Review Essay

IMPROVING LEGAL STATISTICS

DAVID H. KAYE

Morris Degroot, Stephen E. Fienberg, and Joseph B. Kadane (eds.). *Statistics and the Law*. New York: John Wiley & Sons, 1986. xviii + 484 pp. Index.

Stephen E. Fienberg (ed.). *The Evolving Role of Statistical Assessments as Evidence in the Courts*. Report of the Panel on Statistical Assessments as Evidence in the Courts. New York: Springer-Verlag, 1989. xiii + 357 pp. Appendixes, bibliography, index.

Applications of mathematics to social phenomena and human behavior have a lengthy but not entirely respectable pedigree. In his day, Auguste Compte (1809: 168) decried "ponderous algebraic verbiage" in social and political theory. John Stuart Mill (1872) complained that "misapplications of the calculus of probabilities" have made it "the real opprobrium of mathematics." Echoing such sentiments, lawyers have voiced misgivings about occasional invasions of their turf by probabilists and statisticians. A little over thirty years ago, one trial lawyer, opposing "the vogue in recent years" of trying to "bolster sagging cases by applying the law of probability," insisted that "the law of probability has absolutely no application on the forensic field" (Houts, 1956). Nearly twenty years ago, a young law professor, Laurence Tribe (1971), eloquently warned his profession of the evils of using probability theory to quantify the probative value of certain forms of evidence. Since then, the attacks—some flashy and some baffling—on probability theory as a model of proof (e.g., Tillers and Green, 1988; Thompson, 1986; Jaffee, 1985) and on particular applications of probability of statistics to legal proof (e.g., Kaye, 1989, 1986a; Walker and Monahan, 1988; Sugrue and Fairley, 1983) have multiplied to the point of distraction.

The siren call of statistical expertise, however, has been irresistible. In recent decades, the use of quantitatively inclined expert witnesses—economists, sociologists, statisticians, and others—has swelled. As one observer of the litigation scene remarked, "[w]hat demonstrative evidence was to the 1960s and early 1970s, statistics have become to the 1980s—the hottest way to prove a

LAW & SOCIETY REVIEW, Volume 24, Number 5 (1990)

complicated case" (Lauter, 1984: 10). Now, in the 1990s, there are signs that the nascent discipline of legal statistics is coming of age and developing a sense of identity. The first conference devoted specifically to forensic statistics occurred last year (Royal Statistical Society, 1991). The latest textbooks on statistics and law (Finkelstein and Levin, 1990; Gastwirth, 1988a) display a new level of statistical erudition. And works examining the forensic aspects of legal statistics have emerged, among them Degroot, Fienberg, and Kandane's *Statistics and the Law* and Feinberg's *The Evolving Role of Statistical Assessments as Evidence*.

This review examines the last of these developments. It tries to indicate how statistical expertise relates to the judicial process and what can be done to improve this relationship. The focus is very much on the applied side of legal statistics, and I conclude with a personal experience that highlights some of the difficulties in making statistical assessments for and in the courtroom and underscores the need for reforms like those recommended in the Fienberg volume.

I. STATISTICIANS LOOK AT LEGAL STATISTICS

The legal statistics literature, one statistician waggishly reported, "is filled with articles and books in which lawyers explain to each other what statisticians are saying" (Aickin, 1986: 159). *Statistics and the Law* is different. In the words of its editors, Morris DeGroot, Stephen Fienberg, and Joseph Kadane of Carnegie-Mellon University:

> This book has been developed by statisticians for statisticians. Its basic purpose is to provide useful background and information to statisticians who may serve as expert witnesses or consultants in the future, and to give statisticians a better understanding of the legal process and philosophy to which the lawyers with whom they work must adhere. (P. ix)

A quick survey of its contents reveals the breadth of the field. The first chapter discusses the problem of establishing a prima facie case of employment discrimination in the selection of candidates for a position or benefit. Here, Paul Meier, Jerome Sacks, and Sandy Zabell cogently explain crucial concepts in the application of simple statistical techniques to employment discrimination cases[1] and argue that a prima facie case can be established by a dis-

[1] An earlier version of this chapter appeared as Meier *et al.*, (1984). Additional discussions of these issues may be found in the accompanying comment by Stephen Fienberg (pp. 41–46) and in Gastwirth (1988b) and Kaye (1985, 1986b).

parity in the flow of applicants that is significant at the .05 level[2] and that is substantial—as indicated by the "four-fifths rule."[3]

An essay by Benjamin King on statistics in antitrust litigation (pp. 49–78)[4] uses a hypothetical case involving a plaintiff's clearly fallacious analysis of price levels to reveal the obstacles to exposing to a court the defects in even an obviously misguided regression study.[5] Michael Finkelstein and Hans Levenbach examine the use of multiple regression and time series to estimate damages in price-fixing cases,[6] offering suggestions to promote wiser judicial reliance on this methodology (pp. 79–106). A final chapter on antitrust by Martin Geisel and Jean Mason (pp. 289–303) describes how the efficient market theory[7] may be applied in industry-wide studies of mergers and in firm-specific studies.

Delores Conway and Harry Roberts (pp. 107–68) move the discussion of regression into the employment discrimination domain[8]

---

[2] A disparity of this size (or larger) would arise no more than 5 percent of the time if the selection process gave members of the groups being compared equal opportunities to be chosen. An obvious alternative to insisting on this conventional but arbitrary significance level is simply to provide the court with the probability that if the selection process were independent of group membership, a disparity at least as large as the one actually observed would be seen. Even if this probability were to exceed .05, a court still could consider the statistical evidence along with plaintiff's other proof in deciding whether a prima facie case exists. The authors maintain, however, that failing to treat the .05 level as a prerequisite to giving the statistical evidence any weight "departs markedly from Fisher's usage and seems more likely to confuse than clarify in the legal setting" (p. 13). These assertions will not convince all observers that resort to the .05 convention is optimal in the legal context. (See Kadane, 1991; cf. *De Luca v. Merrill Dow Pharmaceuticals*, 1990 ($p < .05$ in separate epidemiological studies does not preclude expert opinion that a drug is teratogenic)).

[3] The four-fifths rule, originally promulgated by the Equal Employment Opportunity Commission, involves the ratio of selection rates for two groups. Let A be the rate at which members of a protected class are selected for employment, and let W be the rate for white applicants. If A/W is less than 4/5, the rule generally requires the employer to demonstrate that the selection procedure is justified by job-related considerations.

[4] Despite the title "Statistics in Antitrust Litigation," this chapter is by no means limited to antitrust cases. It describes, among other things, the pretrial discovery process for all civil litigants, and it offers much practical advice to statisticians preparing for any type of civil litigation. As Aickin (1986: 159–60) noted, the legal statistics literature is filled "to a lesser extent with articles in which statisticians explain to each other what lawyers are saying."

