ARE JURIES COMPETENT TO EVALUATE STATISTICAL EVIDENCE?

WILLIAM C. THOMPSON*

I

INTRODUCTION

The issue of jury competence has arisen due to concerns about the ability of jurors to deal appropriately with the increasing complexity of evidence presented in trials. While most attention has focused on complex civil litigation,[1] criminal trials have grown more complex as well, due in part to revolutionary advances in the forensic sciences.[2] New procedures for criminal identification, such as protein gel electrophoresis, DNA typing, gas chromotography, and neutron activation analysis have recently become available for use at trial.[3] Additionally, the technology behind many of the more traditional identification techniques, such as bite mark comparison and hair comparison, has advanced in recent years.[4] For a juror to understand and evaluate the technology underlying these techniques is often a formidable task in itself.

Adding further to the difficulty is the probabilistic nature of much of this new evidence. The results of forensic tests are often meaningful only if they are accompanied by statistical data. For example, evidence that the defendant in a rape case has genetic markers matching those in semen recovered from a rape victim cannot be evaluated without statistical information on the frequency of the matching markers in the population. Because forensic tests are often less than perfectly reliable, statistical data on the error rate of the test may be necessary as well.[5] Hence, jurors may hear that a criminalist compared a sample of the defendant's blood to a semen sample taken from the rape victim using a procedure known as protein gel electrophoresis and found that the two samples contain a common set of genetic markers that collectively occur in only 1.5 percent of the population. The jurors may also hear, however, that proficiency tests have found that criminalists misclassify

---

* Associate Professor, Program of Social Ecology, University of California, Irvine.

1. *See, e.g.*, M. SAKS & R. VAN DUIZEND, THE USE OF SCIENTIFIC EVIDENCE IN LITIGATION (1983); Austin, *Jury Perceptions on Advocacy: A Case Study*, 8 LITIGATION 15 (1982); Lempert, *Civil Juries and Complex Cases: Let's Not Rush to Judgment*, 80 MICH. L. REV. 68 (1981).

2. P. GIANNELLI & E. IMWINKELRIED, SCIENTIFIC EVIDENCE at xxi (1987).

3. *Id.*

4. *Id.* at 369-83, 1013-39.

5. Among forensic scientists there is a "growing recognition that in many cases the results obtained yield their maximum information only if statistical methods and calculations of probability are used." Walls, *Ten Years of Forensic Science—1964-73*, 1974 CRIM. L. REV. 504, 505.

genetic markers in blood in semen at rates ranging from 1 to 6 percent per marker. What should jurors make of such evidence? What do they make of it?

The use of such statistical data in court is growing rapidly.[6] According to one authority, "our criminal justice system is now at the threshold of an explosion in the presentation of mathematical testimony."[7] The complexity of such testimony has raised concerns about the ability of jurors to deal with such evidence appropriately.[8] Empirical studies of the ability of lay individuals to use the type of statistical evidence presented in criminal trials have emerged only recently.[9] While this new literature is small and full of gaps, it is beginning to define the strengths and weaknesses of statistical reasoning by laypersons in ways that should prove quite helpful to courts facing decisions about the admissibility of statistical evidence and about the manner in which statistics should be presented to the jury.

II

THE NATURE OF STATISTICAL EVIDENCE IN CRIMINAL TRIALS

A discussion of jurors' competence to evaluate statistical evidence must necessarily begin with a description of the statistics jurors may encounter in a criminal trial and with a discussion of how jurors should evaluate those statistics.[10] There are two basic types of statistics: base rates and error rates. A base rate measures the frequency at which an event or characteristic occurs in a population.[11] An error rate measures the frequency at which a test or procedure produces wrong results. Although an error rate is a type of base rate, error rate statistics raise special issues distinct from those surrounding

---

6. E. IMWINKELRIED, THE METHOD OF ATTACKING SCIENTIFIC EVIDENCE (1982); Jonakait, *When Blood is Their Argument: Probabilities in Criminal Cases, Genetic Markers, and Once Again, Bayes' Theorem*, U. ILL. L. REV. 369 (1983); Walls, *supra* note 5, at 505.

7. Jonakait, *supra* note 6, at 369.

8. Saks & Kidd, *Human Information Processing and Adjudication: Trial by Heuristics*, 15 LAW & SOC'Y REV. 123 (1980); Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329 (1971).

9. *See, e.g.*, J. GOODMAN, PROBABILISTIC SCIENTIFIC EVIDENCE: JURORS' INFERENCES (1988); Faigman & Baglioni, *Bayes' Theorem in the Trial Process*, 12 LAW & HUM. BEHAV. 1 (1988); Thompson, Britton & Schumann, *Jurors' Sensitivity to Variations in Statistical Evidence*, J. APPLIED SOC. PSYCHOLOGY (in press); Thompson & Schumann, *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy*, 11 LAW & HUM. BEHAV. 167 (1987); J. Goodman, Jurors' Comprehension of Scientific Evidence (June 1988) (paper presented at the Meeting of the Law & Society Association, Vail, Colo.); E. Schumann & W. Thompson, Effects of Attorney's Arguments on Jurors' Interpretation of Statistical Evidence (June 1989) (paper presented at the Meeting of the Law & Society Association, Madison, Wis.); W. Thompson, J. Meeker & L. Britton, Recognizing Conditional Dependencies in Evidence: Effects of Group Deliberation (June 1987) (paper presented at the Meeting of the Law & Society Association, Washington, D.C.).

10. The information reported here was gleaned from a review of appellate opinions and from a recent survey in which fifty forensic scientists in California were questioned about the types of statistical evidence they present in court and the ways in which they present it. N. Miller, The Role of Statistical Scientific Evidence in Criminal Trials: A Survey of Criminologists (1986) (unpublished thesis, University of California, Irvine). A cogent discussion of the various types of statistical evidence is also provided by Kaye, *The Admissibility of "Probability Evidence" in Criminal Trials—Part II*, 27 JURIMETRICS J. 160, 161-64 (1987).

11. *See generally*, Koehler & Shaviro, *Veridical Verdicts: Increasing Verdict Accuracy Through the Use of Overtly Probabilistic Evidence and Methods*, 75 CORNELL L. REV. (in press).

other statistics on the frequency of events. Accordingly, base rates and error rates will be discussed separately.

A. Base Rates

Base rate statistics are usually developed from empirical studies in which a population, or a sample drawn from the population, is surveyed to determine the frequency of the event or attribute. For example, a survey showing 40 percent of a sample of Caucasians have type A blood establishes a base rate of type A blood among Caucasians. A study showing that 15 percent of the taxicabs in a city are green establishes a base rate of green cabs in the city. A study showing that 90 percent of defendants tried for burglary are convicted establishes a base rate of convictions among burglary defendants. The base rate of an event or attribute is equal to the probability that it will be present in a randomly selected member of the relevant population.[12]

1. *Directly Relevant and Indirectly Relevant Base Rates.* Base rate statistics can be used to prove a fact in two distinct ways. In some instances, the base rate is directly relevant to a target outcome because it directly expresses the frequency of that outcome. When a pedestrian is struck by a bus of unknown origin, evidence that a particular company operated 90 percent of the buses on that route is directly relevant to the question of who owned the bus. Similarly, when a defendant possessing heroin has been charged with concealing an illegally imported narcotic, evidence that 98 percent of all heroin is illegally imported is directly relevant to the question of whether the heroin possessed by the defendant was illegally imported. In such instances, the base rate is said to establish a prior probability of the target outcome.[13] If 90 percent of the buses that could have been involved in the accident are owned by a particular company, then there is a prior probability of .90 that that company owned the offending bus.

In other instances, the base rate is only indirectly relevant to a target outcome, and must be combined with other information before any probabilistic assessment of the target outcome is possible. When forensic tests link a criminal defendant to a crime by showing his blood type matches that of the perpetrator, evidence that the blood type is found in 5 percent of the population is relevant to the ultimate issue of the defendant's guilt, but only indirectly. The base rate of the blood type does not, by itself, reveal anything about the likelihood of the target outcome—the defendant's guilt—and thus, unlike a directly relevant base rate, does not establish a prior probability of the target outcome. Instead, it speaks to the likelihood the defendant might, by chance, have a "matching" blood type if innocent, and thus helps to establish the value of the forensic evidence.

---

12. *Id.*

13. The prior probability of a target outcome is the probability a reasonable person would assign to that outcome prior to receiving any case-specific or individuating information.

The use of "directly relevant" base rates as evidence in court has been controversial, particularly where the base rate is the sole evidence of a target outcome. Base rates of this sort have been labeled "naked statistical evidence"[14] and have generally been held inadmissible.[15] A few courts, however, have admitted such evidence.[16] The most widely discussed case involving "naked statistical evidence" is *Smith v. Rapid Transit,*[17] in which plaintiff was struck by a hit-and-run bus and based her claim that the bus was the defendant's solely on evidence that the defendant operated 90 percent of the buses in the city. The Massachusetts Supreme Judicial Court sustained defendant's motion for summary judgment on grounds that the base rate statistic was insufficient to make a case against the defendant in the absence of more particularized proof of the ownership of the offending bus.

While most commentators agree with this holding, they disagree on the rationale. One group, which has been labeled "anti-Bayesian,"[18] argues that base rates are inherently inferior to more particularized evidence and have little or no relevance unless they reflect a "suitably narrowed down reference class."[19] By this account, the frequency of defendant's buses among all buses in the city is merely "background information" that does not necessarily reflect the likelihood that the hit-and-run bus was defendant's, and therefore is an insufficient basis for a holding in plaintiff's favor.[20] Other commentators, sometimes labeled Bayesians, maintain that base rates need not meet any standard of specificity in order to be relevant.[21] By their account, the fact that defendant operates 90 percent of the buses in the city is highly relevant because it establishes a prior probability of .90 that the offending bus was defendant's. This estimate is subject to modification in light of additional evidence, of course; but the most accurate estimate one can make of the likelihood the bus was the defendant's, in the absence of other evidence, is .90. While many commentators in the Bayesian camp agree with the holding in *Smith,* they do so on grounds of policy considerations unrelated to doubts about the evidentiary value of base rates.[22]

---

14. Kaye, *The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation,* 2 AM. B. FOUND. RES. J. 487, 488 (1982).

15. *See, e.g.,* Smith v. Rapid Transit, 317 Mass. 469, 58 N.E.2d 754 (1945).

16. *E.g.,* Turner v. U.S., 396 U.S. 398, 414-16, *reh'g denied,* 397 U.S. 958 (1970); Sindell v. Abbott Labs, 26 Cal. 3d 588, 607 P.2d 924, 163 Cal. Rptr. 132 (1980).

17. 317 Mass. 469, 58 N.E.2d 754 (1945).

18. *See* Koehler & Shaviro, *supra* note 11.

19. Cohen, *Subjective Probability and the Paradox of the Gatecrasher,* 1981 ARIZ. ST. L.J. 627, 633. *See also* Brilmayer & Kornhauser, *Review: Quantitative Methods and Legal Decisions,* 46 U. CHI. L. REV. 116 (1978).

