

1991

The Siren Songs of Science: Toward a Taxonomy of Scientific Uncertainty for Decisionmakers

Vern R. Walker

Maurice A. Deane School of Law at Hofstra University

Follow this and additional works at: https://scholarlycommons.law.hofstra.edu/faculty_scholarship

Recommended Citation

Vern R. Walker, *The Siren Songs of Science: Toward a Taxonomy of Scientific Uncertainty for Decisionmakers*, 23 Conn. L. Rev. 567 (1991)
Available at: https://scholarlycommons.law.hofstra.edu/faculty_scholarship/59

This Article is brought to you for free and open access by Scholarly Commons at Hofstra Law. It has been accepted for inclusion in Hofstra Law Faculty Scholarship by an authorized administrator of Scholarly Commons at Hofstra Law. For more information, please contact lawcls@hofstra.edu.

THE SIREN SONGS OF SCIENCE: TOWARD A TAXONOMY OF SCIENTIFIC UNCERTAINTY FOR DECISIONMAKERS

*Vern R. Walker**

"Come this way, honored Odysseus, great glory of the Achaians,
and stay your ship, so that you can listen here to our singing;

.....

Over all the generous earth we know everything that happens."

—HOMER, *THE ODYSSEY*¹

I. INTRODUCTION

AS the population increases, as natural resources decrease, and as economic competition intensifies, it becomes ever more important for private parties, regulatory agencies, and courts to make decisions that are factually correct and efficient as well as effective. The ability to make such decisions about public health, safety, and the environment, for example, often depends upon the best use of available scientific information. It is not surprising, therefore, that during the last several decades administrative agencies with scientific expertise have multiplied and the areas regulated by those agencies have broadened. Nor is it surprising that the courts have been inundated with scientific information in both civil and criminal proceedings. Parties want to use the most recent scientific advances to forecast the effects of decisions upon their interests and to argue against decisions with adverse consequences. The sophistication of scientific methodology, the amount of scientific information about the world, and the desire of parties to use those methods and that information in legal decisionmaking all seem to

* Associate Professor of Law, Hofstra University School of Law; Ph.D., University of Notre Dame, 1975; J.D., Yale University, 1980. I wish to thank Bernard Jacob, James Hickey, Jr., Eric Freedman, Gary Liberson, Norman Silber, and Wei-Yann Tsai for their comments on an early draft of this Article. The ideas explored here were first presented in a preliminary form at the interdisciplinary conference, "The Environment," Hofstra University, June 7-9, 1990.

1. HOMER, *THE ODYSSEY*, Book XII, lines 184-91 (R. Lattimore trans. 1967) (the Sirens' song to Odysseus).

be growing apace.

As the pressures increase to make more and better use of scientific information, the legal profession, regulators, and the judiciary struggle to assimilate scientific information into legal processes. It takes only a few examples to suggest the breadth of that effort. Agencies, courts, and commentators have tried to determine the effects of pharmaceuticals,² pesticides,³ and air emissions⁴ on public health. They have puzzled over the relevance of observed differences in group composition to claims of illegal discrimination⁵ and voting power dilution.⁶ They have

2. See, e.g., *Lynch v. Merrell-National Laboratories Div. of Richardson-Merrell, Inc.*, 646 F. Supp. 856 (D. Mass. 1986), *aff'd*, 830 F.2d 1190 (1st Cir. 1987) (in a Bendectin case involving child with congenital birth defect, summary judgment for defendant drug manufacturer on issue of causation was appropriate where plaintiffs' evidence consisted of reanalysis of epidemiological studies, extrapolations from *in vivo* and *in vitro* animal studies, and studies of analogous chemical structures); *In re Richardson-Merrell, Inc. "Bendectin" Products Liability Litigation*, 624 F. Supp. 1212 (S.D. Ohio 1985) (providing extensive details of causation phase of bifurcated products liability action); United States Dept. of Health & Human Services, Food and Drug Administration, Draft Guideline Patient Package Insert, Bendectin and Other Combination Drugs Containing Doxylamine and Vitamin B₆, 45 Fed. Reg. 80,740 (1980), *withdrawn*, 47 Fed. Reg. 39,249 (1982).

3. For example, see the controversies surrounding proof of causation of cancer by chlorinated hydrocarbon pesticides and the waste byproducts of their manufacture. *National Coalition Against the Misuse of Pesticides v. EPA*, 867 F.2d 636 (D.C. Cir. 1989) (holding that the EPA's decision to settle with pesticide manufacturer for voluntary cancellation of pesticide registration, but to allow continued use and sale of existing stocks of pesticide, was not arbitrary or capricious in view of divergence of scientific opinion on risk to humans and lack of reliable data on environmental impact); *Sterling v. Velsicol Chem. Corp.*, 855 F.2d 1188 (6th Cir. 1988) (analyzing numerous causation issues associated with exposure to chlorinated hydrocarbon waste products through contaminated groundwater); *Dine v. Western Exterminating Co., Prod. Liab. Rep. (CCH) ¶ 11,714* (D.D.C. 1988) (evidence that chlordane used as termiticide was carcinogenic and that airborne concentrations could infiltrate treated structures presented sufficient evidence to raise jury question whether product was defective under risk/utility test of defectiveness); *Rabb v. Orkin Exterminating Co.*, 677 F. Supp. 424 (D.S.C. 1987) (plaintiffs failed to establish sufficient likelihood of future disease from exposure to chlordane, where they could offer no testimony that their increased risk of disease was greater than 50%).

4. See, e.g., *Final Determination under Section 126 of the Clean Air Act (Interstate Pollution Abatement)*, 49 Fed. Reg. 48,152 (EPA 1984), *reviewed in New York v. United States EPA*, 852 F.2d 574 (D.C. Cir. 1988) (denying petitions of Pennsylvania and Maine and remanding New York's petition for submission of new data, after the states claimed that air emissions from mid-western states prevented petitioners from attaining and maintaining national ambient air quality standards, impermissibly consumed a portion of petitioners' prevention of significant deterioration increments, interfered with visibility in petitioning states, and caused acid rain; EPA found that petitioners did not adequately demonstrate their claims of causation and injury).

5. E.g., *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642 (1989) (employment discrimination); *Bazemore v. Friday*, 478 U.S. 385 (1986) (employment discrimination); *Hazelwood School Dist. v. United States*, 433 U.S. 299 (1977) (employment discrimination); *Castaneda v. Partida*, 430 U.S. 482 (1977) (discrimination in grand jury selection). See generally Finkelstein, *The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases*, 80 COLUM.

wrestled with attempts to use advanced statistical modeling to help resolve economic disputes⁷ and charges of unconstitutional imposition of the death penalty.⁸ Courts are repeatedly confronted with issues ranging from novel forensic evidence of identification⁹ to epidemiological evidence of causation.¹⁰ Legal theorists have tried to determine how to improve the accuracy of verdicts,¹¹ how to integrate quantitative information with more traditional qualitative evidence,¹² and how to under-

L. REV. 737 (1980) [hereinafter *Regression Studies*]; Finkelstein, *The Application of Statistical Decision Theory to the Jury Discrimination Cases*, 80 HARV. L. REV. 338 (1966) [hereinafter *Statistical Decision Theory*].