[5] Regression analysis involves fitting a function $Y = f(X_1, X_2, \ldots, X_m)$ to a set of observed values for these variables. The fitted function is said to be the regression of the dependent, or response, variable Y on the independent, or carrier, variables X_1 through X_m.

[6] This chapter is a slightly revised version of Finkelstein and Levenbach (1983).

[7] This theory implies that in an efficient capital market the prices of securities of firms that engage in anticompetitive conduct will reflect the anticipated monopoly profits.

[8] The chapter discusses both statistical issues and "legal considerations in the development of regression models" (p. 129) along the lines of the more legally sophisticated treatment in Finkelstein (1980).

and advance controversial claims[9] about "reverse regression."[10] Regression enters into yet a third legal context when John Pincus and John Rolph (pp. 257–86), who testified for the state of Washington in a constitutional challenge to its financing of public elementary and secondary schools, explain how they constructed a model of school resources and student achievement, how they presented their findings to the court, and what effect their work may have had.[11]

Several chapters concern applied probability models. These models can be important in litigation involving gambling machines. For example, Pennsylvania law subjects gambling equip-

---

[9] When one regresses job performance on variables that are proxies for productivity, one obtains an equation that predicts performance on the basis of these proxies. This ordinary regression equation is usually thought to establish that the use of the proxies in selecting or compensating employees is fair if, on average, minority and majority applicants with the same predictors do equally well on the job. In "reverse regression," one regresses a predictor on job performance. This addresses the question of whether minority and majority applicants who perform equally well on the job have, on average, the same predictors.

As the rejoinders by Stephan Michelson (pp. 169–82) and Arthur Goldberger (pp. 182–83) reveal, reverse regression has been criticized repeatedly. Despite cogent objections to reverse regression (e.g., Ash, 1986; Peterson and Novick, 1976), a National Research Council committee study of the General Aptitude Test Battery recently concluded that it could be unfair to select a majority applicant for a job over a lower scoring minority applicant even when the test scores of both groups correlate equally well with job performance (Hartigan and Wigdor, 1989). In fact, the committee proposed adding points to minority scores to compensate for the fact that when minority test takers tend to score below majority test takers on a test that does not predict job performance perfectly, most of the people with low scores who would perform adequately on the job will be minority applicants.

Whether such pursuit of "performance fairness" at the expense of "prediction fairness" is desirable or coherent is, at best, unclear. The Panel on Statistical Assessments as Evidence in the Courts notes that no "professional statistical consensus on the relevance and validity of reverse regression" exists (Fienberg, p. 98).

[10] The chapter is notable also for what it does not say. Although it describes or reanalyzes data used in an administrative proceeding involving the Harris Bank in Chicago, Roberts does not explicitly inform his readers that he consulted and testified for the bank in the case. Evidently, this failure to disclose consulting associations is not unique. Geisel and Masson (p. 290) speculate that "[p]ublications [on statistical issues in antitrust] may also appear in 'disguised' fashion so that one cannot immediately ascertain that the work began in consulting activity." A consulting relationship does not necessarily preclude subsequent academic objectivity, but it can indicate (1) that the author has had better access to one source of information than another, (2) that this access may have been channeled or controlled by the client, and (3) that the author may not be disposed to publish new conclusions adverse to the former employer (and that would contradict the researcher's own testimony or written work product). For such reasons, a researcher retained by a litigant should disclose that fact in subsequent academic publications that discuss the litigation or reach conclusions about the data in the case.

[11] As they acknowledge, and as Gastwirth (1988b) indicates, their work had its limitations. Their intention was not to disprove the existence of a relationship between student achievement and school resources, but merely to confirm that the failure of researchers in other states to discern such a relationship could apply to Washington schools.

ment to forfeiture, and a device that has no other reasonable use—a gambling device per se—may be confiscated without proof that it was used for gambling. Thus, when the state police seized an electronic draw poker machine in 1980, its owner sought proof that skill, more than chance, determines the outcomes of the plays. Jay Kadane (pp. 333–43) concluded from a quick empirical analysis of such a machine that a particular "dumb" strategy[12] is inferior to a "smart" one.[13] As Kadane observes, this analysis does not yield a quantitative statement of the relative importance of chance and skill, and John Lehoczsky's penetrating criticism (pp. 344–51) further illustrates the difficulties that can arise in devising a suitable analysis under an ambiguous legal standard.[14]

A probability analysis more commonly seen in the courts involves the "paternity index" and the "probability of paternity" derived from serological and other genetic tests.[15] These quantities often are shrouded in confusion,[16] but the survey chapter by Seymour Geisser and Donald Berry (pp. 353–82) on paternity probabilities and Donald Ylvisaker's commentary (pp. 383–90) should help dispel the most obvious misunderstandings.

Although this list does not exhaust the contents of *Statistics and the Law*,[17] it should illustrate the fact that judges, lawyers,

---

[12] The "dumb" strategy he tested consists of always standing pat with each hand.

[13] For the "smart" strategy, Kadane says, "I did my best to win, knowing what I do about probability" (p. 335).

[14] The empirical analysis was dispositive under the standard applied at trial and on appeal to the superior court. These courts held that a gambling device per se involves a game of pure chance. The Pennsylvania Supreme Court, however, held that although "[s]kill can improve the outcome," the fact that "the element of chance predominates" established that the machine was a gambling device per se (p. 342).

[15] The characteristics detected in the laboratory—the phenotypes or genotypes—occur with some probability $P(T|\beta=1)$ if the tested man is the biological father, and another conditional probability $P(T|\beta=0)$ if he is not. ($PT(T|\beta=0)$ is the p-value that would be used in a significance test, but that procedure makes little sense in this context.) The ratio of these likelihoods is the paternity index, and the "probability of paternity" is the resulting posterior probability given by Bayes's rule with some prior probability $P(\beta=1)$. As this notation suggests, the prior probability must be evaluated before any test results are considered. If the value of this probability is some number π, then the posterior probability is

$$P(\beta=1|T) = \frac{\pi P(T|\beta=1)}{\pi P(T|\beta=1) + (1-\pi)P(T|\beta=0)}. \quad (1)$$

[16] The literature includes claims that a posterior probability of .95 corresponds to a significance level of .05 (Saldeen, 1981; Silver and Schoppmann, 1987; cf. Li and Chakravarti, 1988), that the probability of paternity computed using the fact that the alleged father has types that do not exclude him as a biological father is an appropriate prior probability π to use in equation (1) of note 15 (Steinberg, 1987), and that a posterior probability of .88 makes paternity improbable (Peterson, 1982). For a dissection of such errors as displayed in several leading cases, see Kaye (1989).

[17] Dennis Guilland and Paul Meier (pp. 391–411) exchange ideas with Herbert Robbins (pp. 412–14) about the best way to model the effects of illegal

statisticians, and social scientists must address issues such as the evaluation and interpretation of quantitative evidence, the ethical and professional obligations of expert witnesses, and the roles of court-appointed witnesses. This is not an easy task. It has been made somewhat easier, however, by the work of a distinguished panel of judges, attorneys, statisticians and social scientists. I turn now to this report.