20. Koehler & Shaviro, *supra* note 11, offer a cogent critique of the anti-Bayesian position.

21. "From the perspective of verdict accuracy, it is unjustifiable to ignore, by reason of its unspecificity, the best available base rates." *Id.*

22. For example, they fear that allowing a party to prevail based solely on "naked statistical evidence" may lead to strategic behavior, in which case-specific evidence is suppressed by the party favored by the base rate, and other "feedback effects" involving opportunistic responses to the knowledge that such evidence will be used. These concerns would generally not apply where base rates are offered in conjunction with more particularized evidence; hence most of these commentators argue that "directly relevant" base rates should be admitted where they are not "naked." But even "non-naked" base rates are sometimes excludable on policy grounds. For

The use in trials of "indirectly relevant" base rates has been more common. Base rates of this type may be used to show the weight that should be accorded a piece of forensic evidence. For example, where the perpetrator and the defendant are shown to have the same blood type, the prosecutor often presents statistics on the frequency of that blood type in a relevant population to prove that the match is unlikely to have occurred by chance.[23] Statistics on the percentage of the population possessing a given blood group are routinely admitted in evidence in most states.[24] Statistics also have been admitted in conjunction with forensic evidence showing a match between samples of hair,[25] glass and paint,[26] fibers,[27] particles,[28] and teeth marks.[29]

2. *Sources of Base Rate Statistics.* Because the value of associative evidence depends, in part, on the rarity of the characteristic or trace that links the defendant to the crime, forensic scientists have devoted much effort in recent years to studying the rarity of characteristics likely to be important in criminal identification. Efforts are being made in the United States and the United Kingdom to collect and store frequency data in a central location.[30] Base rate statistics literature is increasingly finding its way into criminal trials.[31]

The studies in this area are of two types. One type of study simply reports the relative frequency of various characteristics or traces in a sample drawn from some population. Most studies on the frequency of serological

---

example, it would be inconsistent with the constitutionally based presumption of innocence for a prosecutor to present base rate evidence showing that a high percentage of defendants in similar cases are convicted in order to show that a particular defendant is likely to be guilty.

23. P. GIANNELLI & E. IMWINKELRIED, *supra* note 2, at 605-31; N. Miller, *supra* note 10.

24. P. GIANNELLI & E. IMWINKELRIED, *supra* note 2, at 589-92, 605-38; Jonakait, *supra* note 6, at 639; Annotation, *Admissibility, Weight and Sufficiency of Blood Grouping Tests in Criminal Cases*, 2 A.L.R. 4th 500 (1980). The major exception, for a number of years, was New York, where in 1970 the state's highest appellate court found error in the admission of evidence that a defendant and perpetrator shared a blood type (Type A) found in 40% of the population. People v. Robinson, 27 N.Y.2d 864, 265 N.E.2d 543, 317 N.Y.S.2d 19 (1970); *accord* People v. Macedonio, 42 N.Y.2d 944, 366 N.E.2d 1355, 397 N.Y.S.2d 1002 (1977). The *Robinson* court found that this evidence was "of no probative value in the case against defendant in view of the large proportion of the general population having blood of this type" and expressed concern that jurors might give such evidence more weight than it deserves. 27 N.Y.2d at 865, 265 N.E.2d at 543, 317 N.Y.S.2d at 20. The court subsequently admitted evidence of a match on a set of blood group markers found in 1% of the population, however, arguing that "the relative rarity of the . . . type of blood relegates arguments as to remoteness to the realm of weight rather than admissibility." *In re* Abe A, 56 N.Y.2d 288, 299, 437 N.E.2d 265, 271, 452 N.Y.S.2d 6, 12 (1982). Then, in 1985, the court disavowed *Robinson*, citing near unanimous opposition to the holding by commentators, *e.g.*, MCCORMICK ON EVIDENCE § 205, at 619 (3d ed. 1984), and other courts, and recognizing that while proof of a match on a common characteristic has little value by itself, such evidence "may acquire great probative value when considered cumulatively." People v. Mountain, 66 N.Y.2d 197, 203, 486 N.E.2d 802, 805, 495 N.Y.S.2d 944, 947-48 (1985).

25. *E.g.*, United States *ex rel.* DiGiacomo v. Franzen, 680 F.2d 516 (7th Cir. 1982).

26. State v. Menard, 331 S.W.2d 521 (Mo. 1960).

27. People v. Trujillo, 32 Cal. 2d 105, 194 P.2d 681 (1948).

28. People v. Coolidge, 109 N.H. 403, 260 A.2d 547 (1969), *rev'd on other grounds*, 403 U.S. 443, *reh'g denied*, 404 U.S. 874 (1971).

29. State v. Garrison, 120 Ariz. 255, 585 P.2d 563 (1978).

30. Saferstein, *Criminalistics—A Look Back at the 1970s, A Look Ahead to the 1980s*, 24 J. FORENSIC SCI. 925 (1979).

31. P. GIANNELLI & E. IMWINKELRIED, *supra* note 2, at 423-504, 605-31.

characteristics are of this type.[32] Studies have also been undertaken to determine the frequency of various types of paint,[33] glass,[34] fibers,[35] and soil;[36] the frequency of wear characteristics in men's footwear; and the frequency with which blood and semen stains[37] and glass and paint particles[38] are found on outer clothing and shoes.

A second type of study looks directly at the likelihood of a coincidental match between samples rather than the proportion of various characteristics in the population. Two Canadian forensic scientists, for example, conducted a study in which thousands of hairs from 100 unrelated individuals were compared under a microscope with respect to twenty-three different characteristics, such as color, pigment distribution, diameter, and scale count. In approximately one of every 4500 comparisons of hairs from different individuals, a match was found with respect to all twenty-three characteristics. Hence, the researchers reported that the chances of a coincidental match between two unrelated individuals on a microscopic comparison of scalp hairs is one in 4500.[39] Based on a subsequent study, the probability of a coincidental match for pubic hairs was estimated to be one in 800.[40] Data on the probability of coincidental match have also been collected on dental characteristics, DNA print patterns,[41] and even lipstick—the finding that there

---

32. *E.g.*, Grunbaum, Selvin, Myhre & Pace, *Distribution of Gene Frequencies and Discrimination Probabilities of 22 Human Blood Genetic Systems in Four Racial Groups*, 25 J. FORENSIC SCI. 428 (1980); Steadman, *Blood Group Frequencies of Immigrant and Indigenous Populations for South East England*, 25 J. FORENSIC SCI. SOC'Y 95 (1985).

33. Ryland, Kipec & Somerville, *The Evidential Value of Automobile Paint, Part II: Frequency of Occurrence of Topcoat Color*, 26 J. FORENSIC SCI. 64 (1981).

34. Fong, *Value of Glass as Evidence*, 18 J. FORENSIC SCI. 398 (1973).

35. Home & Dudley, *A Summary of Data Obtained From a Collection of Fibres From Casework Materials*, 20 J. FORENSIC SCI. SOC'Y 253 (1980).

36. P. GIANNELLI & E. IMWINKELRIED, *supra* note 2, at 1080-86.

37. Owen & Smalldon, *Blood and Semen Stains on Outer Clothing and Shoes Not Related to Crime: Report of a Survey Using Presumptive Tests*, 20 J. FORENSIC SCI. 391 (1975).

38. Pearson, May & Dabbs, *Glass and Paint Fragments Found in Men's Outer Clothing-Report of a Survey*, 16 J. FORENSIC SCI. 283 (1971).

39. Gaudette & Keeping, *An Attempt at Determining Probabilities in Human Scalp Hair Comparison*, 19 J. FORENSIC SCI. 599, 604 (1974). This study has been heavily criticized. *See, e.g.*, Barnett & Ogle, *Probabilities and Human Hair Comparison*, 27 J. FORENSIC SCI. 272 (1982); Note, *Splitting Hairs in Criminal Trials: Admissibility of Hair Comparison Probability Estimates*, 1984 ARIZ. ST. L.J. 521.

40. Gaudette, *Probabilities and Human Pubic Hair Comparisons*, 21 J. FORENSIC SCI. 514, 517 (1976).

41. DNA typing procedures produce "prints" consisting of a pattern of bands somewhat analogous to a supermarket bar code. *See generally* Thompson & Ford, *DNA Typing: Admissibility and Weight of the New Genetic Identification Tests*, 75 VA. L. REV. 45 (1989). Individuals differ in the position of the bands on their print. To determine the likelihood of a coincidental match between DNA prints of two unrelated individuals, where the prints in question had fifteen distinct bands, Jeffreys, Wilson, and Thein made prints of twenty unrelated individuals, laid the prints side-by-side, and counted the number of instances in which a band in one print was matched by a band in the adjacent print. Jeffreys, Wilson & Thein, *Individual-specific "Fingerprints" of Human DNA*, 316 NATURE 76 (1985). Overall, about 21% of the bands were matched by a band on an adjacent print. Accordingly, Jeffreys and colleagues concluded that there is about a 21% chance that a given band in a DNA print will be matched by a band in the print of an unrelated individual. To calculate the probability that two unrelated individuals will match on all fifteen bands, the researchers applied the product rule and concluded that the probability of a coincidental match on fifteen bands is approximately 0.21^{15} or one in thirty billion.

was a one-in-707 chance that samples of lipstick chosen at random would match was reportedly used as evidence in two criminal cases.[42]

3. *Drawing Conclusions from Base Rate Statistics: Potential Problems and Complexities.* Although base rate statistics are often highly relevant and informative, their probative value depends on a number of factors. To deal competently with base rate statistics, jurors must take these factors into account. However, several potential problems with base rate statistics may make these statistics misleading if jurors fail to understand their defects.

One problem is that base rates may be derived from inaccurate or uninformative data. For example, regional or geographic variations in the frequency of various types of fibers, paints, or soil types may render data based on samples in one area unrepresentative of frequencies in other areas. Kaye notes that for blood typing "the sampling is sufficiently extensive and variegated that the statistic should be reliable" while for other types of forensic evidence "scientific knowledge of the population parameter usually is . . . more sketchy."[43] Kaye favors admitting into evidence even these "sketchy" statistics on the grounds that they "can provide some clue as to the frequency of trace evidence in the population at large."[44] To draw appropriate conclusions from such statistics, however, jurors must appreciate the implications of sampling variability and sample bias. Research on the judgmental ability of untrained individuals raises some doubts about jurors' competence in this area.[45]

A second potential problem concerns computing the frequency of the joint occurrence of multiple characteristics. Where forensic evidence shows a match on several characteristics—for example, three distinct genetic markers in blood—forensic experts typically present statistics on the joint frequency of those characteristics—for example, the proportion of the population that possesses all three. These statistics are typically estimated from data on the frequency of the individual markers rather than from direct observation. Forensic scientists operate on the assumption that the genetic markers they use are independent of one another.[46] Accordingly, they compute the frequency of a combination of genetic markers by applying the product rule, which holds that the frequency of several independent events occurring simultaneously may be determined by simply multiplying the probability that each event will occur. If a match is found on three markers occurring in 5, 10,

---

42. Barker & Clarke, *Examination of Small Quantities of Lipsticks,* 12 J. FORENSIC SCI. SOC'Y 449 (1972).

43. Kaye, *supra* note 10, at 162. Suppose, for example, that the defendant is linked to the crime by the presence of red clay soil on his boots matching the soil at the scene of the crime. Statistics on the prevalence of red clay soil will be relevant only if they apply to the areas where the defendant might have been. The frequency of a given type of soil in a study in California may not be representative of the frequency of that soil type in Georgia.