6. See, e.g., *Campos v. City of Baytown, Tex.*, 840 F.2d 1240 (5th Cir. 1988), *cert. denied*, 109 S. Ct. 3213 (1989) (using regression analysis to determine correlation between percentage of minority voters in population and vote for minority candidates as evidence of political cohesiveness of minority for purposes of § 2 of Voting Rights Act of 1965, as amended).

7. See, e.g., Finkelstein, *Regression Models in Administrative Proceedings*, 86 HARV. L. REV. 1442 (1973) (discussing the types of regulatory proceedings in which regression analysis has been used as economic evidence); Fisher, *Multiple Regression in Legal Proceedings*, 80 COLUM. L. REV. 702 (1980) (discussing effect of cable television licensing upon revenues and growth of broadcast television stations, and antitrust damages in price-fixing cases). An illustrative use of regression models to determine economic effects is provided by Spiller, *The Differential Impact of Airline Regulation on Individual Firms and Markets: An Empirical Analysis*, 26 J. L. & ECON. 655 (1983).

8. E.g., *McCleskey v. Zant*, 580 F. Supp. 338 (N.D. Ga. 1984), *rev'd sub nom. McCleskey v. Kemp*, 753 F.2d 877 (11th Cir. 1985), *aff'd*, 481 U.S. 279 (1987) (deciding relevance and probative value of statistical arguments that death penalty was imposed because of race).

9. For example, see the historical controversy surrounding the admissibility of voiceprint spectrographic evidence in criminal proceedings. E.g., *United States v. Williams*, 583 F.2d 1194 (2d Cir. 1978), *cert. denied*, 439 U.S. 1117 (1979) (admissible); *People v. Kelly*, 17 Cal. 3d 24, 549 P.2d 1240, 130 Cal. Rptr. 144 (1976) (not admissible on record in specific case); *Cornett v. State*, 450 N.E.2d 498 (Ind. 1983) (not admissible, but harmless error to admit); *People v. Collins*, 94 Misc. 2d 704, 405 N.Y.S.2d 365 (Sup. Ct. 1978) (not admissible); *Pope v. State*, 756 S.W.2d 401 (Tex. Ct. App. 1988) (error in admitting evidence, if any, held harmless).

10. See, e.g., Black & Lilienfeld, *Epidemiologic Proof in Toxic Tort Litigation*, 52 FORDHAM L. REV. 732 (1984); Dore, *A Commentary on the Use of Epidemiological Evidence in Demonstrating Cause-in-Fact*, 7 HARV. ENVTL. L. REV. 429 (1983); Hall & Silbergeld, *Reappraising Epidemiology: A Response to Mr. Dore*, 7 HARV. ENVTL. L. REV. 441 (1983).

11. See, e.g., Callen, *Notes on a Grand Illusion: Some Limits on the Use of Bayesian Theory in Evidence Law*, 57 IND. L.J. 1 (1982); Kaye, *The Paradox of the Gatecrasher and Other Stories*, 1979 ARIZ. ST. L.J. 101 [hereinafter *Gatecrasher Paradox*]; Kaye, *The Laws of Probability and the Law of the Land*, 47 U. CHI. L. REV. 34 (1979) [hereinafter *Laws of Probability*]; Koehler & Shaviro, *Veridical Verdicts: Increasing Verdict Accuracy Through the Use of Overtly Probabilistic Evidence and Methods*, 75 CORNELL L. REV. 247 (1990). The foundational work on this topic is the following historical interchange: Finkelstein & Fairley, *A Bayesian Approach to Identification Evidence*, 83 HARV. L. REV. 489 (1970) [hereinafter *Identification Evidence*]; Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329 (1971) [hereinafter *Trial by Mathematics*]; Finkelstein & Fairley, *A Comment on "Trial by Mathematics"*, 84 HARV. L. REV. 1801 (1971) [hereinafter *Comment*]; Tribe, *A Further Critique of Mathematical Proof*, 84 HARV. L. REV. 1810 (1971) [hereinafter *Mathematical Proof*].

12. See, e.g., Kaplan, *Decision Theory and the Factfinding Process*, 20 STAN. L. REV. 1065

stand traditional legal doctrines within a scientific framework.¹³ The problems encountered in such efforts vary tremendously—depending upon the nature of the scientific methodology or information involved, the policies and structure of the relevant substantive law, and the objectives and constraints of the procedural rules, not to mention the more elusive demands of politics.

Throughout such efforts at assimilation, however, there runs at least one common thread: the available scientific information upon which a decision must be made is almost always a mixture of scientific knowledge and scientific uncertainty. Regardless of the information or methodology, there is potential for error. This is true whatever the relevant science, whether archeology or aeronautics, economics or engineering, pharmacology or toxicology or epidemiology. Achieving the best social decisions requires not only understanding and using that which we know, but also appreciating and weighing the extent of our uncertainty.¹⁴

In making the best use of scientific information in legal decision-making, it is still true that the beginning of wisdom is knowing what it is we do not know.¹⁵ Although significant progress has been achieved

(1968); sources cited *supra* note 11.

13. See, e.g., Brilmayer & Kornhauser, *Review: Quantitative Methods and Legal Decisions*, 46 U. CHI. L. REV. 116 (1978); Callen, *supra* note 11; L. Cohen, *Subjective Probability and the Paradox of the Gatecrasher*, 1981 ARIZ. ST. L.J. 627 [hereinafter *Subjective Probability*]; L. Cohen, *The Logic of Proof*, 1980 CRIM. L. REV. 91 [hereinafter *Logic of Proof*]; N. Cohen, *Conceptualizing Proof and Calculating Probabilities: A Response to Professor Kaye*, 73 CORNELL L. REV. 78 (1987); N. Cohen, *Confidence in Probability: Burdens of Persuasion in a World of Imperfect Knowledge*, 60 N.Y.U. L. REV. 385 (1985) [hereinafter *Confidence in Probability*]; *Identification Evidence*, *supra* note 11; Kaplan, *supra* note 12; Kaye, *Apples and Oranges: Confidence Coefficients and the Burden of Persuasion*, 73 CORNELL L. REV. 54 (1987); Kaye, *Is Proof of Statistical Significance Relevant?*, 61 WASH. L. REV. 1333 (1986); Kaye, *Statistical Significance and the Burden of Persuasion*, 46 LAW & CONTEMP. PROBS., Autumn, 1983, at 13; Kaye, *Paradoxes, Gedanken Experiments and The Burden of Proof: A Response to Dr. Cohen's Reply*, 1981 ARIZ. ST. L.J. 635 [hereinafter *Paradoxes*]; Kaye, *Probability Theory Meets Res Ipsa Loquitur*, 77 MICH. L. REV. 1456 (1979) [hereinafter *Res Ipsa Loquitur*]; *Laws of Probability*, *supra* note 11; *Gatecrasher Paradox*, *supra* note 11; Kornstein, *A Bayesian Model of Harmless Error*, 5 J. LEGAL STUD. 121 (1976); Lempert, *Modeling Relevance*, 75 MICH. L. REV. 1021 (1977); *Trial by Mathematics*, *supra* note 11; Williams, *The Mathematics of Proof*, 1979 CRIM. L. REV. 297, 340.

14. Judge Bazelon has urged: "Finally, I would especially stress the need for an agency to disclose the uncertainty that surrounds its determinations. And by uncertainty, I mean the agency's ignorance as well as its quantitative estimates of error." Bazelon, *Science and Uncertainty: A Jurist's View*, 5 HARV. ENVTL. L. REV. 209, 212 (1981).