II. A BLUE-RIBBON PANEL REPORT[18]

The Evolving Role of Statistical Assessments as Evidence in the Courts, or *ERSA*, as I shall abbreviate it, is the report of the Panel on Statistical Assessments as Evidence in the Courts. Assembled by the National Research Council, financed by the National Science Foundation, and working jointly with other committees, this panel has produced a thoughtful report that will be of interest not merely to applied statisticians and members of the legal profession, but to all concerned with how the legal system does and should make use of scientific and statistical expertise.

The principal purpose of *ERSA* is to suggest reforms to enhance the value of statistical assessments in litigation. Before considering the panel's recommendations, however, some of the report's more scholarly components should be noted. For one, to illustrate the issues in the use of statistical assessments as evidence, *ERSA* includes six case studies of varying depth drawn from the fields of employment discrimination, environmental law, criminalistics, and antitrust. In most of these cases *ERSA* detects significant misunderstandings about statistical methods in the opinions of the judges. Such confusion often results from misstatements from experts (p. 198), and similar misconceptions can be found in early textbooks and articles on statistics for lawyers (Kaye, 1987a: 57–58; Kaye 1986b: 1348).

Cautioning that these discoveries of statistical *faux pas* in court opinions do not "suggest that judicial tribunals are necessarily inferior to other methods that might be used to resolve disputes

votes on elections. G. A. Whitmore (pp. 197–219) presents a case study of the role of statistical experts who applied such techniques as probability plots, the analysis of censored data, and life table methodology in litigation following a 1967 strike at a Quebec aluminum smelter. William Fairley and Jeffery Glen (pp. 221–39) urge lawyers to make more use of statisticians to prove damages in contract and tort cases, and they describe analyses estimating the losses to New York City from pilfering employees of a company hired to collect coins from the municipal parking meters. Gordon Apple, William Hunter, and Soren Bisgaard (pp. 417–47) provide a nontechnical introduction to the use of scientific data in environmental proceedings. Robert Coulan and Stephen Fienberg (pp. 305–32) give a case study of a court-appointed statistical expert in a federal employment discrimination case. In the concluding chapter Herbert Solomon (pp. 455–73) describes a number of cases involving such disparate matters as underpaid taxes, welfare eligibility, uninsured automobiles, and tire wear, in which confidence intervals figured prominently.

[18] A part of this section appeared in Kaye (1990).

that turn on statistical evidence" (p. 74), *ERSA* nevertheless looks for institutionally based explanations of the problems, and it generates many interesting hypotheses. For instance, *ERSA* questions the willingness of many courts "to accept claims that individuals with training in such diverse areas as economics, forensic medicine, and psychology all have sufficient background in statistical methodology to . . . present statistical testimony" (p. 74, citations omitted). It sees in courts a powerful urge to deny the existence of uncertainty, leading to inappropriate "all-or-nothing judgment[s] of statistical models and procedure," as in cases involving multivariate analyses of salaries (p. 78) and the acceptance of "research that should be suspect because it is too precise," as in prosecutions relying on a study purporting to show that scalp hairs that match one another in a microscopic examination have only a 1/4,500 chance of not being from a common source (p. 79).

In addition to the case studies and the speculations derived from them, *ERSA* reviews the use of statistics generally in employment discrimination litigation, antitrust litigation, and environmental law and toxic torts (pp. 85–137).[19] It elaborates, with effective examples, on the "two-culture" problem of law and statistics (pp. 139–48). It identifies some "psychological problems" with statistical testimony and offers some suggestions, informed by experimental studies in cognitive psychology, as to when judges and jurors may overvalue or undervalue statistical evidence (pp. 149–54; see also Kaye and Koehler, 1991; Thompson, 1989).

The heart of the report, however, is a series of recommendations intended to reform the law and to improve the practice of forensic experts. These are addressed to many audiences: courts, lawyers, law schools, professional societies, journal editors, and others. *ERSA* calls for the increased use of court-appointed experts; for an end to the practice of hiding the work of consulting experts from opposing counsel; for pretrial procedures to clear away readily resolvable differences among experts; for instruction in statistical concepts during and after law school; for retained experts to be more independent of the clients who pay them; and for regular publication in legal, statistical, and scientific journals of critical statistical reviews of expert presentations and judicial opinions.

ERSA describes two polar roles for expert witnesses: determined advocate and impartial educator. The advocate-expert puts "forward the strongest possible case for the employing attorney's client," while the impartial educator is "as neutral as possible" (p.

[19] Moreover, almost half the volume consists of appendixes and bibliographies. A few, like those excerpting or describing rules of evidence and procedure, apparently are intended as background to some of the panel recommendations or as a resource for expert witnesses who, like most, lack legal training. Others are considerably more original, evaluative, or esoteric. All are described in Kaye (1990).

157). Although *ERSA* concedes that it is impossible for a retained expert to be completely impartial, it emphasizes ethical considerations and professional standards that should propel the expert in this direction. "[E]xperts are looked on to present their own views based on the principles and standards of their professional community" (pp. 163–64). To represent their profession in this manner, experts are urged to maintain a high degree of professional autonomy.

This general exhortation translates into some specific "minimal standards" and safeguards. For example, *ERSA* suggests that experts secure an engagement letter explicitly recognizing their right to perform whatever analysis and have access to whatever data are required to address the issue in a professionally respectable fashion, to consult with colleagues who have not been retained by any party, and, depending on the expert's personal standards, to present as much of the analysis as is required to give a fair and full picture even if that works to the detriment of the employing party (p. 164).

Of course, all the entreaties and all the good intentions in the world will not eliminate the "hired gun" mentality among the pool of experts, especially among some whose livelihood comes from consulting fees. Such experts can present some of the most slanted and distorted views partly because almost no one who is in a position to judge their forensic work ever sees it. McNamara and St. George (1979) present a shocking example. For the trial of a Los Angeles mail-order firm for sending obscene photographs to New Mexico, an educational psychologist conducted a survey using descriptions of the allegedly obscene photographs. He testified that this survey showed the photographs to be within acceptable bounds according to community standards in Albuquerque. This expert had collected over $135,000 for similar work in thirty-seven trials in the past five years. In the Albuquerque case, however, an attempt was made to verify his findings. A replication of his survey using the actual photographs for a random sample of the original respondents revealed that 12 percent of the residential addresses allegedly visited did not exist, that 80 percent of the alleged respondents could not recall being interviewed, and that none of those who did remember the interview found the photographs acceptable. When the prosecution presented these facts, the jury convicted after an hour of deliberation, and an angry defense attorney demanded the return of the consultant's fee, holding him responsible "for putting my two clients behind bars." This disgruntled lawyer notified other attorneys of the incident, and it appears that the psychologist has not been employed again in such cases.