44. *Id.*

45. R. NISBETT & L. ROSS, HUMAN INFERENCE: STRATEGIES AND SHORTCOMINGS IN JUDGMENT 77-88, 256-61 (1980).

46. Independence presumes that possession of a given phenotype on one marker system is not associated with the possession of any particular phenotype on any other system.

and 20 percent of the population respectively, they typically report to the jury that the percentage of the population possessing all three markers is .05 x .10 x .20 = .001 or 0.1 percent.

The use of the product rule is well accepted for computing the frequency of protein markers in blood because there is extensive evidence that these markers are independent of one another.[47] Use of the product rule is inappropriate, however, where the characteristics are not independent. If the product rule is applied to events which are partially dependent, it may significantly underestimate the frequency of their joint occurrence. A number of courts have refused to allow computations based on the product rule unless the proponent can show that the characteristics being multiplied are independent,[48] but there are exceptions.[49] Hence, jurors sometimes must evaluate whether the use of the product rule to compute the statistics in the case was appropriate; where it is not, they must somehow take that problem into account.

In some areas the courts must decide whether to admit computations based on the product rule in the face of scientific uncertainty about the independence of the relevant characteristics. For example, considerable controversy has developed recently over the use of the product rule to compute the frequency of so-called DNA fingerprints when conclusive data have not appeared demonstrating that the genetic markers that make up the print are independent of one another.[50] While some courts have excluded the DNA statistics on this ground,[51] others have admitted the statistics, holding that any dispute over their accuracy should go to weight rather than admissibility.[52] Consequently, in a number of cases in which DNA evidence was presented, the issue of the independence of DNA markers has been thrown to the jury.[53]

---

47. P. GIANNELLI & E. IMWINKELRIED, *supra* note 2, at 605.

48. *E.g.*, People v. Collins, 68 Cal. 2d 319, 438 P.2d 33, 66 Cal. Rptr. 497 (1968).

49. *E.g.*, State v. Garrison, 120 Ariz. 255, 585 P.2d 563 (1978) (product rule applied to determine the likelihood of matching bite marks in absence of demonstration of independence of matching features).

50. In hearings on the admissibility of DNA statistics, some scientists have argued forcefully that the relevant markers ought to be independent, but others have questioned whether this opinion should be accepted in the absence of data to demonstrate its truth. Lander, *DNA Fingerprinting on Trial*, 339 NATURE 501, 503-04 (1989); Thompson & Ford, *Is DNA Fingerprinting Ready for the Courts?*, 125 NEW SCIENTIST 38, 43 (1990).

51. *See, e.g.*, State v. Schwartz, 447 N.W.2d 422 (Minn. 1989) (ruling DNA test results by a commercial laboratory inadmissible based, in part, on the laboratory's failure to comply with a request for information about its data base that would have allowed the defense to assess the independence of the genetic markers); State v. Pennell, IN88-12-0051 (Del. Super. Ct. Sept. 20, 1989) (1989 WL 167430) (ruling statistics reported by the same lab inadmissible based, in part, on the failure of the laboratory to produce adequate documentation for its claim that the markers are independent).

52. *E.g.*, People v. Wesley, 140 Misc. 2d 306, 533 N.Y.S.2d 643 (Albany County Ct. 1988). *See* Thompson & Ford, *supra* note 41, at 81-87 for a thorough discussion of this issue.

53. The jurors' evaluation of the scientific evidence concerning independence may be crucial in such cases. In a hearing on the admissibility of DNA evidence in People v. Axell, CR23911 (Ventura Super. Ct. May 22, 1989), experts for the prosecution, who assumed independence and applied the product rule, testified that the frequency of defendant's DNA print was approximately one in 6 billion; experts called by the defense challenged the assumption of independence and testified that if

A third potential problem arises because the probative value of associative evidence (that is, evidence linking the defendant to the crime by showing a match) sometimes depends on how the defendant has been selected as a suspect. Where the defendant is selected for reasons unrelated to the likelihood of a match linking him to the crime, the frequency of the matching characteristic in the population from which the defendant was drawn is a reasonable estimate of the likelihood of a coincidental match. In other words, the frequency would provide the rough likelihood that the defendant would have the characteristic if innocent. If one person in 100 has the blood type on which the defendant and perpetrator match, for example, the probability is 1 percent that the defendant would happen by chance to have this blood type, if he is innocent. The frequency of the characteristic thus provides an index of the likelihood of a coincidental match. Where the defendant is selected for reasons that render him more or less likely than most people to have the matching characteristic, however, the frequency of the characteristic in the population does not reflect the likelihood of a coincidental match. If jurors fail to appreciate this fact, they may misjudge the likelihood of a misidentification and thereby over- or underestimate the value of the associative evidence.

As an illustration of this selection phenomenon, imagine a hypothetical murder case in which a long red hair (presumably from the killer) is found clenched in the fist of the victim. The police apprehend the defendant because he lives near the victim and has long red hair. Microscopic analysis reveals a match between hair samples taken from the defendant and the hair in the victim's hand with respect to thirteen distinguishable qualities such as color, length, and coarseness. A forensic expert reports a research study showing that the likelihood of a coincidental match between two hairs randomly drawn from different people is one in 4500.[54] What is the likelihood the defendant's hair would happen to match, as it does, if he is not the source of the hair in the victim's fist? Assuming one trusts the forensic report and the research, one is tempted to conclude the likelihood of a coincidental match is one in 4500, but this is demonstrably wrong. The defendant was selected, at least in part, because he has long red hair; thus, the likelihood that his hair would match the hair in the victim's fist, even though he is innocent, is undoubtedly far higher than the likelihood a randomly drawn individual would match. The figure of one in 4500 greatly underestimates the likelihood of a coincidental match in this case.

Perhaps the most blatant example of a selection effect occurred in the infamous case of *People v. Collins.*[55] A robbery was committed by a black man with a beard and a mustache and a blond woman with a ponytail, who both fled in a yellow convertible. Defendants were a couple fitting this description

---

the markers are not independent the frequency of the defendant's DNA print could be as high as one in 50. *See id.* testimony of expert Lawrence Mueller.

54. For an actual case with similar facts, see State v. Carlson, 267 N.W.2d 170 (Minn. 1978).

55. 68 Cal. 2d 319, 438 P.2d 33, 66 Cal. Rptr. 497 (1968).

apprehended in the vicinity of the robbery. To bolster a shaky eyewitness identification, the prosecutor called to the stand a college mathematics instructor and asked him to compute the frequency in the general population of a couple having various characteristics possessed by defendants, for example, a man with a beard, a man with a mustache, a blond woman, a woman with pony tail, and a yellow convertible. The prosecutor supplied "conservative" estimates of the frequency of these characteristics, and the mathematician, applying the product rule, multiplied the frequencies together to obtain a joint frequency of one in twelve million.

The California Supreme Court reversed the resulting conviction, concluding that it was error to admit the frequency estimate when that figure was not only likely to be overvalued by the jury but was computed on the unsupported assumption that the characteristics were independent.[56] Although the evidence in *Collins* certainly suffers from these problems, the major difficulty with the one in twelve million figure is that it purports to measure the probability of a coincidental match between the defendants and perpetrators but in fact does nothing of the sort. Assume that the mathematician was correct in computing the frequency of a couple with the stated characteristics to be one in twelve million. This figure might reflect the likelihood of a coincidental match if the defendants had been selected for reasons unrelated to the likelihood that they would possess the "matching" characteristics. However, it clearly does not reflect the likelihood of a coincidental match in the actual case, where defendants were selected precisely because they possessed the relevant characteristics. The likelihood of a match if these defendants are innocent is not one in twelve million, it is one in one; that is to say that it is certain.[57]

Courts and commentators have generally distinguished statistical testimony like that in *Collins* from testimony regarding the frequency of characteristics identified by forensic tests.[58] As the long red hair example illustrates, however, the same sort of selection effect that renders the *Collins* statistics problematic may also operate in cases involving forensic evidence, albeit in a more subtle manner.[59] As a result, base rate statistics in these cases

---

56. *Id.* at 327, 438 P.2d at 38, 66 Cal. Rptr. at 502.

57. If correct, the one in 12 million figure is not irrelevant; it suggests a low likelihood that another such couple would be found in the area and thus supports the conclusion that defendants are the guilty couple. In a city of several million people, however, the likelihood of finding two couples matching defendants' description might be reasonably high. In *Collins* the probability of a second couple in the Los Angeles area matching the characteristics of the perpetrators was computed to be .40. *Id.* at 333-35, 438 P.2d at 42-43, 66 Cal. Rptr. at 506-07.

58. "[E]xpert testimony [in *Collins*] merely told the jury how to think—how to evaluate the fact that the Collins's were in the vicinity of the crime . . . . Such 'no evidence' cases do not dictate the outcome when meaningful statistical evidence permits a computation of the probability of a coincidental misidentification." Kaye, *supra* note 10, at 167.

59. This problem is particularly likely to occur in cases in which forensic evidence shows a match on some observable characteristic (e.g., hair, paint, fibers). In such cases the suspect is often selected in a manner that renders him more likely than most people to "match" with regard to the relevant characteristic. On the other hand, where the "match" is on a characteristic that is not easily observed (e.g., blood type), it is less likely (though not inconceivable) that the characteristic played a role in

will not always reflect the probability of a coincidental misidentification. A key issue with regard to jury competence is whether jurors appreciate and take into account such phenomena.

4. *Presentation of Base Rate Statistics.* The difficulty jurors face in interpreting base rate statistics is compounded because these statistics may be presented in several different ways. Where forensic evidence shows a match, experts in most cases simply report the frequency of the matching characteristic or set of characteristics in a reference population, using percentages or incidence rates. It is common for experts to report, for example, that "type O blood is found in 44 percent of Caucasians" or that "among Hispanics, three persons in 100 possess both ABO type O and PGM-Type 2 enzyme markers." But other formulations are sometimes used. In *People v. Harbold,*[60] for example, serologist Mark Stolorow testified regarding the probability of a coincidental match between *two* individuals: " 'the chances of selecting any two people at random from the population and having them accidently [sic] have identical blood types in each one of these factors is less than one in 500, that is, what we call the probability of an accidental match is less than one in 500.' "[61] While it is tempting to assume that the frequency of a characteristic is equivalent to what Stolorow calls the probability of an accidental match, this is not the case. The probability of an accidental match actually equals the square of the frequency. If 10 percent of the population has blood type B, for example, the probability of selecting two people at random and finding they both have type B is .10 x .10 = .01, or one in 100. Thus, the impressive sounding conclusion that there is one chance in 500 of an accidental match is equivalent to the somewhat less impressive statement that the matching characteristics would be found in approximately one person in twenty-two.[62] Whether jurors appreciate this distinction between frequency and probability of accidental match is unclear.