15. See PLATO, *APOLOGY* 21 (F. Church trans. 2d ed. 1956):

So when I [Socrates] went away, I thought to myself, "I am wiser than this man: neither of us knows anything that is really worth knowing, but he thinks that he has knowledge

within isolated areas, efforts at making the best use of scientific information have been hampered by the lack of a comprehensive understanding of the nature and structure of scientific uncertainty. Therefore, my purpose in this Article is to provide a taxonomic scheme¹⁶ for classifying the different kinds of scientific uncertainty—that is, a scheme for identifying the different kinds of potential error associated with descriptive scientific information.¹⁷ It is my intention that the classification scheme be general in scope: that the taxonomy cover all of the important kinds of descriptive uncertainty, and that it apply to information from any of the physical, biological, and social sciences.¹⁸ Such a broad classification scheme should provide decisionmakers with a foundation for understanding the nature of scientific information generally.

In order to focus the discussion and make it concrete, this Article discusses specific examples of scientific information. One example is the assertion that exposure to the magnetic fields generated by an alternating current electric transmission line can cause childhood leukemia.¹⁹

when he has not, while I, having no knowledge, do not think that I have. I seem, at any rate, to be a little wiser than he is on this point: I do not think that I know what I do not know."

Id.

William Ruckelshaus, then Administrator of the EPA, stated in 1983:

Given the necessity of acting in the face of enormous scientific uncertainties, it is more important than ever that our scientific analysis be rigorous and the quality of our data be high. We must take great pains not to mislead people about the risks to their health. We can help to avoid confusion by [sic] ensuring both the quality of our science and the clarity of our language in explaining hazards.

Ruckelshaus, *Science, Risk, and Public Policy*, 221 *SCIENCE* 1026, 1027 (1983); see also Guidelines for Carcinogen Risk Assessment, 51 *Fed. Reg.* 33,992, 33,998 (EPA 1986) (decisionmakers faced with determining carcinogenic risks should evaluate the level of uncertainty associated with their assessments of potential population exposure, and take measures of such uncertainty into account in order to achieve a clear understanding of the effect of this uncertainty on any final quantitative risk estimate).

16. "Taxonomy" is "the systematic distinguishing, ordering, and naming of type groups within a subject field." *WEBSTER'S THIRD NEW INTERNATIONAL DICTIONARY* 2345 (P. Gove ed. 1976).

17. By "descriptive" scientific information, I mean those scientific statements that purport to describe the world and how it works. Prescriptive or normative uncertainty is addressed briefly *infra* note 227. Detailed exploration of prescriptive uncertainty, however, is left for another time.

18. I do not pretend to demonstrate in any rigorous way that my taxonomic scheme is complete, in the sense that all possible types of uncertainty are covered. What I do hope to accomplish, however, is an analysis of the most significant aspects of scientific methodology as it is currently practiced and of the descriptive information commonly encountered by decisionmakers.

19. Whether this assertion is true is an important issue for regulators, just as it is for those involved in civil litigation. See, e.g., *Houston Lighting & Power Co. v. Klein Indep. School Dist.*, 739 S.W.2d 508, 514-18 (Tex. Ct. App. 1987) (finding sufficient evidence on which jury could have concluded, that transmission lines near school posed a risk to children and that "uncertainty

Of course, this Article is not concerned with whether such an assertion is true. Rather, such examples are used to illustrate the logically distinct ways in which scientific assertions might be in error.

I discuss in this Article six kinds of descriptive uncertainty: (1) conceptual, (2) measurement, (3) sampling, (4) modeling, (5) causal, and (6) epistemic.²⁰ As will become clear, this ordering tracks the following logical levels nested within scientific assertions: (1) the definition and choice of descriptive concepts or variables to be used as predicates; (2) the application of those concepts or variables to specific, individual cases; (3) the generalization from specific, observed cases to unobserved cases; (4) the prediction of one predicate or variable as a mathematical function of other predicates or variables; (5) the inference from certain mathematical functions between variables to conclusions about causal relationships; and (6) the choice of interpretations for fundamental, logical concepts used throughout levels (1) to (5). In this Article, I address each level and each type of uncertainty in turn, suggesting how each is important to social decisionmaking, and illustrating how scientists try to reduce each type and how they express or characterize any residual uncertainty.

Before beginning this detailed analysis, a few general remarks are in order about why these six types of descriptive uncertainty are logically distinct categories. First, distinct scientific activities give rise to each type of error. For example, generating data through observing and measuring particular instances is an activity different from generalizing beyond those observations, or from positing a causal system that explains those observations. As we will see, conceptual uncertainty and measurement uncertainty can attend any particular observation,²¹ but sampling uncertainty arises only with generalization²² and causal un-

over the magnitude of that risk should dictate caution"); EPA, Workshop Review Draft, Evaluation of the Potential Carcinogenicity of Electromagnetic Fields (1990) (Doc. No. EPA 600/6-90/005A) (circulated for comment on technical accuracy and policy implications).

20. The potential for computational error is not discussed as a separate category for several reasons. First, computational uncertainty pervades many of the categories of uncertainty. Moreover, the risk of computational error is widely appreciated, the means of avoiding it are generally understood, and this kind of error presents few conceptual puzzles for decisionmaking. It should be clear without further elaboration that most scientific information carries with it the potential for computational error.

There are similar reasons for not discussing other kinds of purely human error, or deliberately produced error (intentional fraud), as separate categories of scientific error. The potential for such errors occurs in science, as elsewhere.

21. See *infra* Sections II and III.

22. See *infra* Section IV.

certainty only with explanation.²³ A scientific description of a particular object, therefore, might involve only conceptual and measurement uncertainty, while a causal assertion typically involves all six kinds. As a result, a simple causal assertion can serve throughout this Article as an example of descriptive scientific information.²⁴

A second reason that the six categories are distinguishable is that scientists employ different techniques to reduce each different kind of uncertainty. For example, increasing the number of observations or drawing a stratified random sample can reduce sampling error,²⁵ while other control techniques reduce causal uncertainty.²⁶ Finally, the categories are distinguishable because scientists typically use different ways to measure, characterize, or communicate the extent of *residual uncertainty* in each category—that is, the uncertainty that remains even after efforts have been made to reduce the potential for error.²⁷

My objective in this Article is a broad one: to provide a conceptual overview of the kinds of uncertainty associated with any descriptive scientific information. Such an overview should prove fruitful to decisionmakers and legal theorists as they continue to explore the ways in which particular decisionmaking processes can be improved.²⁸ This analysis should also help decisionmakers to understand, in a comprehensive way, the descriptive science upon which their decisions rest. It is my hope that such a broad understanding will lead decisionmakers to

23. See *infra* Section VI.

24. I do not want to suggest, however, that this Article is simply about causation. My objective is to catalog the kinds of uncertainty that can be associated with *any* descriptive assertion, including causal assertions.

25. See *infra* text accompanying notes 108-11.

26. See *infra* text accompanying notes 186-98.

27. The six categories of uncertainty are also distinct because they can combine in different ways to produce different aggregate uncertainties. Cf. J. COHRSEN & V. COVELLO, *RISK ANALYSIS: A GUIDE TO PRINCIPLES AND METHODS FOR ANALYZING HEALTH AND ENVIRONMENTAL RISKS* 94 (United States Council on Environmental Quality 1989) ("uncertainties accumulate rapidly in a risk assessment"; uncertainty in the variable of failure rate for particular plant equipment expands the uncertainty in estimates of exposed populations, which in turn contributes to uncertainty in dose estimates and total uncertainty in the final risk estimate). This Article does not provide a system for combining the different kinds of potential error into an aggregate measure. By providing an overview of the kinds of uncertainty significant to decisionmakers, however, the taxonomy set forth here should serve as the foundation for a useful theory of aggregate uncertainty.