More often, however, the testimony is not so overtly fraudulent or perjurious. The expert who resorts to milder distortions currently has little to fear. Even in the unlikely event that the

courtroom performance receives professional scrutiny, this expert can return to the witness box to assure jurors in the next case that the criticism is mere academic or professional carping, or simply one of those legitimate disagreements that surface among experts with different opinions. Such talk would be less likely to prevail, however, if a published critique of the witness's testimony followed by the publication of serious correspondence in a respected journal raised doubts about professional competence. Many a lawyer would look elsewhere before retaining an expert whose forensic performance had been so questioned. For this reason, *ERSA*'s recommendation that "legal, statistics, and scientific journals publish, on a regular basis, critical statistical reviews of expert presentations and judicial opinions" is intriguing not only as a device to "educate judges, lawyers, and statistical experts" about the best resolution of specific methodological controversies, but also as a means to expose abuse and "promulgate higher professional standards" (p. 183). This kind of peer review of courtroom work may have some potential for modifying the behavior or employability of even the more venal or partisan experts for hire.

Would the adoption of these recommendations improve the process or outcomes of litigation involving statistical proof? I think so, and at the cost of adding yet another case study to the burgeoning literature, I hope to elucidate the merits of some of *ERSA*'s recommendations by considering a recent unreported case of conflicting expert presentations.

III. FORENSIC STATISTICS IN ACTION

In 1983–84, a university declined to grant tenure and promotion to an assistant professor whom I shall call Diana Doe.[20] Although she did receive tenure and promotion in the next academic year, Professor Doe brought a Title VII action for retroactive appointment, back pay, and declaratory and injunctive relief. Among other things, the complaint alleged a "pattern of granting merited promotion to women only after subjecting them to two full promotion cycles" (Plaintiff's Second Amended Complaint, p. 26, *Doe v. University*, 1989).

A month or so before the case came to trial, Doe's attorney contacted me, seeking an opinion that might counteract evidence that in the university as a whole, men and women tended to be promoted at roughly equal rates.[21] He supplied a list of the recom-

[20] At plaintiff's request, I am not identifying the parties by name.

[21] Because the attorney was unsure of who, if anyone, could produce a meaningful analysis in the time remaining before trial, he also contacted two other consulting experts. *ERSA* observes that "when alternative forms of data and analyses are known to attorneys, they are intentionally misleading the court, verging on deliberate misstatement, by revealing only those statistical data and analyses that are favorable to the client" (p. 167). To avoid such partial disclosure, it recommends that "if a party gives statistical data to different experts for competing analyses, that fact be disclosed to the testifying expert,"

mendations and actions taken on all college of liberal arts faculty considered for tenure and promotion during approximately a ten-year period, and suggested that the dean of this college discriminated against women in making his recommendations. I agreed to determine whether the data were consistent with this suspicion.

Because plaintiff's resources were limited and the available time very short, I used the simplest approach I could think of. To begin with, I ascertained that men and women candidates were promoted at roughly equal rates. Although this "bottom-line" analysis supported the university's position, I felt that this rather blunt analysis was insufficient. As explained in my written report:

> First, the comparison of simple promotion rates . . . assumes that male and female candidates are equally qualified for promotion. Second, it ignores the possibility that disparate treatment at one stage or by one person in the course of the promotion review process may be balanced or masked by decisions at other points. Finally, this method of analysis has no power to detect discriminatory treatment where bias comes into play in close cases, but most candidates are clearly qualified or unqualified. (Plaintiff's Exhibit P-150, pp. 3–4, footnote omitted)

To remedy these limitations with the data at hand, I used the recommendations of the college committed on promotions to control, to some extent, for the varying qualifications of the candidates for promotion:

> If the committee performs its task conscientiously and without bias, then those candidates whom it recommends for promotion—male and female alike—will be relatively qualified, while those for whom it recommends termination of employment usually will be less qualified. Naturally, the committee's judgments are not indubitably correct, and the dean need not follow the committee's recommendations. However, if the committee's arguable mistakes or misjudgments are not concentrated among male or female candidates, then the dean should not single out men or women for discordant treatment. A pattern of discordant recommendations in which women are substantially underrepresented in the group receiving "lenient" treatment from the dean *vis-à-vis* the committee or in which women are overrepresented in the group receiving "harsh" treatment *vis-à-vis* the committee is thus evidence of discriminatory treatment. (*Id.*, pp. 4–5, footnote omitted)

Since the committee and dean each had three choices for each candidate (terminate, hold for another year, or promote), the out-

---

so that "the statistical expert [may] reveal the history behind the development of the final statistical approach to an opponent when proper inquiry is made" (*id.*). This recommendation seems to have been met here. Plaintiff's attorney advised me that another expert was reviewing the same data, and I developed my analysis and reached by conclusions without seeing his work. As far as I know, the state never asked about the reports of other consultants.

Table 1. Recommendations of the Dean and College Committee

COMMITTEE RECOMMENDATION	DEAN'S RECOMMENDATION: Terminate Men	Terminate Women	Hold Men	Hold Women	Promote Men	Promote Women
Terminate						
1977	9	2	0	0	0	0
1978	5	0	0	0	1	0
1979	4	1	0	1	3	0
1980	6	2	0	0	1	0
1981	5	3	0	0	1	0
1982	6	3	0	0	0	0
1983	9	2	0	0	3	0
1984	13	2	0	0	0	0
1985	6	3	0	0	1	2
1986	2	2	0	0	1	1
Total	65	20	0	1	11	3
Hold						
1977	0	0	2	0	0	0
1978	0	0	1	0	0	0
1979	0	0	1	1	1	0
1980	0	0	2	0	1	0
1981	0	0	0	0	1	1
1982	0	0	1	0	0	1
1983	0	0	0	0	1	0
1984	0	0	0	0	0	0
1985	0	0	1	1	3	0
1986	0	0	0	0	0	0
Total	0	0	8	2	7	2
Promote						
1977	1	0	0	0	5	2
1978	0	0	0	0	3	1
1979	0	2	0	0	7	3
1980	0	1	0	0	4	2
1981	0	0	0	0	6	1
1982	0	0	0	0	4	1
1983	1	1	0	0	7	2
1984	0	0	0	0	8	1
1985	0	1	0	0	7	1
1986	1	0	0	0	9	1
Total	3	5	0	0	60	15

comes in the 167 cases of male candidates and the 54 cases of female candidate can be organized into a 3 × 3 table—that is, into nine cells (committee and dean recommend promotion, committee recommends promotion and dean recommends holding, committee recommends promotion and dean recommends termination, and so on). Table 1 presents these data. As one would expect if most cases were fairly clear, the dean's recommendations usually matched the committee's. In the 32 cases where they did not match, however, the dean usually favored men:

> [T]he "harsh" recommendations fell predominantly on women, while the "lenient" ones went mostly to men. Specifically, a woman receiving a discordant recommendation was more than three times as likely to receive a "harsh" recommendation from the Dean as was a man receiving a

https://doi.org/10.2307/3053668 Published online by Cambridge University Press

Table 2. Discordant Recommendations

	Men	Women
Lenient	18	6
Harsh	3	5

NOTE: "Lenient" = Dean recommends P or H when Committee recommends T, or Dean recommends P when Committee recommends H.
"Harsh" = Dean recommends T or H when Committee recommends P, or Dean recommends T when Committee recommends H.