To complicate matters, the probability of an accidental match on a genetic marker system is not necessarily equivalent to the probability of an accidental match on a particular marker. Consider, for example, the frequency of markers in the well-known ABO system. The probability that two randomly chosen individuals will share the same ABO type (not taking into account which type it is) is approximately 38 percent, while the probability they will both share a specific type ranges from 19 percent for type A to 0.16 percent for type AB.[63] Hence, it is crucial to know whether one is referring to an

---

the selection of the suspect and therefore less likely that the probability of a coincidental match diverges from the frequency of the matching characteristic.

60. 124 Ill. App. 3d 363, 464 N.E.2d 734 (1984).

61. *Id.* at 381, 464 N.E.2d at 748.

62. It is possible that the underlying data Stolorow wished to report actually indicated a frequency of one in 500 for blood characteristics in question and that Stolorow mistakenly assumed that this frequency was equivalent to the probability of an accidental match. The appellate opinion leaves this point unclear.

63. The frequency of ABO types and the probability of an "accidental match" with respect to each type (and any type) is shown in the following table:

accidental match on a system or on a specific marker within that system. But it may be difficult to tell from testimony such as Stolorow's[64] which of these terms is being reported. In addition, several different terms might be described in language that sounds similar. Suppose, for example, that the defendant and perpetrator share blood type AB. An expert might be quite correct in stating any of the following: (1) the probability of a match between two randomly chosen people in this system is 38 percent; (2) the probability of match between two randomly chosen people on this marker is 0.16 percent and (3) the probability a randomly chosen individual will have this marker is 4 percent. Whether jurors can and will appreciate the differences among these similar-sounding formulations is difficult to predict, but clearly this is an important issue underlying jury competence to deal with such data.

The difficulty jurors face in correctly interpreting base rate statistics becomes even greater when those statistics are reported in a misleading manner. Appellate opinions provide examples of erroneous and misleading statistical presentations. One error is to use base rate data to characterize the probability that *someone other than the defendant* was the source of an evidentiary sample. In *State v. Carlson,*[65] for example, forensic hair expert Barry Gaudette testified that there was "a 1-in-800 chance that the pubic hairs stuck to the victim were not Carlson's and a 1-in-4,500 chance that the head hairs found on the victim were not Carlson's."[66]

The problem with Gaudette's testimony is that it draws conclusions about the probability that the hairs "were not Carlson's." Obviously the hairs are either Carlson's or someone else's, so if Gaudette is correct in reporting one chance in 4500 the hairs were not Carlson's, it follows that there is a 4499 in 4500 chance they were Carlson's. Gaudette cannot properly testify to this effect, however, because conclusions about the likelihood that the hairs were Carlson's cannot be drawn from the forensic evidence alone. If one person in 4500 would have hair matching that found on the victim, thousands of people, besides Carlson, must have such hair. To determine the likelihood the hair

---

TABLE 1
FREQUENCY OF ABO TYPES

Type	Frequency in Population	Probability of "Accidental Match"
O	.44	.19
A	.42	.18
B	.10	.01
AB	.04	.0016
Overall (any type)		.3816

Because the frequency of type O is .44, the probability that two individuals drawn at random will both be type O is $.44^2 = .19$; that both will be type A is $.42^2 = .18$; that both will be type B is $.10^2 = .01$; and that both will be AB is $.04^2 = .0016$. Accordingly, the probability that two individuals will both have the *same* ABO type is .19 + .18 + .01 + .0016 = .3816. Selvin & Grunbaum, *Genetic Marker Determination in Evidence Bloodstains: The Effect of Classification Errors on Probability of Non-discrimination and Probability of Concordance,* 27 J. FORENSIC SCI. SOC'Y 57 (1987).

64. *See supra* text accompanying note 60.

65. 267 N.W.2d 170 (Minn. 1978).

66. *Id.* at 173. *See* Gaudette & Keeping, *supra* note 39, at 605; Gaudette, *supra* note 40, at 517.

was Carlson's, we must consider whether the hair is more likely to have come from Carlson than from one of the thousands of other people with matching hair. One cannot make this determination, however, without evaluating the other evidence against Carlson, and therein lies the problem. Gaudette was in no position to evaluate the strength of the other evidence against Carlson and had no business doing so in any case. Hence, his opinion about the likelihood the hair was not Carlson's is not only unwarranted, but it also invades the province of the jury in a particularly insidious way, because the jurors are unlikely to realize that Gaudette's statistics rest, in part, on assumptions about the strength of evidence unrelated to the hair.

B. Error Rate Statistics

A second type of statistical formulation jurors may encounter in criminal trials concerns the rate of error in forensic tests.

1. *Sources of Error Rate Statistics.* The major source of error rate statistics is proficiency testing. A typical proficiency test is a blind trial in which forensic analysts are asked to classify specimens of known origin in order to check their accuracy. Studies of this type have provided considerable evidence that forensic testing is less than perfectly reliable. In 1974, the Law Enforcement Assistance Administration ("LEAA") of the Justice Department undertook a large-scale study of the proficiency of crime labs in the United States.[67] Between 235 and 240 laboratories took part in the blind trial, and the results, in the words of one commentator, were "shocking."[68] Over 20 percent of the labs inaccurately or incompletely identified samples of hair and paint, while over 30 percent of the labs inaccurately or incompletely identified glass and soil samples. Furthermore, less than 30 percent accurately or completely identified one sample of blood.

Error rates in blood typing are probably the best documented. Nationwide proficiency tests were conducted by the LEAA in 1975 and by the Forensic Sciences Foundation between 1978 and 1983.[69] Hundreds of blood samples of known type were sent to crime laboratories, which were asked to classify the samples while remaining "blind" to their type. The rate of classification errors varied among the different genetic marker systems used, ranging from 0.3 percent for the Adenosine deaminase system to over 6 percent for the familiar ABO system.[70] Because crime labs typically "type" blood on up to eight different systems, the likelihood of an error cumulates. Based on the proficiency test data, Selvin and Grunbaum concluded that

---

67. J. PETERSON, E. FABRICANT & K. FIELD, CRIME LABORATORY PROFICIENCY TESTING RESEARCH PROGRAM—FINAL REPORT TO U.S. DEPT. OF JUSTICE (1978).

68. Imwinkelried, *A New Era in the Evolution of Scientific Evidence: A Primer on Evaluating the Weight of Scientific Evidence*, 23 WM. & MARY L. REV. 261, 268 (1981).

69. G. SENSABAUGH & D. NORTHEY, WHAT CAN BE LEARNED FROM THE PROFICIENCY TRIALS? AN ANALYSIS OF THE ELECTROPHORETIC TYPING RESULTS, 1975-83 PROCEEDINGS OF THE INTERNATIONAL SYMPOSIUM ON THE FORENSIC APPLICATIONS OF ELECTROPHORESIS 184 (1986). *See also* Selvin & Grunbaum, *supra* note 63, at 59.

70. *See* Selvin & Grunbaum, *supra* note 63, at 59.

"blood group evidence employing [all] eight systems will be incorrect in some way in excess of 20 percent of the time."[71]

Blind trials have also been conducted recently to determine the proficiency of three commercial laboratories doing DNA typing of forensic samples. Asked to test approximately fifty unknown blood and semen samples, two of the labs had a "false-positive"; that is, they mistakenly declared a match in an instance where the pair of samples being compared actually came from different people.[72]

In addition to proficiency tests administered by outside agencies, many forensic laboratories engage in routine internal proficiency testing. These studies are potentially another source of data on error rates.

2. *Drawing Conclusions from Error Rate Statistics: Potential Problems and Complexities.* Like base rate statistics, error rate statistics are often highly relevant and informative, but must be interpreted with care because their probative value depends on a variety of factors. Jurors may be misled by such statistics if they fail to take these factors into account. However, the weight jurors should give to these factors is often unclear in a given instance. One important factor jurors should consider is whether aggregate data produced by proficiency testing of many labs accurately represent the rate of error in any particular lab. Large-scale proficiency tests such as those of the LEAA, for example, typically involve a number of laboratories, which are not individually identified. Consequently, it has been suggested that errors on proficiency tests result from inadequate training of a minority of analysts and tend to cluster in a few "bad" labs. As a result, aggregate data on error rates from proficiency tests greatly overstate the likelihood of an error by a competent analyst at a "good" lab.[73] A second consideration is whether error rates on proficiency tests reflect error rates in routine casework. In most proficiency tests, the laboratory personnel know they are being tested and may therefore be on their best behavior. Finally, jurors must consider whether error rates in the past predict the rate of errors in the future. One purpose of proficiency testing is to detect inadequacies in laboratory procedure that may contribute to error. Laboratories which have been "caught" making errors on proficiency tests sometimes change procedures in an effort to improve future performance.

---

71. *Id.* at 61.

72. One of the laboratories also had three false negatives, although these errors were initially covered up by the agency doing the testing. Ford & Thompson, *A Question of Identity: Some Reasonable Doubts About DNA "Fingerprints,"* THE SCIENCES, Jan./Feb. 1990, at 37, 41. The third lab had no false positives, but was unwilling to make a call on 14 of the samples and, in a follow-up study, twice failed to detect that mixed stains contained the DNA of two individuals. M. Graves & M. Kuo, DNA: A Blind Trial Study of Three Commercial Testing Laboratories (Feb. 1989) (paper presented at the meeting of the American Academy of Forensic Sciences, Las Vegas, Nev.).

73. G. SENSABAUGH & D. NORTHEY, *supra* note 69. On the other hand, aggregate data may underestimate the rate of error at a "bad" lab. Hence, jurors must evaluate whether a particular lab is more or less error-prone than average to draw appropriate conclusions from aggregate error rate data.

A more subtle issue concerns the connection between the error rate of a test and the likelihood of a result that would falsely incriminate an innocent defendant. Not every error is of the sort that incriminates; therefore, the error rate of a test is not necessarily equivalent to the likelihood that an innocent person would be falsely incriminated. Suppose a bloodstain found at the scene of a crime is tested to see whether it matches that of a suspect, who is known to have type A blood. Assuming the stain is actually type O, the suspect will be falsely incriminated if the stain is misclassified as type A, but not if it is misclassified as type B or AB. The key issue, then, is not the overall error rate of the test but the rate at which types other than A are misclassified as type A. This error rate is sometimes called the false-positive rate for A. If errors are distributed randomly across the different blood types, the false-positive rate for a particular phenotype, such as type A, will be lower than the overall error rate for the test because only a subset of errors will be false-positives. If errors do not occur at random, however, the false-positive rate may be either higher or lower than the error rate. Suppose, for example, that there is an error rate of 6 percent in ABO typing, but that all of the errors occur when type O is misclassified as type A. In this instance, the false-positive rate for A would be higher than 6 percent while the false-positive rate for O, B, and AB would be zero. Although the connection between the error rate and the false-positive rate is not obvious in many instances, it is common for forensic scientists to report proficiency data in a form that allows inferences only about the overall error rate. The ability of jurors to draw appropriate conclusions from such data is open to question.