28. This Article suggests different ways that decisionmakers might take the various kinds of scientific uncertainty into account. A detailed analysis of this problem, however, would require considering the substantive and procedural details of each regulatory or judicial task. A general article such as this cannot provide a sufficiently detailed basis for adequately addressing any particular problem.

approach the reduction of scientific uncertainty and the use of scientific information in a wise and cost-effective manner.

II. CONCEPTUAL UNCERTAINTY

The logical structure of any descriptive assertion, whether scientific or not, is predication: the subject of an assertion identifies what is being discussed, and the predicate describes or characterizes what is identified.²⁹ For example, the assertion "That is a transmission line" predicates the property of "being a transmission line" of an object that is being pointed out by the speaker.³⁰ The assertion "That transmission line is generating a magnetic field" contains a further predication about the object identified as a transmission line,³¹ as does "That transmission line caused this case of leukemia."³² Generic information is also predication. For example, the assertion "Electric current generates a magnetic field" predicates of every electric current the property of creating a magnetic field.³³ Predication, therefore, is the essential structure of descriptive information; the propositions about the world formed by predication are either true or false.

Conceptual uncertainty, or the potential for conceptual error, arises whenever predication occurs. Whenever a concept is used to de-

29. See C. LEWIS & C. LANGFORD, *SYMBOLIC LOGIC* 263-67 (2d ed. 1959); Garver, *Subject and Predicate*, in 8 *THE ENCYCLOPEDIA OF PHILOSOPHY* 33 (P. Edwards ed. 1967). I do not intend or need to claim, of course, that the only use of language is to describe things, or that all utterances are either true or false. See, e.g., J. AUSTIN, *HOW TO DO THINGS WITH WORDS passim* (1962) (exploring the performative use of language).

30. In the symbolism of quantification logic, the logically significant components of the assertion can be represented by "T" (symbolizing the property of being a transmission line) and "a" (denoting the particular physical object being identified). The proposition being asserted is then symbolized as "Ta." See, e.g., I. COPI, *SYMBOLIC LOGIC* 64-65 (4th ed. 1973).

31. The proposition asserted here is more complicated logically. It would be symbolized by "Ta · Ma," where "M" represents the additional predicate "generating a magnetic field," and the dot "·" stands for the logical operation of conjunction ("and"). *Id.* at 8-9, 64-65.

32. The proposition asserted here is symbolized by "Ta · Lb · Cab," where "L" is the additional predicate "is a case of leukemia," "b" denotes the particular condition or instance asserted to be leukemia, and the relational predicate "C" relates two subjects and stands for "is the cause of." See *id.* at 112-14. The logical structure displays the three distinct elementary propositions that are asserted with the single English sentence. The English sentence might be in error in one of at least three ways: it might be false that this is a transmission line; it could be that this medical condition is not leukemia; or it is possible that object "a" (the transmission line) is not causally related to "b" (the case of illness). Such decomposition of English sentences into their logical components is a familiar task to attorneys, especially litigators, even if the logical notation is not familiar.

33. In logical notation, the proposition asserted would be rendered as " $(x) Ex \supset Mx$," which can be read "for any thing x, if x is an electric current, then x is generating a magnetic field." See *id.* at 64-68.

scribe something, the use of certain concepts instead of others begins to structure the way that we understand the object, event, or instance under discussion.³⁴ Predication or conceptualization generates useful information about things, but it also can inhibit our ability to think about those same things with concepts other than those selected. The concepts actually used may not be the most fruitful or the best designed—either for scientific purposes or for the purpose of making wise, fair, effective, and efficient decisions.

A high-voltage transmission line, for example, may be studied within a variety of scientific disciplines, with a corresponding variety of conceptual frameworks. To a physicist, a transmission line is, among other things, a source of electromagnetic fields, while to the behavioral scientist it is a visual stimulus for a psychological reaction, perhaps an object of fear or pride. To a utility company's electrical engineer responsible for the reliability of the bulk power supply system, the line is conceptualized in terms of its capacity to transfer electrical energy reliably. To an epidemiologist, exposure to the magnetic fields generated by the line might be a potential risk factor for childhood leukemia, while to an economist, the line might be an investment. Restricting consideration of a transmission line to certain scientific disciplines, therefore, can influence the decisions made concerning that power line.³⁵

In addition to conceptual alternatives between scientific disciplines, differences often occur within the same scientific discipline. In most fields, conceptual frameworks evolve over time.³⁶ Some fields, such as

34. I assume that the subject of scientific predication may be any "thing": for example, a physical object or group of such objects, a quality of an object, a theoretical entity or construct, or an event or situation. Cf. E. CARMINES & R. ZELLER, *RELIABILITY AND VALIDITY ASSESSMENT* 9-10 (1979) (social scientists typically measure abstract phenomena, in addition to objects or events). The neutral term "instance" is sometimes used to denote any individual subject of predication.

35. An example of a judicial statutory interpretation relieving agencies of the legal need to consider an entire field of scientific information is *Metropolitan Edison Co. v. People Against Nuclear Energy*, 460 U.S. 766 (1983), in which the United States Supreme Court held that the Nuclear Regulatory Commission was not required by the National Environmental Policy Act to consider what effects restarting the Three Mile Island nuclear plant might have on the psychological health of those residents around the plant who were concerned over the risk of nuclear accident.

36. See generally G. ALLEN, *LIFE SCIENCE IN THE TWENTIETH CENTURY* (1975); H. BUTTERFIELD, *THE ORIGINS OF MODERN SCIENCE 1300-1800* (1957); W. COLEMAN, *BIOLOGY IN THE NINETEENTH CENTURY: PROBLEMS OF FORM, FUNCTION, AND TRANSFORMATION* (1971); T. KUHN, *THE STRUCTURE OF SCIENTIFIC REVOLUTIONS* (1962); S. TOULMIN & J. GOODFIELD, *THE ARCHITECTURE OF MATTER* (1962).

the various behavioral sciences, have been notorious for generating competing conceptual frameworks at any given time. Scientific change alters the way scientists think about the world, and it often involves inventing, refining, or discarding the concepts used to describe that world.

Scientists prefer to talk about *variables* rather than concepts, thus focusing attention on the object being studied. A variable is a property that varies from individual to individual. The "value" or "score" of the variable can change from instance to instance (for example, height or age).³⁷ Any descriptive predication can, in turn, be viewed as an assertion about the score or value of a variable in a particular instance. For example, to assert that "That is a transmission line" is to score the object referred to on the variable "being a transmission line": "yes" or "no." Thus, any information—even the most qualitative—can be viewed as the result of scoring or assessing instances with respect to variables.