$\Pr(\text{Harsh} \mid \text{Woman, Discordant}) = 0.45$
$\Pr(\text{Harsh} \mid \text{Man, Discordant}) = 0.14$

Relative risk of women for harsh treatment = .45/.14=3.18

> discordant recommendation. [W]hile women represented only about a quarter of those receiving "lenient" treatment, they constituted nearly two-thirds of those receiving "harsh" treatment. (*Id.*, p. 9–10)

These outcomes are presented in Table 2. Consequently, I concluded that "[t]he pattern of decisions of the dean *vis-à-vis* the college committee is largely consistent with the hypothesis of bias against women by the dean. Nevertheless, the statistical analysis itself cannot exclude other conceivable explanations, and the disparities noted here should be considered in light of the other pertinent, nonstatistical evidence" (*id.*, p. 10).

Plaintiff disclosed this report, which included tables more detailed than those set out here and p-values,[22] to the university's counsel, who then arranged for Pete Wolf, Jr., the president of Analytic Services, Inc.,[23] to testify for the university. However, the university did not make the substance of its expert's proposed testimony available to plaintiff.[24] This secretive procedure may

[22] In this context, a p-value is the probability of a difference of at least the magnitude of the observed disparity favoring men under a statistical model that assumes that decisions do not systematically favor men or women (cf. notes 2, 16).

[23] This Houston firm describes itself as "analytic experts" and lists such major corporations as Amoco, Northrop, Grumman Aerospace, and Gulf Oil among its clients.

[24] A few days before the trial, the university reported:

> Mr. Wolf will testify as an expert witness about certain statistical analyses . . . . He will testify about the proper uses of statistical inference . . . . The statistical procedures and test results that Mr. Wolf will testify about will include "standard deviation analysis" commonly employed in Title VII cases, the Fisher Exact Test, and the Probability Integral Transformation Test. . . . Mr. Wolf will testify that mistakes in procedures and interpretation of results in D. H. Kaye's reports . . . have led Professor Kaye to erroneous conclusions. (Defendant's Supplementary Answers to Plaintiff's First Set of Interrogatories, May 10, 1989)

The portions of these answers not quoted here did nothing to clarify this cryptic description. The vagueness of these answers may reflect counsel's incomplete understanding of the analysis the firm had conducted. Or it may be the product of a deliberate discovery tactic. After all, the art of vagueness in answering interrogatories is hardly confined to questions about statistics or the substance of expert testimony.

have been a consequence of the last-minute injection of the experts into the case,[25] but it was entirely at odds with *ERSA*'s recommendations for pretrial discovery and was to have unfortunate consequences at trial.

That trial occurred in federal district court in May 1989. To curtail its length, the trial judge limited each party to a total of four hours of direct and cross-examination.[26] The statistical testimony consumed about half of this time. After some skirmishing about the qualifications of a law professor to perform a statistical analysis,[27] I outlined the logic and results of my matching study and noted that, depending on the details of the analysis, male-female disparities as adverse or more adverse to women than those apparent in the data had "low to moderate" probabilities, ranging from .03 to about .10, of arising even if the dean treated women no differently than men.

On cross-examination, counsel spent considerable time going through the records of the female applicants for promotion one at a time, to show that a particular subtotal was erroneous. Of course, there is no reason to consume trial time with disagreements about simple numbers;[28] had the parties followed procedures like those proposed in *ERSA*, this diversion could have been avoided. The court should require an exchange of written reports before trial and a statement as to which parts there is disagreement on (cf. pp. 251–52).

Counsel also followed more significant lines of cross-examination. For example, she pointed out that the results can be sensitive to slight changes in cells of tables that contain small numbers—an appealing argument, at least superficially.[29] She further established that the apparent influence of gender in the discrepancies between the recommendations of the committee and the dean does

[25] Wolf presented fifteen transparencies to illustrate and detail his alternative analyses and criticisms, but none were entered as exhibits or given to plaintiff at or before the trial. I am grateful to him for providing me with copies after the trial.

[26] Oddly, Judge Nowlin did not rule on the admissibility of depositions and certain other evidentiary questions until the morning of the trial, making it impossible for the parties to know which witnesses they would need to call.

[27] I testified that I had some training in the physical sciences and applied statistics, that I studied and wrote about applications of statistics to law, that I had taught elementary statistics to law students, and that I used statistics in my work. The court ruled that I was qualified to testify as an expert in "that field." Plaintiff did not dispute the qualifications of defendant's consultant, who also had no degrees in statistics, who had never published anything on the subject, but who testified that he had taught "on the order of 20 courses" "in the area of statistics" and had testified as an expert witness in other cases.

[28] The number was not discussed on direct examination, and there were no errors in the tables contained in the report admitted into evidence.

[29] Critics of this reasoning include Baldus and Cole (1987: 180–81) (characterizing this argument as "double counting the influence of the small sample" when the extent of statistical error has been estimated) and Kadane (1991) ("statisticians should analyze the data sets they have, not make up new ones whose conclusions they like better").

not necessarily imply that the dean discriminated on the basis of gender; the numbers are also consistent with the hypothesis that the committee favored women over men and that the dean merely corrected its disposition toward such "affirmative action" in promotions.[30]

The next day, the university's expert testified that the data showed no gender bias. Although I cannot pretend to be entirely objective, I think it fair to say that he chose to operate in the partisan "forensic social science" mode criticized by the Panel on Statistical Assessments. He characterized the analysis of the discordant recommendations as "poppycock" and "data mining" because it "threw out 84% of the data," presupposed that the committee was always right, and aggregated the year-by-year data in violation of "the real world" and of precedents established in "many courts." Instead, he urged an analysis of the differences in the unmatched rates at which the dean recommended men and women for promotion, stratified by year. Strangely, he testified that the annual tables with a p-value of .5 showed that the dean was "neutral" in his treatment of women, while those with $p > .5$ or $p < .5$ showed decisions favoring or disfavoring women.[31] Applying some unusual procedures to test for the overall significance of the ten p-values,[32] he reached the same conclusion that plaintiff had reported from a simple pooling of the records over the ten-year period: the un-

[30] This explanation for the discrepancy between the committee and the dean was enumerated in a preliminary version of my report. At the request of plaintiff's counsel, I deleted it, leaving the reference to "other conceivable explanations" to cover this point in the version distributed to counsel and used in court. In retrospect, it would have been wiser to have retained the more explicit presentation, if only to cushion the cross-examination. As the Panel observes, the pressures on retained experts are hard to resist, and an expert who does not present the complete story risks exposure on cross-examination (pp. 158–59). In this instance, I did not resist or question counsel's suggestion because I saw the statistical analysis as establishing no more than the fact of (*a*) a disparity in the decisions of the dean and the committee (*b*) that operated to the detriment of women candidates and (*c*) that could not readily be dismissed as a statistical fluke. Thus, I was content to leave the expression of rival hypotheses that might explain this fact to counsel.