3. *Presentation of Statistics.* Although error rate statistics of this type are available in the published literature, they apparently are presented infrequently in criminal trials. A group of fifty forensic scientists surveyed by Miller[74] reported that they rarely presented data on error rates in court. Error rate data may be presented infrequently, in part, because attorneys are simply unfamiliar with it. Forensic scientists who are called to present the findings of forensic tests are unlikely to be examined extensively about error rates by the proponent of the evidence. While lawyers who cross-examine forensic experts are advised to probe extensively regarding the reliability of the procedure,[75] experts may be unwilling to phrase their estimates of error rates in statistical terms, or even to admit the possibility of error. In *State v. Spencer*,[76] for example, an expert responded to questions about the reliability of neutron activation analysis, a notoriously unreliable procedure,[77] by declaring "[t]here is no unreliability as far as we are concerned."[78] To challenge or even to detect such overstatements may require the attorney to

74. N. Miller, *supra* note 10.
75. E. IMWINKELRIED, THE METHODS OF ATTACKING SCIENTIFIC EVIDENCE (1982).
76. 298 Minn. 456, 216 N.W.2d 131 (1974).
77. George, *Statistical Problems Relating to Scientific Evidence*, in SCIENTIFIC AND EXPERT EVIDENCE 128 (E. Imwinkelried 2d ed. 1981).
78. *Spencer*, 298 Minn. at 459, 216 N.W.2d at 134.

seek the assistance of another expert, whose services may be difficult to obtain or beyond the financial means of the defendant.[79]

Error rate statistics may also be used sparingly due to confusion about the meaning of errors on proficiency tests in relation to the reliability of a given procedure. Like data on the frequency of trace characteristics, error rate data are sketchy. Nevertheless, like frequency data, they provide some clue as to the likelihood of a wrong result. Whether jurors draw the appropriate conclusions from such data depends on their ability to appreciate the many subtle ways in which such statistics may be misleading. Whether lay individuals are capable of this task is an unexplored issue.

III

JURORS' USE OF STATISTICAL EVIDENCE: MAJOR CONCERNS AND RESEARCH STRATEGIES

Social scientists have recently begun studying whether lay individuals can draw appropriate conclusions from statistical evidence of the type presented in criminal trials. The standard research strategy is the jury simulation study, in which individuals read summaries of evidence and are asked to judge the guilt of a hypothetical criminal defendant. The nature of the evidence can be varied to determine, for example, how variations in the manner in which statistical evidence is presented affect people's judgments.

A major goal of the social scientists is to assess whether people use statistical evidence appropriately. To answer this question there must, of course, be some standard of appropriateness against which people's judgments can be compared. The benchmark used by researchers has been a set of mathematical models based on Bayes' theorem.[80] These models can specify how much one should revise one's estimate of a criminal suspect's probability of guilt after receiving forensic evidence accompanied by statistics.[81] Assuming a juror initially thinks that there is a 20 percent chance

---

79. M. Saks & R. Van Duizend, *supra* note 1, at 89.

80. For a general discussion of the use of Bayes' theorem to model legal judgments, see R. LEMPERT & S. SALTZBURG, A MODERN APPROACH TO EVIDENCE 148-53 (1st ed. 1977); Kaplan, *Decision Theory and the Factfinding Process*, 20 STAN. L. REV. 1065 (1968); Kaye, *What is Bayesianism? A Guide for the Perplexed*, 28 JURIMETRICS J. 161 (1988); Lempert, *Modeling Relevance*, 75 MICH. L. REV. 1021 (1977); Schum & Martin, *Formal and Empirical Research on Cascaded Inference in Jurisprudence*, 17 LAW & SOC'Y REV. 105 (1982).

81. Where H and $\bar{H}$ designate the suspect's guilt and innocence respectively, and E designates evidence of a match between the suspect and perpetrator on some characteristic, Bayes' theorem states:

$$p(H/E) = p(H)p(E/H) / [p(H)p(E/H) + p(\bar{H})p(E/\bar{H})]$$

The term p(H) is read "the probability of H"; this term is called the prior probability and reflects the decisionmaker's initial estimate of the probability the suspect is guilty in light of everything that is known before receiving E. The term p(H/E) is read "the probability of H given E"; this term is called the posterior probability and indicates what the decisionmaker's revised estimate of probable guilt should be in light of everything known after receiving E. The formula indicates that the evidence of a match, E, should cause the decisionmaker to revise his opinion of the suspect's guilt to the extent p(E/H) differs from $p(E/\bar{H})$. If the suspect and perpetrator are certain to match if the suspect is guilty, p(E/H) = 1.00. If an innocent suspect is no more likely than anyone else to possess

a particular suspect is guilty, for example, the models can tell him how much he should revise this estimate after receiving additional evidence that the suspect has genetic markers matching those found in the perpetrator's blood and that those markers occur in only 5 percent of the population.[82]

A major research strategy, then, is to determine whether people revise their judgments to the extent that Bayes' theorem dictates after receiving statistical evidence. Researchers are not particularly concerned with whether people's judgments correspond exactly to the predictions of Bayesian models. No one argues that jurors must be perfect intuitive Bayesians to be considered competent to deal with statistical data. Instead, the research has focused on three major concerns: first, whether people evaluate statistical evidence using inappropriate judgmental strategies that could lead to serious errors in estimating the value of the evidence, and therefore to dramatic divergence of human judgment from Bayesian norms; second, whether people are insensitive to important statistical variations in evidence and therefore fail to distinguish strong and weak evidence as effectively as the Bayesian models suggest they should; finally, whether people are insensitive to nonstatistical factors that affect the value of statistical evidence, such as partial redundancies between statistical evidence and other evidence in the case. By comparing actual judgments to those specified by Bayesian models, one can test sensitivity to such factors. In the sections that follow, each of these concerns will be discussed in some detail in light of the available empirical research and commentary.

A. Inappropriate Judgmental Strategies: Fallacious Interpretation of Statistical Evidence

1. *The Prosecutor's Fallacy.* One of the major concerns that has been raised about population proportions and statistics on the probability of a match is that jurors will mistakenly assume these statistics directly measure the probability of the defendant's innocence. A juror who hears that the defendant and perpetrator share a blood type found in 10 percent of the population, for example, may reason that there is only a 10 percent chance that the defendant would happen to have this blood type if innocent. The juror may then jump to the mistaken conclusion that there is therefore a 90 percent chance that the defendant is guilty. Thompson and Schumann,[83] who

---

the matching characteristic, $p(E/\bar{H})$ is equal to the frequency of the matching characteristic in the population from which the suspect was drawn.

82. The prior probability of guilt, p(H) is equal to .20, and because the suspect must be either guilty or innocent, $p(\bar{H}) = .80$. Because the suspect is certain to have the perpetrator's genetic markers if he is the perpetrator, $p(E/H) = 1.00$; and because the suspect is no more likely than anyone else to have those genetic markers if he is not guilty, $p(E/\bar{H}) = .05$, the frequency of the markers in the population. These probabilites may be plugged into the Bayesian formula in note 81, *supra*, allowing one to solve for p(H/E), which in this case equals .83. In other words, learning that the suspect and perpetrator match on a characteristic found in 5% of the population should cause the decisionmaker to revise his estimate of likelihood of guilt from 10% to 83%.

83. Thompson & Schumann, *supra* note 9.

call this mistake "the Prosecutor's Fallacy," explain the error by applying the underlying logic to a different problem:

> Suppose you are asked to judge the probability a man is a lawyer based on the fact he owns a briefcase. Let us assume all lawyers own a briefcase but only one person in ten in the general population owns a briefcase. Following the [fallacious] logic, you would jump to the conclusion that there is a 90 percent chance the man is a lawyer. But this conclusion is obviously wrong. We know that the number of nonlawyers is many times greater than the number of lawyers. Hence, lawyers are probably outnumbered by briefcase owners who are not lawyers (and a given briefcase owner is more likely to be a nonlawyer than a lawyer). To draw conclusions about the probability the man is a lawyer based on the fact he owns a briefcase, we must consider not just the incidence rate of briefcase ownership, but also the a priori likelihood of being a lawyer. Similarly, to draw conclusions about the probability a criminal suspect is guilty based on evidence of a "match," we must consider not just the percentage of people who would match but also the a priori likelihood that the defendant in question is guilty.[84]

The possibility that jurors might confuse population proportions with the probability of innocence was first raised by Laurence Tribe in his classic article, "Trial by Mathematics: Precision and Ritual in the Legal Process."[85] Discussing a murder case in which a partial palm print matching the defendant's is found on the murder weapon and a forensic expert testifies that such prints appear in no more than one case in a thousand, Tribe notes: "By itself, of course, the 'one-in-a-thousand' statistic is not a very meaningful one. It does not . . . measure the probability of the defendant's innocence—although many jurors would be hard-pressed to understand why not."[86] Tribe sees no problem with the admissibility of the forensic evidence of the match, but argues that the presentation of frequency data in connection with this evidence creates a serious danger of prejudice:

> To be sure, the finding of so relatively rare a print which matches the defendant's is an event of significant probative value, an event of which the jury should almost certainly be informed. Yet the *numerical index* of the print's rarity, as measured by the frequency of its random occurrence, may be more misleading than enlightening, and the jury should be informed of that frequency—if at all—only if it is also given a careful explanation that there might well be many other individuals with similar prints.[87]

Relying largely on Tribe's arguments, the Minnesota Supreme Court, in an interesting series of cases, has greatly limited the admissibility of statistics in connection with forensic evidence. No other appellate court has been as restrictive. The Minnesota opinions are worth examining in some detail because they raise a number of key issues about the competence of jurors to deal with statistics.

In *State v. Carlson*,[88] a rape and murder case in which hairs and semen were recovered from the victim, the court held that it was error (although nonprejudicial error) to admit statistical testimony on the probability of a match of characteristics. Specifically, the court found error in the admission of testimony by forensic hair expert Barry Gaudette that there was "a 1-in-800

84. *Id.* at 170.
85. Tribe, *supra* note 8.
86. *Id.* at 1355.
87. *Id.*
88. 267 N.W.2d 170 (Minn. 1978).

chance that the pubic hairs stuck to the victim were not Carlson's and a 1-in-4,500 chance that the head hairs found on the victim were not Carlson's."[89]

Carlson argued that Gaudette's testimony "goes one step too far toward an ultimate conclusion of fact and therefore invades the province of the jury," and the court apparently agreed.[90] As the court saw it, however, the problem was not so much that Gaudette had used statistics improperly as that he had used statistics at all.

> Our concern over this evidence is not with the adequacy of its foundation, but rather with its potentially exaggerated impact on the trier of fact. Testimony expressing opinions or conclusions in terms of statistical probabilities can make the uncertain seem all but proven, and suggest, by quantification, satisfaction of the requirement that guilt be established "beyond a reasonable doubt." *See* Tribe, *Trial by Mathematics,* 84 HARV. L. REV. 1329 (1971).
>
> Diligent cross-examination may in some cases minimize statistical manipulation and confine the scope of probability testimony. We are not convinced, however, that such rebuttal would dispel the psychological impact of the suggestion of mathematical precision, and we share the concern for "the substantial unfairness to a defendant which may result from ill conceived techniques with which the trier of fact is not technically equipped to cope." People v. Collins, 68 Cal. 2d 332, 66 Cal. Rptr. 505, 438 P.2d 41. For these reasons we believe Gaudette's [statistical] testimony . . . was improperly received.[91]

Although Carlson left it unclear whether the court's objection is to all frequency data or only statistics on the probability of a match, a subsequent case, *State v. Boyd,*[92] reveals that the court was troubled by a broad range of statistical formulations.

