second, the process of analysis shopping can result in a deceptive analysis being presented by an honest and scientifically scrupulous expert witness. Third, the current rules of discovery facilitate the concealment of many such deceptions.

While the law can do little, if anything, to make the scientific method more transparent, it can take some fairly modest steps to ensure that the parties and the gatekeeping trial judge have all of the relevant information in front of them. In our judgment, there is only one way to avoid the kinds of abuses of scientific evidence that we have been discussing: All scientific work commissioned by any party must be fully discoverable, regardless of whether it is used at trial. Some may argue that this approach undercuts the fundamental premises of the adversary system—"that each side's informal evaluation of its case should be protected, that each side should be encouraged to prepare independently, and that one side should not automatically have the benefit of the detailed preparatory work of the other side."[96] But the adversary principle has never been absolute and has often yielded to concerns about justice.

Expert discovery is itself a compromise. Discovery of experts as of right was not available until 1970, when Rule 26(b)(4) was amended to allow an interrogatory seeking the identities of testifying experts and the substance and bases of their opinions.[97] The change was made in response to the concern that "a prohibition against discovery of information held by expert witnesses produces in acute form the very evils that discovery has been created to prevent," especially shooting-in-the-dark cross-examination.[98] The rule remains a work in progress. By 1993, it had become clear that the "expert interrogatory" allowed by the 1970 amendment was inadequate to the task, and the right to depose testifying experts was added, "conforming the norm stated in the rule to the actual practice followed in most courts."[99] The advent of mandatory disclosure of testifying expert reports and other matters, also in 1993, came in the face of vigorous opposition from adversary purists.[100] As in the case of previous changes, the objections were overcome by the desire to make litigation more efficient, more rational, and less expensive.[101] In the present situation, the adversary principle

96. FED. R. CIV. P. 26(b)(3) advisory committee's notes to 1970 amend.

97. *See id*. 26(b)(4) advisory committee's notes.

98. *Id.*

99. *Id*. 26(b) advisory committee's notes.

100. *See, e.g.,* Supreme Court of the United States, Amendments to the Federal Rules of Civil Procedure, 146 F.R.D. 507, 510-12 (1993) (Scalia, J., dissenting from Court's promulgation of mandatory disclosure rules).

101. *See* FED. R. CIV. P. 26(a) advisory committee's notes.

should yield once again—this time to the very real threat of deceptive "science" slipping through the *Daubert* net.

As this brief history suggests, demands for reform in expert disclosure are nothing new. In addition to the incremental work of the Advisory Committee on Civil Rules, legal academics have regularly proposed more fundamental changes. In 1973, Michael Finkelstein proposed a set of procedural reforms that he called "protocols."[102] These reforms were directed specifically at the use of statistical evidence in administrative proceedings. He suggested, for example, that the decisionmaker make a preliminary determination of the data that should be analyzed and that a party objecting to an opponent's statistical model quantify the objection and advance a superior alternative model. While Finkelstein's article talks specifically about ratemaking proceedings and econometric regression models, his protocols are readily adaptable to statistical analyses in a wide variety of cases. Requiring the decisionmaker to specify the relevant data would begin to address some of the problems we have raised. Yet we have been unable to find a single case that has adopted any part of his recommendations.[103]

More recently, in 1991, Samuel Gross wrote a lengthy and thoughtful commentary on the state of expert evidence.[104] According to Gross, the need for special procedural treatment of experts derives from the fact that "expert information is categorically different from other types of information that we use in litigation, and that the functions of expert witnesses are fundamentally different from those of other people who provide information in court."[105] Gross proposed "minor changes" that trial judges already had the authority to impose. Among these "minor" changes were presenting all expert testimony at the same time; permitting experts to question each other; requiring that experts exchange reports and meet face-to-face for an "unmediated discussion" before trial; and ordering opposing experts to produce a joint report that defines key terms, lists points of agreement and disagreement, and jointly addresses such issues as the identification of qualified non-partisan experts.[106] Although the pre-trial exchange of expert reports has become part of Rule 26(a)'s mandatory disclosures, Gross's other suggestions, like Finkelstein's, have been ignored by the courts.[107]

---

102. *See* Finkelstein, *supra* note 78.

103. Although it does not cite his work, the MANUAL FOR COMPLEX LITIGATION adopts the general thrust of Finkelstein's protocols in its recommendations for dealing with complex statistical evidence. *See supra* note 78.

104. *See* Samuel R. Gross, *Expert Evidence*, 1991 WIS. L. REV. 1113 (1991).

105. *Id.* at 1208.

106. *See id.* at 1211-13. Gross also advocated a number of other changes that he characterized as "major," including the elimination of oral testimony by experts, the elimination of juries, and the restriction of expert testimony to court-appointed neutrals. *See id.* at 1213-30.