Thinking in terms of variables is more than simply a change of nomenclature. Once the world, as seen through the eyes of the scientist, is conceptualized as the object of scoring, then we can go further than mere logical structure and can evaluate variables for their adequacy from the standpoint of measurement. For example, variables should be defined in such a way that every possible instance (object or event) has some value for each variable. If some instances have no score, then they are indeterminate with respect to that variable, and predications of that variable cannot be determined to be either true or false.³⁸ The scoring categories for a variable, therefore, must be exhaustive before the variable can escape indeterminacy.³⁹

37. See, e.g., E. GHISELLI, J. CAMPBELL & S. ZEDECK, *MEASUREMENT THEORY FOR THE BEHAVIORAL SCIENCES* 9-10 (1981); H. LOETHER & D. MCTAVISH, *DESCRIPTIVE AND INFERENTIAL STATISTICS: AN INTRODUCTION* 14 (2d ed. 1980) (discussing measurement techniques employed by sociologists).

38. This would be radical indeterminacy, not just a reflection of practical ignorance. Even if I do not know the location of a particular transmission line, I know in principle how to resolve the question. Such ignorance is simply a limitation in my factual knowledge, not a problem with the definition of the variable "being a transmission line."

39. The value categories for a variable are "exhaustive" or "inclusive" if they are defined in such a way that every instance fits into *some* category. E.g., H. LOETHER & D. MCTAVISH, *supra* note 37, at 16; Reynolds, *Nominal Data*, in 6 *ENCYCLOPEDIA OF STATISTICAL SCIENCES* 256 (S. Kotz & N. Johnson ed. 1985). Thus, if the variable is color and the categories are defined as "red" and "non-red," those categories are exhaustive. If, however, the categories were defined as "red" and "blue," these two categories would not be exhaustive because yellow things could not be classified. In order for a concept or variable to be well-defined, it must provide categories in which to place every instance. See *id.*; H. LOETHER & D. MCTAVISH, *supra* note 37, at 16.

In addition to being exhaustive, the scoring categories of a variable should be mutually exclusive: each instance should fit into *only* one category.⁴⁰ If the categories are not mutually exclusive (for example, if an instance could go into either of two categories), results may be inconsistent or misleading.

So far, the illustrations of scientific concepts and variables have been primarily *qualitative* or *nominal*,⁴¹ emphasizing that scientific information of any type can be thought of in terms of variables and measurement. Unlike the values of qualitative variables, those of *quantitative variables* are related to each other on a scale. The simplest level of quantitative variable, the *ordinal variable*,⁴² merely orders or ranks the categories relative to an increase of the property under consideration. For example, while the variable "carcinogenic to humans" might be defined as a nominal variable (a chemical agent would be then regarded simply as either carcinogenic to humans or not),⁴³ the variable "hazardous to humans" might be defined as an ordinal variable ranking degrees of hazard ("low," "medium," and "high").⁴⁴ Finally, *scalar*

40. See H. LOETHER & D. MCTAVISH, *supra* note 37, at 16; Reynolds, *supra* note 39, at 256. With regard to the variable of color, the categories "yellow" and "non-red" are not mutually exclusive because we can properly place a yellow object into either category.

41. Qualitative variables generate *nominal data*. There is no ranking or ordering among the value categories. E.g., E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 23-24; H. LOETHER & D. MCTAVISH, *supra* note 37, at 16-17 (nominal variables in sociology include marital status, gender, and religious affiliation); Reynolds, *supra* note 39, at 256.

Dichotomous variables have two value categories. E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 27-28. There is no theoretical limit, however, to the number of value categories that can be used to define a variable, although there are practical and psychological reasons for limiting the number. As long as the subcategories are exhaustive and mutually exclusive, multivalued variables can improve our ability to classify phenomena. For example, in a particular study it might be useful to define the variable "color" as possessing only six categories: "red," "yellow," "blue," "black," "white," and "other."

42. E.g., E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 12-13, 24; Agresti, *Ordinal Data*, in 6 *ENCYCLOPEDIA OF STATISTICAL SCIENCES* 511 (S. Kotz & N. Johnson eds. 1985). Unlike the quantitative scales used to define scalar quantitative variables, see *infra* text accompanying note 45, ordinal scales order categories, but do not measure the degree or amount of the property from category to category. See, e.g., E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 12; H. LOETHER & D. MCTAVISH, *supra* note 37, at 17 (giving example of social class grading schemes as ordinal variables).

43. See, e.g., Guidelines for Carcinogen Risk Assessment, 51 Fed. Reg. 33,992, 33,994-96 (EPA 1986) (hazard identification part of risk assessment is qualitative assessment of "whether or not an agent may pose a carcinogenic hazard"); NATIONAL RESEARCH COUNCIL, *RISK ASSESSMENT IN THE FEDERAL GOVERNMENT: MANAGING THE PROCESS* 19 (1983) (defining the "hazard identification" step in risk assessment as the process of determining whether exposure to an agent can cause an increase in the incidence of an adverse health effect—theoretically, a "yes-no" question).

44. For an example of an ordinal variable in a regulatory context, see EPA Pesticide Labeling

variables are quantitative variables whose categories are related by some measure of incremental frequency, degree, or amount of the relevant property.⁴⁵ For example, the variable "magnetic field strength" is measured on a real-number scale, typically in gauss (G) or milligauss (mG).⁴⁶ A three-milligauss field, therefore, has twice the strength of a 1.5-milligauss field.

Regardless of the type of variable employed, the most basic descriptive information provided by scientists to decisionmakers takes the form of *data*, a collection of reports about observations or measurements of particular instances.⁴⁷ But even such elementary items of scientific information come laden with conceptual uncertainty. The variables used might be badly defined because their categories are not exhaustive or are not mutually exclusive. Variables might be defined nominally, when ordinal or scalar definitions later prove more appropriate.⁴⁸ Some variables might not be fruitful, or might lead to wrong theories, and might be abandoned or modified later when new variables are invented. The conceptual tools of the different scientific disciplines are frequently in flux. The problem for decisionmakers is that, while

Requirements for Pesticides and Devices, 40 C.F.R. § 156.10(h)(1) (1989) (defining Toxicity Categories I to IV for pesticide labeling purposes).

45. E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 12. It is common to classify scalar variables as either *interval variables* or *ratio variables*, *id.* at 13-15; H. LOETHER & D. MCTAVISH, *supra* note 37, at 17-19, although this distinction is not of conceptual importance here. Interval variables are those whose quantitative scales measure the intervals between subcategories by means of a standard unit of measurement, but which do not have a true zero value (a subcategory for instances that do not possess the property identified by the variable at all). An example is a psychological scale measuring arithmetic ability without testing for the total absence of such ability. E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 14 (stating that "many if not most psychological variables" are interval rather than ratio variables). The values of *ratio variables* are related by unit intervals and include a true zero category as well. For example, scales measuring length typically have true zero values, indicating the absence of length. Having a true zero value allows measured individuals to be compared to each other in terms of possessing a ratio (proportion or percentage) of the property. *See id.* at 13-14; H. LOETHER & D. MCTAVISH, *supra* note 37, at 18-19 (giving examples of ratio variables in sociology, such as "size of family").

46. *See, e.g.*, Scientific Advisory Panel, New York State Power Lines Project, Final Report on Biological Effects of Power Line Fields 31 (1987).

47. H. LOETHER & D. MCTAVISH, *supra* note 37, at 15. Reports of individual observations are sometimes legally important in themselves, for example, where an investigating scientist reports the measurements of magnetic field strength under a particular transmission line. Such reports can also form the basis for collective or statistical information, or provide support for a scientific theory.