In *Boyd,* a rape case, the prosecutor sought to show that the defendant had fathered the victim's child in order to prove he had achieved sexual penetration. Deciding a pretrial appeal of a trial court's decision to suppress the results of a paternity test, the court allowed evidence that a paternity test had failed to exclude the defendant as a possible father, but rejected accompanying statistical testimony on the percentage of men in the general population that the test would also exclude, as well as a statistical calculation of the "probability of paternity." The court again cited Tribe as support for its conclusion that

> there is a real danger that the jury will use the [statistical] evidence as a measure of the probability of the defendant's guilt or innocence, and that the evidence will thereby undermine the presumption of innocence, erode the values served by the reasonable doubt standard, and dehumanize our system of justice.[93]

The most recent case in this line is *State v. Joon Kyu Kim,*[94] another rape case, in which the prosecution appealed the trial court's decision to exclude statistics offered in connection with serological tests performed on the defendant and a semen sample extracted from the victim. The test results showed a match on a set of genetic markers that occur in only 3.6 percent of

89. *Id.* at 173.
90. *Id.* at 175.
91. *Id.* at 176 (footnote omitted).
92. 331 N.W.2d 480 (Minn. 1978).
93. *Id.* at 483.
94. 398 N.W.2d 544 (Minn. 1987).

the population. The frequency of the set of markers was computed by determining the frequency of each marker separately and then multiplying those frequencies together in accordance with the product rule. The court rejected the use of this frequency calculation on the grounds that it might be mistaken by the jury for the probability of Kim's innocence:

> [T]he expert called by the state . . . should not be permitted to express an opinion as to the probability that the semen is Kim's and should not be permitted to get around this by expressing the opinion in terms of the percentage of men in the general population with the same frequency of *combinations* of blood types.[95]

Retreating slightly from its previous rejection of all frequency statistics, however, the court allowed testimony as to the percentage of men in the population who possess each of the individual matching genetic markers.[96] The court apparently believed these constituent probabilities were less likely to be prejudicial.

2. *Underutilization of Statistical Evidence.* In striking contrast to the concerns of Tribe and the Minnesota Supreme Court about the prejudicial potential of statistical evidence, other commentators have raised the opposite concern—that jurors will give statistics too little weight. Saks and Kidd criticise Tribe's analysis, calling it "a Swiss cheese of assumptions about human behavior—in this case human decision-making processes— which are asserted as true simply because they fall within the wide reach of the merely plausible, not because any evidence is adduced on their behalf."[97] Based on an extensive review of psychological studies on human judgment and decisionmaking, Saks and Kidd challenge arguments that statistics are inordinately persuasive and suggest that the reverse is true.[98]

A major psychological finding underlying Saks and Kidd's conclusion is the existence of the so-called base rate fallacy: the tendency for people, when judging the likelihood of an event, to ignore or underutilize statistical information on the base rate frequency of the event.[99] This tendency has been observed in a large number of studies.[100] When asked to judge whether a man described in a short vignette is a lawyer or an engineer, for example, people are nearly as likely to say he is a lawyer when told he was selected at random from a group consisting of 70 lawyers and 30 engineers as when told he was selected from a group consisting of 30 lawyers and 70 engineers.[101] The base rate data have relatively little impact on the judgment. "Only at the extremes of the distributions, where the group approaches 100 lawyers and 0

---

95. *Id.* at 549.

96. *Id.*

97. Saks & Kidd, *supra* note 8, at 125.

98. *Id.* at 149.

99. For reviews, see 3 E. BORGIDA & N. BREKKE, THE BASE-RATE FALLACY IN ATTRIBUTION AND PREDICTION: NEW DIRECTIONS IN ATTRIBUTION RESEARCH (1981); Bar-Hillel, *The Base-Rate Fallacy in Probability Judgments*, 44 ACTA PSYCHOLOGICA 211 (1980).

100. 3 E. BORGIDA & N. BREKKE, *supra* note 99; Bar-Hillel, *supra* note 99.

101. Kahneman & Tversky, *On the Psychology of Prediction*, 80 PSYCHOLOGICAL REV. 237, 241-43 (1973).

engineers (or the converse) do the decision makers become sensitive to the information about group composition."[102]

The example of the base rate fallacy most relevant to jury competence is people's response to the well-known cab problem developed by psychologists Daniel Kahneman and Amos Tversky:[103]

> A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:
>
> (a) 85 percent of the cabs in the city are Green and 15 percent are Blue.
>
> (b) a witness identified the cab as Blue . . . .
>
> [U]nder the same circumstances that existed on the night of the accident . . . the witness correctly identified each one of the two colors 80 percent of the time and failed 20 percent of the time.
>
> What is the probability that the cab involved in the accident was Blue rather than Green?[104]

This problem, and a number of similar problems, have been posed in dozens of psychological research studies.[105] As Saks and Kidd report, the typical probability response is 80 percent, although in actuality, the evidence given leads to a probability of 41 percent that the responsible cab was blue.[106] By judging the probability to be 80 percent, people are, in effect, ignoring the low base rate of blue cabs.

People's insensitivity to base rates in hypothetical problems of this type leads Saks and Kidd to suggest, in direct contradiction to Tribe and the Minnesota Supreme Court, that jurors are likely to pay little heed to base rates in actual legal proceedings. "[S]tatistical data need not be regarded as so overwhelming as some have supposed, and therefore they ought not to be considered prejudicial. The more realistic problem is presenting statistical evidence so that people will incorporate it into their decisions at all."[107]

It is important to notice, however, that the cab problem, as well as other problems revealing a "base rate fallacy," concerns the use of what this article has called "directly relevant" base rates, while the type of statistical evidence of concern to the Minnesota Supreme Court is "indirectly relevant" base rates. Thompson and Schumann have suggested that the two types of statistics "are likely to play a different role in the people's inferences" and therefore that the tendency to underutilize directly relevant base rates may

---

102. Saks & Kidd, *supra* note 8, at 128.

103. *See* Tversky & Kahneman, *Causal Schemata in Judgments Under Uncertainty*, in PROGRESS IN SOCIAL PSYCHOLOGY 49 (M. Fishbein ed. 1980); Tversky & Kahneman, *Evidential Impact of Baserates*, in JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES (D. Dahneman, P. Slovic & A. Tversky eds. 1982) [hereinafter *Evidential Impact*].

104. Tversky & Kahneman, *Evidential Impact*, *supra* note 103, at 156-57.

105. *See* 3 E. BORGIDA & N. BREKKE, *supra* note 99; Bar-Hillel, *supra* note 99.

106. Saks & Kidd, *supra* note 8, at 128. The correct solution to the problem may be obtained by applying Bayes' theorem. *See supra* notes 81, 82. The prior probability that the cab was blue, p(H), is .15 (and $p(\bar{H}) = .85$) because 15% of the cabs in the city are blue. This prior probability must be revised in light of evidence, E, that the witness identified the cab as blue. Because the witness is accurate 80% of the time, p(E/H) - .80 and $p(E/\bar{H}) = .20$. Plugging these values into the Bayesian formula in note 81, *supra*, one can solve for p(H/E) and see that it is .41.

107. Saks & Kidd, *supra* note 8, at 149.

not generalize to indirectly relevant base rates.[108] Thus, although Saks and Kidd marshal impressive evidence that statistics in general, and directly relevant base rates in particular, tend to be underutilized,[109] their analysis does not rule out the possibility that jurors may sometimes give too much weight to indirectly relevant base rates by falling victim to the Prosecutor's Fallacy of confusing the frequency of a matching characteristic with the probability of innocence.

3. *Jury Simulation Studies.* Additional light has been cast on the issue by recent studies of how simulated jurors react to statistical evidence when judging the guilt of hypothetical criminal defendants.[110] In a typical study, mock jurors are asked to estimate the probability that a hypothetical defendant is guilty based on a description of the evidence against him. The jurors are then asked to revise their initial estimate after receiving additional evidence indicating that the defendant and perpetrator share a characteristic, for example, a blood type, and specifying the statistical frequency of that characteristic. The weight jurors give to the evidence of the match is inferred from the extent to which they revise their estimates of probability of guilt after receiving it. A juror who initially thought the probability of guilt was 10 percent but revised his estimate to 90 percent after hearing about the match has given more weight to the evidence than a juror who revised the initial estimate from 10 percent to 15 percent. One advantage of assessing the weight mock jurors give to evidence in this manner is that it allows a direct comparison of human judgments to Bayesian predictions.[111] By using this approach, researchers can identify situations in which people give more or less weight to evidence than would be specified by Bayesian norms. It is also possible to detect instances in which people fall victim to the Prosecutor's Fallacy by determining whether they equate the frequency of a matching characteristic with the probability of innocence. In a case where the defendant and perpetrator matched on a characteristic found in 2 percent of the population, for example, a victim of the fallacy would always say the probability of guilt was 98 percent, regardless of the strength of the other evidence.

The studies generally find that some mock jurors make judgments consistent with the Prosecutor's Fallacy, although victims of this fallacy appear to be a small minority in most instances. For example, in the first study of this type, Thompson and Schuman asked 144 undergraduates to read a description of a hypothetical case involving the robbery of a liquor store by a man wearing a ski mask.[112] The police apprehended a suspect near the store who matched the store clerk's description of the robber's height, weight, and

108. Thompson & Schumann, *supra* note 9, at 169.
109. Saks & Kidd, *supra* note 8, at 132-45, 148-49.
110. Faigman & Baglioni, *supra* note 9; Thompson & Schumann, *supra* note 9; J. Goodman, *supra* note 9; E. Schumann & W. Thompson, *supra* note 9.
111. *See supra* notes 81-82.
112. Thompson & Schumann, *supra* note 9, at 172-76.

clothing. In a trash can nearby, the police found the mask and the money. Subjects were asked, at this point, to give an initial estimate of the probability of the suspect's guilt. Then they read a summary of testimony by a forensic expert who reported that hair found in the ski mask matched the suspect's hair and that the frequency in the general population of hair that would match the suspect's was 2 percent. The subjects then made a final estimate of the suspect's probability of guilt. On average, about 13 percent of subjects judged the suspect's probability of guilt to be exactly 98 percent, which is the probability one would obtain by simply assuming that the frequency of the matching characteristic equals the probability of innocence. These subjects' comments during debriefing confirmed that they had fallen victim to the Prosecutor's Fallacy.[113] In other studies of this type, the percent of subjects making judgments consistent with the Prosecutor's Fallacy has been lower, ranging from about 1 percent to 8 percent.[114]

These studies also find evidence of a second fallacy, which Thompson and Schumann have labeled the "Defense Attorney's Fallacy."[115] This fallacy is the erroneous assumption that evidence of a match between the defendant and perpetrator on a rare characteristic is irrelevant to the defendant's likelihood of guilt.[116] For example, in a case where the defendant and perpetrator match on a characteristic found in 1 percent of the population, victims of this fallacy might reason that in a city of one million people there would be approximately 10,000 people with the relevant characteristic. They then might erroneously conclude that there is little, if any, probative value in the fact that the defendant and perpetrator both belong to such a large group. This reasoning is fallacious because the great majority of the 10,000 people with the relevant blood type are not suspects in the case and because the blood-test evidence drastically narrows the group of individuals who are or could be suspects without eliminating the very individual on whom suspicion has already focused. Victims of the Defense Attorney's Fallacy give no weight to evidence of a match on a rare characteristic. Consequently, in simulation studies their initial and final judgments of probability of guilt are identical. In the same study by Thompson and Schumann, described above, in which 13 percent of subjects made judgments consistent with the Prosecutor's Fallacy, another 12 percent made judgments consistent with the Defense Attorney's Fallacy.[117] Hence, one quarter of the subjects made judgments consistent with fallacious reasoning.