107. A LEXIS search of both state and federal databases revealed a single citation to Gross's article, on the ancillary point of how hard it is to cross-examine experts. *See* State v. Porter, 698 A.2d 739, 748 (Conn. 1997).

This past summer, Stephen Easton published yet another exhaustive review of the problems of expert discovery and testimony.[108] He, too, offered a set of proposed reforms. The centerpiece of his proposal was a substantial addition to the mandatory disclosure provisions of Rule 26(a).[109] Specifically, he argued that experts retained to give testimony should be required to disclose all documents, data compilations, and tangible things furnished to them by a party or a lawyer, and the contents of all communications between them and a party or a lawyer.[110] Easton's amendment would certainly, as he suggests, make cross-examination better informed.[111] Furthermore, it would address one of the problems we have identified, that of counsel giving leading instructions to the expert. It would not, however, assist counsel and the courts in uncovering questions that were never asked, analyses that were left undone, or unfavorable results from now non-testifying experts who had been dispatched to the evidentiary graveyard. It is too early to tell whether Easton's proposal will commend itself to the courts and the rules authorities.

Despite the lack of response to these earlier efforts, we will argue nonetheless for a simple but significant revision of Rule 26(b)(4)(B). Initially, we wanted to amend that paragraph quite radically, as follows:

> (B) A party may also, through interrogatories or deposition, discover the identity of, as well as facts known or opinions held by, any expert retained, specially employed, or consulted by another party in anticipation of litigation or preparation for trial and who is not expected to be called as a witness at trial.

Whereas the present paragraph (B) requires a showing of "exceptional circumstances" to discover the work of a non-testifying expert, our initial idea was to create an entitlement. And whereas the present rule limits even this court-ordered discovery to experts retained or specially employed, our proposed discovery as of right would have extended to the broader category of all those who had been consulted in any manner. Finally, our proposal would have given an explicit affirmative answer to a question not dealt with at all by the present discovery rules: whether a party can be forced to identify its non-testifying experts.

As we presented this proposal at two pre-publication conferences,[112] we argued for it on the basis of its obvious capability to prevent the sorts of abuses we have discussed above. More generally, we argued that this proposal would aid trial judges in the discernment of the scientific method. We remain convinced of that capability. The principal value of this radical approach is that it would force scientists appearing in court to behave like scientists. Science, like

---

108. *See* Easton, *supra* note 83.

109. *See id.* at 544-49.

110. *See id.* at 545-49.

111. *See id.* at 550.

112. The first was a conference on science and the law at Seton Hall University School of Law in Newark, New Jersey, on October 27, 2000, which involved practicing lawyers, judges, and scholars. The second was the editorial conference for this volume, held at Duke University School of Law on November 10-11, 2000. At the latter conference, we benefited especially from the always-perceptive comments of Duke professor Francis McGovern, who has vast experience managing and mediating complex scientific evidence cases for courts throughout the country.

the law, is an adversarial process, with an approximation of the truth emerging from a dialectic.[113] Science differs from the law, however, in that it is open, at least with respect to its methods, procedures, and data.[114] Whereas the law falls into the category—with sausages and, now, elections—of things that one does not want to watch being made, the essence of science is that one *does* watch it. As Douglas Crawford-Brown has pointed out elsewhere in this syposium, science is best defined by its practice.[115] Our initial, radical proposal had the preeminent virtue of bringing that practice to the courtroom.

At the two conferences, the proposal was subjected to withering criticism from several lawyers in attendance. Much of this critique seemed little more than defense of the status quo for its own sake. Some attacked our initial proposal solely because it went directly against the grain of the adversarial principle that each side should do its own work. Our response to this point is that separating the use of science from its abuse is difficult enough without having to contend with deceptive gamesmanship. Thus, this is simply another instance in which adversarial purity must yield to the interests of justice. Others thought that full disclosure would disproportionately disadvantage either plaintiffs or defendants—with the choice strongly correlated with the commentator's own practice background. But we think that the only parties who are systematically disadvantaged by full disclosure are those who rely on bad science. We have no solicitude for them, be they plaintiffs or defendants. Some of these same people also complained about the possible proliferation of depositions that our proposal portended. While this is a valid point, the answer depends on comparing the cost of expanded discovery to the benefit of complete disclosure in individual cases; on balance, we did not judge this problem to be a fatal flaw.[116]

Another element of the critique was more telling, however. Several plaintiffs' lawyers remarked that, at the outset of a personal injury or medical malpractice case, they routinely had a brief and informal conversation with a respected doctor of their acquaintance. This discussion came immediately after the initial conference with the client; its purpose was to help the lawyer decide whether to take the case and begin formal consultations with prospective testifying experts. Although we heard only from the plaintiffs' bar, it is plausible that similar conversations occur on the defense side for the purpose of making

113. *See* Goodstein, *supra* note 11, at 74.