[31] This use of p-values is confused and pointlessly confusing. (Cf. Peterson, 1986: 333–34) ("One can only marvel that this proposal [involving a p-value of .50] has made it into print twice"). A p-value in this context is the probability that a disparity at least as large as that implicit in a table of the outcomes for men and women would arise if gender and outcome were statistically independent variables. Because this quantity depends on sample size as well as the direction and extent of the disparity, however, any thoughtful statistics text warns its readers against thinking of a p-value as a measure of the size of a disparity. A far simpler and more direct way to see whether women or men fared better is to consider whether a higher proportion of women as opposed to men were denied promotion each year. For a fuller discussion of statistics for describing the degree of discrimination in a selection process, see, e.g., Gastwirth (1988a: 206–10); Kaye (1985).

[32] Wolf applied the chi-square test to the set of ten p-values derived from the Fisher exact test, after subjecting these values to a probability integral transformation. I computed p-values for the pooled data with a simple t-test of the difference in proportions as recommended by D'Agnostino, Chase, and Belanger (1988). The literature on the merits of different procedures for compar-

matched promotion rates showed no discrimination against women.

The university's expert applied a similar analysis to the discordant cells of the dean-committee cross-tabulation. He used the magnitude of the p-values in each year as a measure of the extent of disparate treatment, then applied unusual methods (see supra note 32) to conclude that this series of p-values was not significant. Despite the several p-values given in plaintiff's expert report and the direct examination about this report, the university's expert blithely insisted that plaintiff did not report the statistical significance of the disparity in the thirty-two cases of disagreement between the dean and the committee. Playing the often criticized (e.g., Kaye, 1986b, 1987b, 1982) "magic number" game, he claimed that the p-value was .07, which "does not in any way statistically indicate discrimination."[33] Finally, he undertook a sensitivity analysis of the aggregated data and reported that if one changed the numbers in the various cells slightly, the p-values ranged from .02 to .26.[34]

Perceiving the obstacles to an effective response to the details of this presentation as overwhelming, plaintiff made little effort to challenge the testimony. The attorneys and the court probably did not understand the meaning of a p-value, let alone the subtleties of which procedure should be used to compute it (cf. p. 94) or its relationship to a sensitivity analysis; the university did not reveal its unusual methods of analysis in advance of trial; and the court had imposed severe time constraints on the parties. Recalled as a witness, I testified that matching studies are not "data mining," but are a commonly used statistical procedure that sacrifices a certain amount of sample size for the sake of an enhanced power to detect differences in treatments and that I had settled on this approach before analyzing the data for statistical significance. I explained

---

ing two binomial proportions is subtle and apparently unrelenting (e.g., Storer and Kim, 1990; Little, 1989).

Although none of the procedures for computing p-values involved a normal distribution, Wolf also treated the p-values for chi-square as if it represented the area under a given number of standard deviations about the mean of a normal curve. This idiosyncratic procedure permitted him to claim that "under the law" (which he interpreted as requiring "two or three standard deviations"), the differences were "not significant." From the statistical standpoint, computing a test statistic, finding that p-value associated with it, then computing an inapplicable test statistic that would have the same p-value, and using this inapposite statistic to test for significance is bizarre. It is also inadvisable from a legal standpoint (cf. Baldus and Cole, 1987: 309).

[33] It would not be unfair to describe this testimony as exaggerated; e.g., Baldus and Cole, 1987: 185 ("the .07 P-value represents an intermediate level of significance, giving partial but not decisive support to the inference that the observed disparity was more than a consequence of chance factors"); Baldus and Cole, 1980: 308 ("if a level is set to determine what is or is not statistically significant, that judgment is a legal determination properly made by a court and not by an expert").

[34] As mentioned in note 28, the appropriateness of this procedure is debatable.

that I did not make any particular assumptions about the superiority of the committee or the dean in making promotion recommendations, but simply investigated whether the differences between these two sets of recommendations were independent of gender. Finally, I suggested that stratification by year was fine but that the yearly outcomes needed to be evaluated with different procedures than those adopted by the university's expert (Kaye, 1985). After a brief cross-examination, the trial ended and counsel tendered final arguments in written form.

Four months later, the court entered its judgment finding plaintiff's proof of disparate treatment wanting. The opinion made no attempt to describe the university's expert testimony but concluded that plaintiff's statistical analysis provided "little evidence" of sex discrimination (Order, Sept. 13, 1989, p. 20). The court gave three reasons for this conclusion. First, as the university had urged, it reasoned that the college committee was not "an appropriate standard" with which to measure the dean's decisions because "the committee may make mistakes" (*id.*, p. 19).

This criticism misperceives the point of the analysis, which is intended only to identify a gender-based pattern of discordant decisions that calls for some explanation from the university. This explanation might be that all the committee's "mistakes" happened to disadvantage men, but unless there is some a priori reason for this to be so, the burden should be on the university to perform a further analysis that would support the hypothesis that the committee's mistakes are correlated with gender (Kaye, 1988).

Adherence to *ERSA*'s proposals could have led the court to a clearer understanding of the logic of the matching studies. A court-appointed expert (p. 171) probably would have seen that the matching analysis does not assume that every committee decision is correct and would have explained that the logic is merely that if the mistakes made by the dean and the committee are not biased against men or women, then the differences in the decisions of these bodies will not systematically favor men or women. Pretrial procedures to narrow the statistical disputes (p. 166) would have allowed the court to devote more attention to this point and avoided the distraction of a dispute over the meaning of a p-value of .07 and the relative merits of such esoterica as the Fisher exact test, the t-test, and the twenty other procedures that were mentioned for testing the difference between two binomial proportions. Quite possibly, the innovation of having "both experts placed under oath and, in effect, [engaged] in a dialogue" (p. 174) would have clarified the disagreement over the appropriateness of focusing on the disagreements between the dean and the college committee.

Second, the court relied on a point that the university did not emphasize at trial. It noticed that "the committee recommendations were not always unanimous decisions, and the difference be-

tween a recommendation to promote or to terminate may be one or two votes" (*id.*, p. 20). The use of a simple 3 × 3 table to categorize the data, the court complained, did not account for such variations.