Susceptibility to the fallacious reasoning appeared to depend in part on how the statistical evidence was presented. When the frequency of the

113. *Id.* at 173 n.5.

114. In a second study reported by Thompson and Schumann, only 4% of subjects made judgments consistent with the prosecutor's fallacy. *Id.* at 177-81. Two similar studies reported by Goodman found rates of 1.6% and 8%. J. Goodman, *supra* note 9, at 8, 20.

115. Thompson & Schumann, *supra* note 9, at 171.

116. *Id.*

117. *Id.* at 173.

matching characteristic was presented as a conditional probability,[118] the percentage of judgments consistent with the Prosecutor's Fallacy was higher (22 percent) and the percentage consistent with the Defense Attorney's Fallacy was lower (8 percent).[119] When the frequency was presented as a percentage and incidence rate,[120] the percentage of judgments consistent with the Prosecutor's Fallacy dropped to 4 percent but the percentage consistent with the Defense Attorney's Fallacy rose to 17 percent.[121]

Although these findings are suggestive, their practical significance is difficult to judge without more information. One limitation of the materials used in the study is that they included no arguments about the way statistical evidence should be used. In actual cases, jurors are likely to hear such arguments from attorneys or from other jurors during deliberation, so it is important to determine how people respond to them. Can people recognize the flaw in an argument for a fallacious position? Suppose they hear two fallacious arguments for contrary positions. How will they respond?

These questions were explored in a second study by Thompson and Schumann that was similar to the first[122] except that before subjects made final judgments of guilt, they received arguments.[123] One argument advocated the Prosecutor's Fallacy:

> The blood test evidence is highly relevant. The suspect has the same blood type as the attacker. This blood type is found in only 1% of the population, so there is only a 1% chance that the blood found at the scene came from someone other than the suspect. Since there is only a 1% chance that someone else committed the crime, there is a 99% chance the suspect is guilty.[124]

The other argument advocated the Defense Attorney's Fallacy:

> The evidence about blood types has very little relevance for this case. Only 1% of the population has the "rare" blood type, but in a city . . . [l]ike the one where the crime occurred with a population of 200,000 this blood type would be found in approximately 2000 people. Therefore the evidence merely shows that the suspect is one of 2000 people in the city who might have committed the crime. A one-in-2000 chance of guilt (based on the blood test evidence) has little relevance for proving *this* suspect guilty.[125]

Half the subjects read the prosecutor's argument first followed by the defendant's, while the other half read the arguments in reverse order. After reading each set of arguments, subjects were asked to indicate whether they thought either was correct and to judge the suspect's probability of guilt.

Most subjects found at least one of the fallacious arguments convincing. Twenty-nine percent thought the argument for the Prosecutor's Fallacy was

---

118. The expert stated there was "only a two percent chance [that] the defendant's would be indistinguishable from that of the perpetrator if he were innocent . . . ." *Id.*

119. *Id.* at 174.

120. The expert stated that 2 percent of people have hair that would be indistinguishable and that in a city of 1,000,000 people there would be approximately 20,000 such individuals. *Id.* at 173.

121. *Id.* at 174.

122. The case involved evidence of a match between the suspect and perpetrator on a blood type found in 1% of the population. *Id.* at 177-81.

123. *Id.* at 177.

124. *Id.*

125. *Id.* at 178.

correct.[126] Sixty-eight percent thought the argument for the Defense Attorney's Fallacy was correct.[127] Only 22 percent correctly concluded that both arguments were incorrect.[128]

Not surprisingly, after hearing the arguments a much higher percentage of subjects made judgments consistent with fallacious reasoning than in the earlier experiment. However, the Defense Attorney's Fallacy seemed to dominate. Over 50 percent of the subjects made judgments of probable guilt consistent with the Defense Attorney's Fallacy, giving no weight to evidence of a match on a characteristic found in 1 percent of the population,[129] but only 4 percent made judgments consistent with the Prosecutor's Fallacy.[130]

Overall, these findings suggest that people have difficulty detecting fallacious arguments—especially the argument favoring the Defense Attorney's Fallacy—and that these arguments can lead significant numbers of people to make judgments consistent with fallacious reasoning. In other words, it is easy to talk people into using inappropriate judgmental strategies to evaluate "indirectly relevant" base rate evidence presented in conjunction with forensic evidence.

With the exception of individuals who fall victim to the Prosecutor's Fallacy, however, most subjects in these studies appeared to give less weight to evidence of a match than Bayes' theorem says they should. A consistent finding, observed in six experiments,[131] is that subjects, on average, revise judgments of probability of guilt upward by a smaller amount than that required by Bayesian norms. In the first study by Thompson and Schumann, for example, subjects' initial judgments of probability of guilt averaged about 25 percent.[132] According to Bayes' theorem, after learning of a match on a characteristic found in 2 percent of the population, subjects' final judgments should have been about 93 percent.[133] However, subjects' actual judgments averaged only 63 percent,[134] indicating that, on average, they gave the evidence of the match less weight than they should have. This result is typical of findings in other studies.[135]

Researchers in this area sometimes warn that their findings should be viewed as preliminary.[136] Many of the studies are rather rudimentary simulations of trials in which subjects read summaries of evidence rather than see actual testimony. In addition, the subjects make individual judgments instead of deliberating as a group. Subjects are often asked to judge

126. *Id.*
127. *Id.*
128. *Id.* at 178-79.
129. *Id.*
130. *Id.*
131. Faigman & Baglioni, *supra* note 9; Thompson & Schumann, *supra* note 9, at 176, 180, experiments 1, 2; J. Goodman, *supra* note 9, studies 1, 2; Schumann & Thompson, *supra* note 9.
132. Thompson & Schumann, *supra* note 9, at 174.
133. *Id.* at 175.
134. *Id.*
135. Faigman & Baglioni, *supra* note 9; J. Goodman, *supra* note 9, at 30.
136. *See, e.g.*, Thompson & Schumann, *supra* note 9, at 183.

probability of guilt rather than to decide whether to convict or acquit, hence some concerns have been expressed about whether findings of such studies "[go] beyond the articulation of numbers and actually [influence] the sorts of decisions juries are called upon to make."[137]

Recently, however, researchers have begun conducting more realistic studies. For example, Schumann and Thompson[138] recently examined the effects of fallacious statistical arguments in the context of a highly realistic simulated trial. Subjects in the role of jurors viewed a four-hour videotape of a simulated trial based on transcripts of an actual California murder case. Although some of the testimony was abbreviated or replaced with stipulations, the simulated trial included virtually all of the evidence presented in the case on which it was based. The attorneys in the simulated trial were, in fact, experienced criminal lawyers, and the judge was a real judge who gave legally appropriate instructions.

The case involved a robbery and murder in which the perpetrator was injured, leaving blood at the scene. Although a considerable amount of circumstantial evidence was presented, the case against the defendant was weak except for a key piece of forensic evidence—genetic markers in his blood matched those of the blood at the scene, and the relevant markers are found in only 2 percent of the population.

Five different versions of the trial were shown to 116 simulated jurors. Some jurors heard the prosecutor make an argument advocating the Prosecutor's Fallacy, while others heard the prosecutor state only the frequency of the genetic markers. Within each of those two groups, half of the subjects heard the defense attorney make an argument for the Defense Attorney's Fallacy and half did not. Thus one group heard competing fallacious arguments, a second group heard no fallacious arguments, a third group heard just the argument for the Prosecutor's Fallacy, and a fourth heard just the argument for the Defense Attorney's Fallacy. The fifth group was a control condition in which evidence of the matching blood types and the accompanying statistical evidence were not presented. The arguments lasted less than one minute in the context of a ten- to fifteen-minute closing argument in a four-hour trial.

After watching the trial, the simulated jurors individually indicated whether they would vote guilty or not guilty and estimated the probability that the defendant actually committed the crime. Then they deliberated in groups of six for up to an hour before again indicating their choice of verdict and estimate of probability of guilt.

Before deliberation, half of the jurors voted guilty although no significant differences were found among the five conditions.[139] In other words, the evidence of the matching blood markers appeared to have no effect, regardless of how it was argued. Effects of the arguments emerged after

137. *Id.* at 183.
138. Schumann & Thompson, *supra* note 9.
139. *Id.* at 5.

test showing the match can have a large effect on the probative value of the evidence.[144]

Suppose, for example, that a juror initially estimates the probability of the defendant's guilt to be only 10 percent, but then receives new, independent evidence indicating that the defendant and perpetrator have the same blood type. If the incidence rate of the blood type and the false-positive rate of the test showing the match are both 1 percent, then, according to the Bayesian model, the juror should revise the estimate of probability of guilt upward to about .85 percent. If the incidence rate and false-positive rate are both 5 percent, however, the Bayesian model indicates the juror's revised estimate of probability of guilt should be only .53. Although the difference between 1 percent and 5 percent may appear small, it has a dramatic impact on the probative value of the match between the defendant and perpetrator.[145]

To evaluate the results of a forensic test linking the defendant to the crime in cases where the test is less than perfectly reliable, jurors must evaluate two factors: the possibility that the test result showing a match is a false-positive, and the possibility that the match, if correct, is merely coincidental. Experts on human judgment have suggested that when faced with problems such as this involving two sources of uncertainty, people often proceed in a stepwise fashion, following what has become known as the "best guess" strategy.[146] To reduce the complexity of the judgment, people make their best guess as to whether the evidence is reliable and, if they think it is probably reliable, they proceed to evaluate the evidence as if it were perfectly reliable. They then discount their certainty about their conclusions to take into account their uncertainty about the reliability of the evidence. They often fail to discount this evidence adequately, however. The result is that judgments based on less than fully reliable evidence are often unduly extreme because of the failure to discount adequately for unreliable evidence.[147] If this process influences jurors' evaluations of forensic matching evidence, it could cause jurors to be insensitive to variations in the reliability of the evidence.

The sensitivity to variations in the statistics accompanying forensic matching evidence was examined directly in a series of studies reported by Thompson, Britton, and Schumann.[148] In the standard experimental paradigm, mock jurors were asked to evaluate evidence of a match between the suspect and the perpetrator on a rare blood type. The rarity of the blood type and the false-positive rate of the forensic test were both experimentally varied to be either 1 or 5 percent. In the first study, undergraduate mock

144. The way a juror should respond to forensic evidence involving frequency statistics and false-positive rates may be specified by a mathematical model based on Bayes' theorem. *Id.* at Appendix. Thompson, Britton, and Schumann derived this model and used it as a benchmark against which to compare actual judgments based on such data.