114. *See id.* at 78-79.

115. *See* Douglas Crawford-Brown, *Scientific Models of Human Health Risk Analysis in Legal and Policy Decisions*, 64 LAW & CONTEMP. PROBS. 63 (Autumn 2001).

116. It was also suggested that judges can avoid the kinds of problems we have raised by appointing their own experts under Federal Rule of Evidence 706. Projects such as the Duke Private Adjudication Center's Registry of Independent Scientific and Technical Advisors have made it easier for courts to identify such experts. *See* Stephen Breyer, *Introduction to* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE, *supra* note 11, at 7-8. Nonetheless, all the available evidence continues to indicate that courts almost never use Rule 706. *See* Note, *Improving Judicial Gatekeeping: Technical Advisors and Scientific Evidence*, 110 HARV. L. REV. 941, 947 n.47 (1997) (stating that 80% of judges in the survey had never used a court-appointed expert); Schwarzer & Cecil, *supra* note 14, at 61 (explaining that, according to judges interviewed, such appointments are "infrequent"). Moreover, even a neutral expert will be unlikely to detect deceptive science if the critical evidence remains hidden.

an initial judgment about possible liability. These discussions are not necessarily limited to personal injury cases and might occur in any case hinging on expert advice.

Our interlocutors stressed two points: These conversations serve the interests of justice, and they could never take place if the experts being consulted thought that they might be subject to discovery or even identification. We were persuaded on both points. With respect to the first, there can be no doubt that the conversations described have the capacity to deter frivolous suits with attendant saving in public and private resources. On occasion, they may also expedite the settlement of cases that do have merit. The validity of the second point is self-evident.

As a consequence of this dialogue, we propose the following as a revised version of Rule 26(b)(4)(B):

> A party may, through interrogatories, discover the identity of, as well as facts known or opinions held by, any expert who has been retained or specially employed by another party in anticipation of litigation or preparation for trial and who is not expected to be called as a witness at trial. Further discovery of such experts, including depositions, may be taken only by leave of court, which shall be granted when the interests of justice so require. Discovery of experts consulted by another party but neither retained nor specially employed may be taken only by leave of court upon a showing of exceptional circumstances under which a fair trial on the merits may not be had without such discovery.

The three sentences of this paragraph seek to accomplish three different objectives. First, the interrogatory provision creates a preliminary entitlement to discover the identities and opinions of all experts who have been "retained or specially employed"[117] but will not testify. The interrogatory approach is the same as that followed with respect to testifying experts between 1970 and 1993.[118] We chose it to permit an adversary to learn of the existence and basic function of "hidden" experts without inviting an explosion of fishing-expedition depositions.

The second sentence requires a court order to depose a specially retained but non-testifying expert, as the current rule provides. However, whereas current Rule 26(b)(4)(B) requires a showing of "exceptional circumstances" involving the impracticability of getting comparable opinions from other sources, we propose a more generous "interests of justice" criterion.[119] The intent of the language is that courts allow further discovery when it appears calculated to expose the kinds of problem we have reviewed.

Finally, the third sentence requires leave of court before any discovery may be conducted against casually consulted experts. In this case, the party seeking such discovery must convince the court that the circumstances are so excep-

---

117. This phrase is currently used in Rule 26(b)(4)(B).

118. Prior to 1970, there was no discovery as of right against experts. The 1970 amendments to Rule 26(b)(4) introduced the testifying expert interrogatory. The current right to depose testifying experts was created by the 1993 amendments.

119. This language is borrowed from a very different context: Federal Rule of Civil Procedure 15(a), which deals with amendments to pleadings.

tional that a fair trial would otherwise be impossible. The intent of this approach is twofold. On the one hand, it protects those doctors and other informal experts who perform a valuable screening function. On the other, it allows for the possibility of discovery against more substantial hidden experts who are being disingenuously characterized as non-specially retained. Therefore, a party who proceeds through the discovery permitted under the first two sentences and learns of the existence of, for example, a statistician performing "casual" parallel analyses has the opportunity to persuade a court that further discovery is warranted.

Even this refined proposal does not avoid the fundamental objection that it violates the adversary principle by allowing one party to eavesdrop, as it were, on the other's trial preparation efforts, albeit in a balanced and symmetrical way. We see no less drastic alternative, however, if the actual and potential distortions of science we have chronicled are to be exposed. In the end, one must choose between two values: the sanctity of the adversarial process versus the openness that is a hallmark of science. Were the adversary principle utterly immutable, there would be no expert discovery at all—or any discovery, for that matter. In our judgment, it is time for the adversary principle to yield once again to the interests of fair and rational adjudication.