48. Distinguishing these types of variables becomes extremely important when statistical techniques are to be applied to data. Some techniques apply only to nominal data, ordinal data, or scalar data. *See id.* at 220-65 (describing different measures of association for nominal, ordinal, and interval variables).

scientists go about their task of conceptualizing and classifying things in ways that they hope will prove fruitful for scientific prediction and explanation, litigating parties, courts, and agencies must make their decisions with the scientific information then at hand.⁴⁹

Courts often exhibit the reasonable, although conservative, approach of not accepting novel scientific conceptualizations until the associated theories have been reasonably well accepted by scientists.⁵⁰ That threshold test of acceptance may be difficult both to define and to apply.⁵¹ Such an approach, however, comports with the entrenched disposition of scientists themselves not to adopt concepts or variables as isolated items, but always to develop and define them within broader conceptual systems. The broader the range of phenomena covered by the theory, and the more coherent the inner logical connections of the theory, the more confidence scientists have that the theory provides an adequate representation of the real world.⁵² Moreover, integration of a concept or variable into a broad, established scientific theory might be the only assurance a decisionmaker can receive that conceptual uncertainty is being kept to a minimum, given the current state of science.

An important but unresolved question is how to measure, or even characterize, the residual conceptual uncertainty associated with scien-

49. Tribe has suggested that one consequence of increased reliance upon mathematical proof in the trial setting "may be to shift the focus away from such elements as volition, knowledge, and intent, and toward such elements as identity and occurrence—for the same reason that the hard variables tend to swamp the soft." *Trial by Mathematics*, *supra* note 11, at 1366. In other words, increased reliance upon mathematically structured inference might influence our choice of which variables to consider legally significant because ease of measurement would ease problems of proof. *Cf.* Brillmayer & Kornhauser, *supra* note 13, at 117 ("Counting has invaded, indeed nearly conquered, the social sciences.").

50. *See, e.g.,* *Frye v. United States*, 293 F. 1013, 1014 (D.C. Cir. 1923) (rejecting novel scientific evidence unless premised on a principle "sufficiently established to have gained general acceptance in the particular field in which it belongs"); Giannelli, *The Admissibility of Novel Scientific Evidence: Frye v. United States, a Half-Century Later*, 80 COLUM. L. REV. 1197, 1205, 1207 (1980) (*Frye* test, which dominated admissibility of scientific evidence for more than half a century, established a method for ensuring reliability of scientific evidence). *Cf.* FED. R. EVID. 703 ("If of a type reasonably relied upon by experts in the particular field in forming opinions or inferences upon the subject, the facts or data [upon which an expert witness bases an opinion or inference] need not be admissible in evidence."); J. WEINSTEIN & M. BERGER, WEINSTEIN'S EVIDENCE MANUAL ¶ 13.03[02][c] (1987) (under Rule 703, "the proponent of the expert must establish that experts other than the proposed [expert] witness would act upon the information relied upon, and would do so for purposes other than testifying in a lawsuit").

51. *See* Giannelli, *supra* note 50, at 1208-31.

52. Other components of a complete scientific theory, besides the menu of concepts and variables associated with it, will be discussed throughout this Article, particularly in the sections dealing with measurement, modeling, and causal analysis. Scientific theories are highly complicated logical structures, not simply bundles of concepts or variables.

tific information—the uncertainty that remains once efforts have been made to reduce conceptual uncertainty as much as possible. Perhaps all that is possible, at least at the present time, is to attempt to ensure that all relevant scientific disciplines (and all viable theories within those disciplines) are represented when decisions are deliberated. Doing so, however, does not necessarily assist the decisionmaker in knowing how best to integrate competing conceptual frameworks to arrive at a single decision. Decisionmakers usually must accept the vastly splintered state of scientific knowledge and must try to determine how to make decisions in the face of radical scientific pluralism.⁵³

III. MEASUREMENT UNCERTAINTY

Once descriptive concepts and variables have been identified as potentially useful or appropriate, the “measurement process” introduces another kind of scientific uncertainty. *Measurement*, in this general sense, is the process of classifying or sorting things into the categories of variables.⁵⁴ More specifically, it is the process of determining the value of a variable for a particular instance. Unaided visual inspection, for example, might be the measurement method for sorting balls into the categories of a nominal variable (for example, the variable “color”: “red,” “blue,” “yellow,” and “other”) or an ordinal variable (for example, the variable “size”: “small,” “medium,” and “large”). Calipers can measure the diameters of the balls in centimeters. A gas chromatograph can identify the presence or absence of certain chemical compounds in samples (for example, the nominal variable “contains chlordane”: “yes,” “no”) or their quantity (for example, micrograms of chlordane per sample).⁵⁵

Measurement, then, is a generic term for any process used to classify instances into the various categories of a variable, whether that variable is qualitative or quantitative. Measurement uncertainty is the

53. For a discussion of the problem of optimal decisionmaking and scientific theories concerning decisionmaking itself, see *infra* note 227.

54. See, e.g., E. CARMINES & R. ZELLER, *supra* note 34, at 10; H. LOETHER & D. MCTAVISH, *supra* note 37, at 14. Some authors reserve the word “measurement” for use with quantitative variables, and use “classification” for qualitative variables. See E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 9. While this usage might conform better to ordinary usage, I use the word “measurement” to underscore the logical similarities among all processes of sorting instances into categories, regardless of the nature of the variable.

55. See generally Giuliani, *Gas Chromatographic Analysis in Water Pollution*, in CHROMATOGRAPHIC ANALYSIS OF THE ENVIRONMENT 195-218 (R. Grob 2d ed. 1983) [hereinafter CHROMATOGRAPHIC ANALYSIS OF THE ENVIRONMENT]; Grob & Kanatharana, *Gas Chromatographic Analysis in Soil Chemistry*, in *id.*, at 347-73.

potential for misclassification, for error in placing a particular instance into the wrong category. Such uncertainty is logically distinct from conceptual uncertainty, for the possibility of measurement error arises only after a variable has been selected and the associated risk of conceptual uncertainty has been incurred.⁵⁶

Depending upon the nature and extent of measurement error, misclassification can have legal significance in matters ranging from rulemaking and regulatory enforcement proceedings to verdicts in civil or criminal cases. The legal significance of any measurement uncertainty would vary with the substantive issues being decided, the allocation of the burden of producing evidence, and the degree of uncertainty that can be tolerated in the legal decisionmaking.

Scientists normally evaluate a measurement method by assessing its *validity* and its *reliability*. A measurement method is *valid* to the degree that "it measures what it purports to measure."⁵⁷ If the measurement method introduces a bias or systematic error into the results (systematic overestimate or underestimate), then to that extent the measurement method lacks validity.⁵⁸ However, even a valid measurement method might produce random error or *noise* around the true amount.⁵⁹ To the extent that there is such noise among different measurement scores of the same thing, the measurement technique is *unreliable*.⁶⁰ The total measurement uncertainty is generally a combination of the potential for error in these two dimensions: systematic error and random error.⁶¹ These two aspects of measurement uncertainty will be discussed individually, for they are two different sources of measure-

56. When a variable is not well defined, as when the criteria for category inclusion are vague, the risk of misclassification may increase. Nevertheless, misclassification can occur even when the variable is well defined in principle.