145. *Id.* at 7.

146. *See, e.g.*, Gettys, Kelly & Peterson, *The Best-Guess Hypothesis in Multistage Inference*, 10 ORG. BEHAV. & HUM. PERFORMANCE 364 (1973); Slovic, Fischhoff & Lichtenstein, *Behavioral Disorder Theory*, 28 ANN. REV. PSYCHOLOGY 1 (1977).

147. Gettys, Kelly & Peterson, *supra* note 146.

148. Thompson, Britton & Schumann, *supra* note 9.

jurors were given several hypothetical cases at once in which the frequency and false-positive rate varied. These subjects were able to evaluate accurately the relative strength of the evidence. In other words, they could tell that the evidence was strongest where the frequency and false-positive rate were low, weakest where the frequency and false-positive rate were high, and of intermediate value where one factor was high and the other low. In the second study, however, subjects were given only one piece of forensic evidence to evaluate and the frequency and false-positive rate statistics were experimentally varied among different groups of subjects.[149]

Although the statistical variation should have made a large difference in the value of the evidence, the different groups of subjects did not differ significantly in the value they assigned to it.[150] The group that received the strongest evidence (low frequency, low false-positive rate) did not give the evidence more weight than the group that received the weakest evidence (high frequency, high false-positive rate). It thus appears that people can rank in order several pieces of evidence according to relative strength but have difficulty evaluating the absolute strength of any single piece of evidence.[151] Actual trials are, of course, more analogous to the second experiment than the first, because jurors are called upon to evaluate absolute strength of forensic matching evidence rather than the relative strength of several pieces of evidence.

In a third experiment, Thompson, Britton, and Schumann tested mock jurors sensitivity to these statistical variations in a more realistic study which included group deliberation.[152] Again they found that mock jurors did not differentiate weak from strong evidence. The hypothetical case was devised in such a way that Bayesian predictions of probability of guilt were 64 percent in the strong evidence condition, where forensic tests had a low false-positive rate and found a match on a rare characteristic, but only 22 percent in the weak evidence condition, where forensic tests had a higher false-positive rate and found a match on a more common characteristic.[153] The conviction rate of subjects in the two conditions did not significantly differ, however, and was actually a bit higher in the weak evidence condition (84 percent) than in the strong evidence condition (78 percent).[154] Following deliberation, subjects in both conditions estimated the probability of the defendant's guilt. Among subjects in the strong evidence condition the average estimate was 66 percent, which is very close to Bayesian norms. Among subjects in the weak evidence condition, however, the average estimate was 65 percent, which is much higher than the Bayesian prediction of 22 percent, indicating that

149. *Id.* at 7-10.
150. *Id.* at 10-19.
151. *Id.* at 11.
152. *Id.* at 12-13.
153. *Id.* at 15-20.
154. *Id.* at 17.

subjects' failure to differentiate weak and strong statistical evidence led them to overestimate the value of weak evidence in this instance.[155]

More research is needed in this area to test the generality of these provocative but rather preliminary findings. These studies suggest, however, that under some circumstances jurors may seriously overestimate the value of statistical evidence, as Tribe and the Minnesota Supreme Court feared. The reason for this problem, however, is not jurors' tendency to confuse the frequency of a matching characteristic with the probability of innocence, but their tendency to give equal weight to statistical evidence that varies widely in its probative value.

Recommendations for dealing with this problem may be a bit premature. Until the scope of the problem is better understood, attempted solutions might easily miss the mark. If these preliminary findings are borne out by further research, however, one possible solution would be for the courts to be especially cautious about admitting forensic evidence of questionable reliability. Forensic matching evidence that is relatively weak, by virtue of having a high false-positive rate, might be particularly likely to be overvalued.

C. Dealing With Partially Redundant Evidence

A third problem concerns subjects' reactions to partially redundant evidence, that is, evidence that partly overlaps and recapitulates facts they have already taken into account.[156] Forensic evidence showing a match between the suspect and perpetrator is partially redundant in cases where the suspect was selected in a manner that renders him more likely than most people to match the perpetrator on a certain characteristic. In the *Carlson* case[157] discussed earlier, for example, the defendant's hair matched, in a number of separate qualities, a hair taken from the perpetrator. Assuming the defendant was arrested in part because his hair matched the perpetrator's with regard to color and length, the evidence of a match on all qualities is partly redundant with what is already known.

The inferential complexities created by partial redundancy among multiple items of evidence have been discussed by philosophers of inductive logic,[158] legal evidence scholars,[159] and Bayesian theorists.[160] A variety of terms have been used to describe partially redundant evidence. The terms "cumulative" and "corroborative" are preferred by most legal scholars, but those terms will be avoided here because, as Schum notes, the precise

---

155. *Id.* at 17-19, 25.

156. *See supra* notes 55-59 and accompanying text.

157. State v. Carlson, 267 N.W.2d 170 (Minn. 1978).

158. J. VENN, THE PRINCIPLES OF INDUCTIVE LOGIC (2d ed. 1905); S. TOULMIN, THE USES OF ARGUMENT (1964).

159. J. WIGMORE, THE SCIENCE OF JUDICIAL PROOF (1937); Lempert, *Modeling Relevance*, 75 MICH. L. REV. 1021 (1977).

160. Most notably, D. Schum, On Factors which Influence the Redundancy of Cumulative and Corroborative Testimonial Evidence (1979) (Technical Report #79-02). *See also* Schum & Martin, *supra* note 80.

meaning varies so much among different scholars that their use promotes more confusion than clarity.[161] A precise account of the ways in which evidence may be partially redundant requires an appreciation of the catenated or "cascaded" nature of inductive inference.[162] Accordingly, the best accounts of partially redundant evidence are provided by Wigmore, who uses complex evidentiary diagrams to show connections among various pieces of evidence,[163] and Schum, who uses formal mathematical models of cascaded inference derived from Bayes' theorem.[164] Lempert also provides a clear, though less complete, account of partial redundancy using Bayesian models.[165]

Although considerable attention has been paid to formal descriptions of partial redundancy in evidence, relatively little is known about people's ability to apprehend and deal appropriately with this evidentiary subtlety when evaluating the evidence in criminal trials. The only empirical study that sheds light on this question was reported by Schum and Martin.[166] In this complex and sophisticated study, subjects evaluated evidence in hypothetical criminal cases in three different but formally equivalent ways. In some instances, subjects evaluated the evidence in the entire case "holistically;" in other instances the evidence was either partially or totally "decomposed," and subjects made separate evaluations of its constituent elements. Subjects were trained to evaluate the evidence by estimating likelihood ratios and conditional probabilities and they evaluated the evidence in these terms. The major finding of relevance here is that subjects' evaluations were more sensitive to partial redundancies among items of evidence when the evidence was decomposed than when evaluations were "holistic."[167] This finding raises the possibility that people may be inadequately sensitive to partial redundancy when, as in actual criminal trials, they are evaluating evidence that is not explicitly "decomposed" for them.

Some additional light has been cast on the problem by a study reported by Thompson, Meeker, and Britton.[168] Undergraduate subjects were asked to read written descriptions of evidence in a series of criminal cases. For each case, the description provided an account of the nature of the crime and the manner in which the suspect was identified. It then described two hypothetical pieces of forensic evidence that might be offered against the suspect. Subjects were asked to judge which of the two pieces of evidence would constitute stronger evidence of the suspect's guilt. Each piece of evidence revealed a match between the suspect and the perpetrator in a different but equally rare characteristic. One piece of evidence was partially

161. Schum & Martin, *supra* note 80, at 117.
162. *Id.* at 116-18.
163. J. WIGMORE, *supra* note 159, at 154.
164. Schum, *supra* note 160. *See also* Schum & Martin, *supra* note 80.
165. Lempert, *supra* note 80. *See also* R. LEMPERT & S. SALTZBURG, *supra* note 80, at 148-53.
166. Schum & Martin, *supra* note 80.
167. *Id.*
168. Thompson, Meeker & Britton, *supra* note 9.

redundant because the suspect had been selected in a way that rendered him unusually likely to have that matching characteristic even if innocent. The other piece of evidence was not redundant because the suspect was no more likely than anyone else to have it if he was innocent. For example, one case involved the burglary of a drug store by a black perpetrator who was injured, leaving blood at the scene. The police identified a suspect based, in part, on the fact that he was black. Subjects were then asked which of two pieces of evidence would be stronger: (1) evidence that the suspect and perpetrator both have sickle-cell characteristic, a trait found in about one person in 100 in the United States, but most commonly found among blacks and rarely found in other races, or (2) evidence that the suspect and perpetrator both have a hypothetical genetic characteristic (factor Q), which is found in one person in 100 in the United States but is evenly distributed among the races. The match on factor Q is clearly stronger evidence against the suspect because he was identified based in part on a characteristic (his race) that renders him more likely than the general population to have sickle-cell trait if innocent.

The goal of the study was simply to test whether people realize that the partially redundant piece of evidence deserved less weight than the nonredundant evidence. In general, it appears that they do not.[169] In most of the hypothetical cases, about a third of subjects thought the partially redundant evidence was stronger, about a third thought the two pieces of evidence were equally strong, and a third thought (correctly) that the nonredundant evidence was stronger. This distribution of responses is what would be expected if people could not detect any difference between the partially redundant and nonredundant evidence and simply responded at random. People's ability to appreciate the distinction appears to improve following discussion of the issue with others. Even after discussing the issue for up to twenty minutes with others, however, approximately half of the subjects still chose the incorrect option when asked which piece of evidence was stronger.[170]

Although these findings are preliminary, they raise serious concerns about the ability of jurors to detect partial redundancies in forensic evidence and to take those into account. As a result, jurors may overvalue forensic matching evidence in cases in which it is partially redundant with other evidence.

IV

CONCLUSION

Empirical research on peoples' evaluation of statistical evidence, although preliminary and full of gaps, is beginning to define the strengths and weaknesses of lay statistical reasoning in ways that should prove helpful to the legal system. This research casts some much-needed light on a number of issues that have divided commentators and troubled appellate courts.

169. *Id.*
170. *Id.*

However, it appears that the broad question posed by this article has no single or simple answer.

The research should allay fears that jurors will overvalue statistical evidence by mistakenly equating the frequency of matching characteristics with the probability of innocence. Although some jurors do fall victim to this type of fallacious reasoning, they are a small minority in all studies and the prevailing tendency is toward undervaluing rather than overvaluing such evidence. Arguments for the Prosecutor's Fallacy can have a powerful influence on judgments of guilt, but can be countered effectively by opposing arguments.

Jurors may sometimes overvalue forensic evidence used to link a defendant to a crime, but this potential problem arises not from the Prosecutor's Fallacy but from people's failure to take into account the unreliability and partial redundancy of forensic evidence. Where the value of forensic matching evidence is significantly undermined by the high false-positive rate of a forensic test, and where the defendant was selected in a manner that renders him more likely to possess the matching characteristics than the general population, there appears to be a significant danger that the forensic evidence will be overvalued. Until these judgmental tendencies are better understood, courts would be well advised to use caution when considering the admissibility of statistics in connection with such evidence.