When a court questions a statistical study on grounds like this, several courses of action are open to it. First, it can dismiss the findings as inconclusive. Unless the omission seems likely to bias the analysis in favor of its proponent or unless the defect is so drastic as to render the analysis irrelevant, however, the findings should be given *some* weight. In *Bazemore v. Friday* (1986: 400), the Supreme Court unanimously declared it "plainly incorrect" for lower courts to dismiss as "unacceptable" a regression study of salaries that "accounts for the major factors" merely because the analysis did not include "all measurable variables thought to have an effect on salary level." If speculations about omitted or imperfectly measured variables were enough to vitiate a study, virtually all applied statistical work could be ignored.

A second response to an analysis that seems incomplete is to complete it. Yet, when courts undertake their own statistical analyses, the results can be dismaying, as shown in, for example, Kaye (1985, 1982). Opinions that read "one of my law clerks advised me that, given the size of the two-year sample, there is only about a 5% likelihood" (*Hazelwood School District v. United States* (1977: 318 n.5) (Stevens, J., dissenting)) seem more candid than convincing.

Thus, the *ERSA* Panel recommends a third course of action—that "in general, judges not conduct analytical statistical studies on their own. If a court is not satisfied with the statistical evidence before it, alternative means should be used to clarify matters, such as a request for additional submissions from the parties or even, in exceptional circumstances, a reopening of the case to receive additional evidence" (p. 176).

Post-trial submissions are most appropriate when concerns not previously apparent to the parties emanate from the court. On the other hand, when the criticism originates with the party questioning the statistical evidence, it can and should be addressed before trial. A party should not be permitted to wait until the midst of a trial to suggest refinements to a basically reasonable analysis. Indeed, in *Bazemore v. Friday* (1986:403–4 n.14), the Court noted with dissatisfaction that "Respondent's strategy at trial was to declare simply that many factors go into making up an individual employee's salary; they made no attempt that we are aware of—statistical or otherwise—to demonstrate that when these factors were properly organized and accounted for there was no significant disparity between the salaries of blacks and whites." Whether defendant should undertake the more refined analysis, as the *Bazemore* Court implied, is not always obvious (Kaye, 1988), but insisting on pretrial notice of technical objections greatly en-

hances the chance that a serious criticism will receive expert scrutiny and save the court from having to engage in an unassisted analysis or to speculate about what a proper analysis of the data would reveal. In *Doe*, for instance, had the university raised in a timely manner the concern that the court later voiced, plaintiff could have investigated whether the discordant recommendations disadvantaging women or advantaging men tended to be the results of close votes.

Finally, the court in *Doe* complained that the "analysis of cases with discordant recommendations involved only 32 candidates. When such small numbers are being analyzed, one more or fewer incidents of [Dean] King's treating a woman harshly or a man leniently could significantly change whether the statistics evidence sex discrimination" (Order, p. 20). This point, too, was not fully explored at trial, for the university did not notify plaintiff of this criticism before trial and did not disclose its sensitivity analysis until the final day of the trial. Had *ERSA*'s recommendations for more reasonable pretrial exploration of statistical analyses been followed, it seems likely that the record would have shown that the association between gender and discordant treatment was not overly sensitive to small changes in the data set.[35]

Of course, I cannot prove that had *Doe* been litigated in the manner proposed by the *ERSA* Panel, the outcome would have been different. Even if the data analyses had been sharpened and the more spurious or distracting statistical claims eliminated by pretrial conferences of the opposing experts, by a more revealing format for their testimony, and by more informed cross-examination, the judgment easily could have been the same. After all, the statistical proof was merely suggestive of some discrimination, and the other evidence did little to persuade the court that the dean or anyone else mistreated Professor Doe because she was a woman. Still, the fact that one category of testimony is not necessarily dispositive is no argument against procedures reasonably calculated to promote balanced and unobscured statistical presentations.

---

[35] As it was, the university's evidence on this point largely supported the stability of the statistical association between gender and the type of discordant treatment a candidate received. The table that its expert presented as demonstrating the greatest diminution in this association (as erroneously measured by the p-value) moved one man from the "lenient" to the "harsh" treatment category *and* moved one woman from "harsh" to "lenient." In that scenario 4/21 discordant cases of men and 4/11 discordant cases of women involved "harsh" treatment, making the relative risk of "harsh" treatment from the dean $(4/11) \div (4/21) = 1.91$ instead of 3.18 (see Table 2) and increasing the Fisher exact test p-value from .07 to .26. All other variations of the data considered by the university gave greater relative risks and much smaller p-values, almost always in the neighborhood of the statistics presented by plaintiff on the basis of the real data. In fact, three of the seven alternatives advanced by defendant were *more* favorable to plaintiff's case than were the real data analyzed by plaintiff. The university's successful withholding of its analyses and the court's restrictive time constraints prevented any careful presentation of these points during the trial.

Most evidentiary and procedural rules are designed to encourage the production of probative evidence that will be evaluated for whatever it is worth. Evidence admitted or excluded as a result of this structure sometimes will tip the balance in favor of one side or the other, but often it will not make so crucial a difference. Nevertheless, the objective of structuring trial and pretrial procedures to generate useful evidence that the judge or jury will value properly remains worthwhile.

Thus, I have described *Doe* not to argue that it represents a miscarriage of justice but to show that the creation and fate of statistical evidence is linked with and affected by pretrial and trial procedures. If statistical proof is not developed in a way that promotes the application of appropriate methods to accurate data and is not presented in a way that enhances the chance of its being understood and evaluated intelligently and fairly, then it will remain a rhetorical device for attorneys and judges to manipulate rather than a scientific tool for the discovery of the truth.[36] This is the state of affairs that *ERSA* seeks to correct, and its recommendations and observations are well conceived and sorely needed.

REFERENCES

AICKIN, Mikel (1986) "Issues and Methods in Discrimination Statistics," in D. H. Kaye and M. Aickin (eds.), *Statistical Methods in Discrimination Litigation*. New York: Marcel Dekker.

ASH, Arlene S. (1986) "The Perverse Logic of Reverse Regression," in D. H. Kaye and M. Aickin (eds.), *Statistical Methods in Discrimination Litigation*. New York: Marcel Dekker.

BALDUS, David C., and James W. COLE (1987) *Statistics as Proof of Discrimination: 1987 Cumulative Supplement*. New York: McGraw-Hill.

——— (1980) *Statistical Proof of Discrimination*. New York: McGraw-Hill.

COMPTE, Auguste (1809) 4 *Cours de Philosophie Positive* (3d ed.), 367, reprinted in Jean-Paul Enthoven (ed.), *Physique Sociale: Cours de Philosophie Positive, Leçons 46 à 60* (1975).

D'AGNOSTINO, R. B., W. CHASE and A. BELANGER (1988), "The Appropriateness of Some Common Procedures for Testing Equality of Two Independent Binomial Proportions," 42 *American Statistician* 198.

FINKELSTEIN, Michael O. (1980) "The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases," 80 *Columbia Law Review* 737.