57. E. CARMINES & R. ZELLER, *supra* note 34, at 12; E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 266; H. LOETHER & D. MCTAVISH, *supra* note 37, at 14, 32; see 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, 279 (1981) (in forensic science, a forensic technique's validity depends upon the percentage of cases in which the analyst can make correct determinations). The term "valid" is sometimes applied by extension to the measurement instrument used in the method, or to the measurement result itself.

58. See *infra* notes 63-66 and accompanying text.

59. See *infra* notes 76-88 and accompanying text.

60. E. CARMINES & R. ZELLER, *supra* note 34, at 13; E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 191; see Imwinkelried, *supra* note 57, at 279 (in forensic science, a forensic technique's reliability depends upon the percentage of cases in which independent examiners of the test results will make the same determination).

61. See sources cited *infra* note 62; Peters & Westgard, *Evaluation of Methods*, in TEXTBOOK OF CLINICAL CHEMISTRY 410, 413 (N. Tietz ed. 1986).

ment error.⁶²

A. *Validity*

A measurement method that produces systematic, nonrandom results measures *something*, even if we are mistaken about what we are measuring. For a measurement method to be valid, therefore, we must be correct in our interpretation of exactly what is being measured. When a process actually measures one variable when we think it is measuring another, the results are potentially informative, but we draw our conclusions about the wrong thing. For example, if, in trying to sort red balls from those of other colors, a person consistently selects blue balls instead of red ones, that person's visual perception is not a valid measurement method for redness (though it might be a valid method for blueness).

Lack of validity also occurs when the method measures two variables at once, only one of which is the variable we think we are measuring. In such a case, the "interfering variable" produces a systematic bias in the measurements. For example, if a person includes both orange and red balls in a group that is supposed to contain only red balls, that person's results have a bias toward overestimating the number of red balls.

Problems with measurement validity occur in every scientific area. In the biophysical sciences, sophisticated measurement techniques often raise questions concerning validity—as, for example, when clinical laboratory results are affected by intermediate chemical products.⁶³ In the social sciences, where measurement instruments often take the form of questionnaires or tests administered to people, problems of the proper interpretation of scores often arise.⁶⁴ Private or regulatory decisions can

62. See E. CARMINES & R. ZELLER, *supra* note 34, at 13-15; E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 187-90; Cameron, *Error Analysis*, in 2 *ENCYCLOPEDIA OF STATISTICAL SCIENCES* 545, 550 (S. Kotz & N. Johnson eds. 1982); Currie, *Sources of Error and the Approach to Accuracy in Analytical Chemistry*, in 1 *TREATISE ON ANALYTICAL CHEMISTRY* 95, 119-22 (I. Kolthoff & P. Elving eds. 1978).

63. See Peters & Westgard, *supra* note 61, at 411-12 (example of glucose measurements being affected by hydrogen peroxide when a glucose oxidase reaction is employed).

Analytical specificity is another term used to refer to "the ability of an analytical method to determine solely the component(s) it purports to measure." *Id.* at 411 (quoting Büttner, Borth, Boutwell & Broughton, *Provisional Recommendation on Quality Control in Clinical Chemistry*, 22 *CLINICAL CHEMISTRY* 538 (1976) [hereinafter *Provisional Recommendation*]).

64. The California F Scale, for example, may be interpreted as measuring two different properties at the same time: adherence to authoritarian beliefs and the trait of tending to agree with assertions. E. CARMINES & R. ZELLER, *supra* note 34, at 15.

be significantly affected by bias in environmental measurements.⁶⁵ The Environmental Protection Agency (EPA), for example, has denied petitions for relief from interstate air pollution under section 126 of the Clean Air Act in part because the sulfate data of the petitioning states were probably biased toward overestimating the amount of sulfate in ambient air.⁶⁶ In such examples, errors in the results are due in part to lack of validity.

Assessing validity poses different problems for different scientific disciplines and different variables. In the biophysical sciences, the *criterion validity* of one measurement technique is determined by statistically correlating its results with those of a "criterion method" or "standard method" for measuring the desired variable.⁶⁷ In analytical chemistry, for example, the *inaccuracy* or systematic error of a new analytical method is determined by comparing its results with those of another method whose accuracy has already been accepted.⁶⁸ Criterion validation thus requires independent evidence of the validity of the cri-

65. See Currie, *supra* note 62, at 98-100 (examples of nitrogen dioxide in ambient air or concentrations of cholesterol in blood).

66. Interstate Pollution Abatement, 49 Fed. Reg. 48,152, 48,153 (EPA 1984) (final determination) (petitioners had not corrected sulfate data for artifact formation caused by sampling technique); Interstate Pollution Abatement, 49 Fed. Reg. 34,851, 34,863 (EPA 1984) (proposed determination) (certain glass filters used by petitioning states in their high-volume air samplers were believed to result in overestimation of true sulfate concentrations).

67. Compare Peters & Westgard, *supra* note 61, at 412 (accuracy of analytical method usually established by comparing results with "another method whose accuracy has already been established") with E. CARMINES & R. ZELLER, *supra* note 34, at 19 (no criterion variables for "many if not most measures in the social sciences"). Regulatory agencies sometimes establish officially adopted "reference methods" as criteria. See, e.g., OSHA Regulations on Asbestos, Tremolite, Anthophyllite, and Actinolite, 29 C.F.R. § 1910.1001 (1990); EPA Regulations on Ambient Air Quality Surveillance, 40 C.F.R. §§ 50.1-50.2 (1990); EPA Regulations on National Air Monitoring Stations, 40 C.F.R. §§ 58.30-58.36 (1990).

The statistical correlation is sometimes referred to as a "validity coefficient." E. CARMINES & R. ZELLER, *supra* note 34, at 18; E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 269; *Validity*, in 9 ENCYCLOPEDIA OF STATISTICAL SCIENCES 460 (S. Kotz & N. Johnson eds. 1988). The assessment of method accuracy is also called "method validation." See, e.g., Brame, *Gas Chromatographic Analysis in Air Pollution*, in CHROMATOGRAPHIC ANALYSIS, *supra* note 55, at 121, 123.

68. Peters & Westgard, *supra* note 61, at 412. "The term *inaccuracy* has been recommended to emphasize lack of agreement [between results of the method being evaluated and the criterion method] and is defined as the 'numerical difference between the mean of a set of replicate measurements and the true value.'" *Id.* (quoting *Provisional Recommendation*, *supra* note 63, at 538).

A more generic meaning of "accuracy" is the degree to which measurement results differ from the true value. Accuracy, in that generic sense, means "total error," both random and systematic. *Id.* at 413.

terion method.⁶⁹ When criterion validation is possible and its validity is well grounded, then its empirical basis makes it a particularly strong test of validity.⁷⁰

By contrast, an evaluation of a measurement method's *construct validity* begins at the level of theory.⁷¹ A scientist first identifies the logical relationships between the variable to be measured and other variables within the theory. Then the various measurement methods for each of the theoretically related variables are examined to determine what statistical associations should be observed between data from the different methods. For example, within a theory about causes of anxiety, the variable "common anxiety" might be predicted to be strongly but negatively associated with the variable "self-esteem." That is, when an individual's score is low with respect to "self-esteem," he or she is expected to have a high score with respect to "common anxiety."⁷² If the relevant measurement methods all generate results that are statistically associated in the patterns that would be predicted by the theory, then there is evidence that the new measurement method is valid, in the sense that its results comport with what one expects given the theoretical relationships between variables.⁷³

An assessment of construct validity, then, depends on the ade-

69. See E. CARMINES & R. ZELLER, *supra* note 34, at 19.

70. The level of validity that scientists consider acceptable varies from situation to situation. See generally Currie, *supra* note 62, at 199-209. In the sciences, as in legal settings, the costs and benefits are generally weighed before validity uncertainty is considered acceptable. See *id.*; Peters & Westgard, *supra* note 61, at 413-15 (approximate specifications for allowable analytical errors depend upon medical mission, population being served, particular application of test, and physician's interpretation of test results); Westgard & Klee, *Quality Assurance*, in TEXTBOOK OF CLINICAL CHEMISTRY, *supra* note 61, at 424 (quality goals for laboratories must vary, depending upon such factors as medical missions of health care facilities and cost).