FINKELSTEIN, Michael O., and Hans LEVENBACH (1983) "Regression Esti-

[36] Plainly, I believe that some statistical analyses are demonstrably better than others for assessing particular data and that the legal system should be structured to generate the preferred assessments. This does not mean that there is no room for argument and rhetoric about how the judge or jury should respond to such evidence or that the parties have no role in the development and presentation of expert evidence. It does mean that when we involve statistical and other experts in the resolution of disputed questions of fact, we should do so in a way designed to produce testimony that is not grossly unrepresentative of professional opinion generally. This view should be congenial not only to those who identify as the primary objective of litigation the discovery or reconstruction of truth, but also to those who perceive the overriding goal to be the production of verdicts that litigants will regard as fair or that society will accept as conclusive.

mates of Damages in Price-Fixing Cases," 46 *Law and Contemporary Problems* 145.

FINKELSTEIN, Michael O., and Bruce LEVIN (1990) *Statistics for Lawyers*. New York: Springer-Verlag.

GASTWIRTH, Joseph L. (1988a) *Statistical Reasoning in Law and Public Policy*. Boston: Academic Press.

——— (1988b) "Book Review," 28 *Jurimetrics Journal* 345.

HARTIGAN, John A., and Alexandra WIDGOR (eds.) (1989) *Fairness in Employment Testing*. Washington, DC: National Academy Press.

HOUTS, Marshall (1956) *From Evidence to Proof: A Searching Analysis of Methods to Establish Fact*. Springfield, IL: Charles C. Thomas.

JAFFEE, Leonard (1985) "Of Probativity and Probability: Statistics, Scientific Evidence, and the Calculus of Chance at Trial," 46 *University of Pittsburgh Law Review* 925.

KADANE, Joseph B. (1991) "A Statistical Analysis of Adverse Impact of Employer Decisions," *Journal of the American Statistical Association* (in press).

KAYE, David H. (1990) "Statistical Proof in the Courts," 35 *Contemporary Psychology* 839.

——— (1989) "The Probability of an Ultimate Issue: The Strange Cases of Paternity Testing," 75 *Iowa Law Review* 75.

——— (1988) "Statistical Evidence: How to Avoid the 'Diderot Effect' of Getting Stumped," 2 *Inside Litigation* 21.

——— (1987a) "Apples and Oranges: Confidence Coefficients and the Burden of Persuasion," 73 *Cornell Law Review* 54.

——— (1987b) "Hypothesis Testing in the Courtroom," in A. Gelfand (ed.), *Contributions to the Theory and Application of Statistics*. New York: Academic Press.

——— (1986a) "Ruminations on Jurimetrics: Hypergeometric Confusion in the Fourth Circuit," 26 *Jurimetrics Journal* 215.

——— (1986b) "Is Proof of Statistical Significance Relevant?" 61 *Washington Law Review* 1333.

——— (1985) "Statistical Analysis in Jury Discrimination Cases," 25 *Jurimetrics Journal* 274.

——— (1982) "The Numbers Game: Statistical Inference in Discrimination Cases," 80 *Michigan Law Review* 833.

KAYE, David H., and Jonathan KOEHLER (1991) "Can Jurors Understand Probabilistic Evidence?" 154 *Journal of the Royal Statistical Society Series A* 75.

LAUTER, David (1984) "Making a Case with Statistics," *National Law Journal*, 10 Dec., p. 1.

LI, C. C., and A. CHAKRAVARTI (1988) "An Expository Review of Two Methods for Calculating the Paternity Probability," 43 *American Journal of Human Genetics* 197.

LITTLE, Roderick J. A. (1989) "Testing the Equality of Two Independent Binomial Proportions," 43 *American Statistician* 283.

McNAMARA, Patrick H., and Arthur ST. GEORGE (1979) " 'Porno' Litigation, Community Standards, and the Phony Expert: A Case Study of Fraudulent Research in the Courtroom," 3 *Sociological Practice* 45.

MEIER, Paul, Jerome SACKS, and Sandy L. ZABELL (1984) "What Happened in *Hazelwood*: Statistics, Employment Discrimination, and the 80% Rule," 1984 *American Bar Foundation Research Journal* 139.

MILL, John Stuart (1872) 3 *A System of Logic, Ratiocinative and Inductive* (8th ed.), reprinted in J. M. Robson (ed.), *Collected Works of John Stuart Mill*. Toronto: University of Toronto Press.

PETERSON, David W. (1986) "Book Review," 26 *Jurimetrics Journal* 329.

PETERSON, Nancy S., and Melvin R. NOVICK (1976) "An Evaluation of Some Models for Culture-Fair Selection," 13 *Journal of Educational Measurement* 3.

PETERSON, Robert (1982) "A Few Things You Should Know About Paternity Tests (But Were Afraid to Ask)," 22 *Santa Clara Law Review* 667.

ROYAL STATISTICS SOCIETY (1991) "Proceedings," 154 *Journal of the Royal Statistics Society* Series A (in press).

SALDEEN, Ake (1981) "Jurimetrics and the Ascertainment of Paternity," 25 *Scandanavian Studies in Law* 169.
SILVER, Herbert, and A. SCHOPPMANN (1987) "Limitations of Paternity Testing Calculations," 27 *Transfusion* 288.
STEINBERG, A. G. (1987) "More on Paternity," 41 *American Journal of Human Genetics* 77.
STORER, Barry E. and Kim CHOONGRAK (1990) "Exact Properties of Some Exact Test Statistics for Comparing Two Binomial Proportions," 85 *Journal of the American Statistical Association* 146.
SUGRUE, Thomas, and William B. FAIRLEY (1983) "A Case of Unexamined Assumptions: The Use and Misuse of the Statistical Analysis of Castaneda/Hazelwood in Discrimination Litigation," 24 *Boston College Law Review* 925.
THOMPSON, William C. (1989) "Are Jurors Competent to Evaluate Statistical Evidence?" 52 *Law and Contemporary Problems* 9.
TILLERS, Peter, and Eric GREEN (eds.) (1988) *Probability and Inference in the Law of Evidence: The Limits and Uses of Bayesianism*. Dordrecht: Kluwer Academic Publishing.
THOMPSON, Judith J. (1986) "Liability and Individualized Evidence," 49 *Law and Contemporary Problems* 199.
TRIBE, Laurence (1971) "Trial by Mathematics: Precision and Ritual in the Legal Process," 84 *Harvard Law Review* 1329.
WALKER, Laurens, and John Monahan (1988) "Social Facts: Scientific Methodology as Legal Precedent," 76 *California Law Review* 877.

CASES CITED

Doe v. University, Civ. No. A-87-CA-015 (W.D. Texas, 1989).
Bazemore v. Friday, 478 U.S. 385 (1986).
Hazelwood School District v. United States, 433 U.S. 299 (1977).
De Luca v. Merrill Dow Pharmaceuticals, 911 F.2d 941 (3d Cir. 1990).