71. A third approach to validity, *content validity*, is sometimes used in the behavioral sciences, when researchers evaluate a measurement tool by assessing how well the tool covers some relevant content. E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 274. Examinations in an educational setting provide obvious examples. See E. CARMINES & R. ZELLER, *supra* note 34, at 20 ("This type of validity has played a major role in the development and assessment of various types of tests used in psychology and especially education but has not been employed widely by political scientists or sociologists."). An assessment of content validity would identify the content to be covered by the test (such as facility with different arithmetic concepts and skills), judge the extent to which each item of the test addresses some aspect of that content, and determine whether every aspect of that content is addressed by some item in the test. See *id.* at 20-22; E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 275-77.

72. For discussion of this example, see E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 282-87.

73. See generally E. CARMINES & R. ZELLER, *supra* note 34, at 23-26 (discussing the ideal of having particular measures generate patterns of consistent findings when different theoretical structures are used in a number of studies).

quacy of the theory on which it is based, as well as on the validity of the relevant measurement methods. Ultimately, however, construct validity is probably all we are capable of achieving. Even when criterion validity is available, our confidence in the validity of the criterion methods themselves must be grounded ultimately in construct validity. Construct validity functions as a "coherence theory" of validity, in the sense that the theory, the measurement expectations based on the theory, and the measurements themselves are all expected to cohere.

Sometimes measurement uncertainty can be reduced by carefully evaluating the validity of the measurement techniques. Whether it is possible to increase the validity of a method depends, of course, on the nature of the particular method and the type of validity problems encountered. The residual measurement uncertainty traceable to problems with validity can itself sometimes be measured.⁷⁴ However, our confidence in the validity of a measurement method ultimately depends upon our confidence in the theory that supports the method and in the theory that supports any criterion method used in the validation. There is good reason to expect more measurement uncertainty when dealing with measurement methods that depend upon a novel or narrowly established theory for their validation.⁷⁵

Decisionmakers should not merely assume that measurements offered by scientists are valid, even if those results have an aura of mathematical precision. Even variables with measurement values carried out to many decimal places can lack validity. Validity is always a question *external* to the measurement method itself; it is a question about the proper *interpretation* of the data generated by the method, and involves a scientific judgment that the data in fact give information about the variable of interest.

74. For example, scientists can sometimes provide a validity coefficient relative to some criterion method. See *supra* note 67. An additional complication is sampling uncertainty, which will be discussed in Section IV. The potential for sampling error is present because any statistical tests for criterion validity are based on samples. The sampling uncertainty that is associated with a validity coefficient can be dealt with using techniques discussed in Section IV.

75. See *supra* text accompanying notes 50-52; cf. Giannelli, *supra* note 50, at 1248 (proposing that prosecutors in a criminal case who wish to introduce a novel scientific technique should be required to establish its validity beyond a reasonable doubt, while criminal defendants and civil litigants should be held to a preponderance of the evidence standard in establishing the validity of novel techniques that they wish to introduce).

B. *Reliability*

In contrast to validity error, which is systematic in nature and can be viewed as a problem external to the measurement method, unreliability is a source of error intrinsic to even validly interpreted measurement methods. A method is reliable to the extent that it is capable of generating consistent results when repeatedly applied to the same subjects.⁷⁶ In theory, if a perfectly reliable measurement technique were used to measure the same subject under the same circumstances, without the subject's having changed in any relevant respect, then the results of all the measurements would be identical. In practice, of course, measurement methods are never perfectly reliable. Variations in results arise due to changes in the surrounding circumstances or in the measurement instruments, or because of variations in the performance of the persons conducting the measurements.

If a valid measurement method is used, such that there are no systematic errors in the results, deviations of results from the true value of the variable are expected to be random and to "cancel each other out" in the long run. Under classical test theory, if a large number of repeat measurements are made of the same thing, then the mean of those measurements is expected to equal the "true value."⁷⁷ While errors due to unreliability fall randomly around their mean, the mean of all the actual measurements is expected to equal the true value in the long run. Put differently, there should be no correlation between errors due to unreliability and the true value.⁷⁸

When reliability is increased, the extent of the random scatter around the mean for repeat measurements is decreased.⁷⁹ The difficulty lies in determining whether reliability is in fact being increased and, if

76. See, e.g., E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 184, 191 (reliability of measurement is "the degree of self-consistency among the scores earned by an individual;" reliability of measurement is "the extent of unsystematic variation in the quantitative description of some characteristic of an individual when the same individual is measured a number of times"). The variability between repeated measurements is sometimes divided into *repeatability variance* (variance recorded under identical conditions) and *reproducibility variance* (variance reflecting all random contributions to measurement, such as different conditions, instruments, operators, samples, days, laboratories, or environments). Hunter, *Measurement Error*, in 5 *ENCYCLOPEDIA OF STATISTICAL SCIENCES* 378, 379 (S. Kotz & N. Johnson eds. 1985); Mandel, *Accuracy and Prediction: Evaluation and Interpretation of Analytical Results*, in 1 *TREATISE ON ANALYTICAL CHEMISTRY*, *supra* note 62, at 259.

77. See E. CARMINES & R. ZELLER, *supra* note 34, at 29-30.

78. See, e.g., *id.* at 30; J. COHEN & P. COHEN, *APPLIED MULTIPLE REGRESSION/CORRELATION ANALYSIS FOR THE BEHAVIORAL SCIENCES* 68 (2d ed. 1983).

79. See E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 193-94.

so, by how much. In some situations, and particularly in the physical sciences, it is possible to determine the degree of reliability of a method by using a *test-retest* approach.⁸⁰ In analytical chemistry, for example, it is possible to analyze the same sample of material more than once.⁸¹ Reliability error can be reduced (or "precision" increased) by various techniques, depending upon the relevant measurement method.

The social sciences, however, often encounter methodological difficulties with the notion of measuring the "same" person twice, without intervening changes in the subject due to, among other things, the memory of having taken the same test earlier.⁸² One approach to this problem is to use a multi-item questionnaire or measurement instrument. If the items of such an instrument are "parallel"—that is, they are interchangeable and have no systematic differences in their results—then the different questionnaire items are sometimes treated as though they are replicate measurements of the same variable.⁸³ In such a case, the correlation between scores on parallel items is sometimes used as a *reliability coefficient of internal consistency*.⁸⁴ The advantage

80. See E. CARMINES & R. ZELLER, *supra* note 34, at 37-40 (discussing difficulties with using the retest method in the social sciences).

81. This is sometimes called a *replication experiment*, and the term *precision* is sometimes used to refer to the "agreement between replicate measurements." Peters & Westgard, *supra* note 61, at 412 (citing *Provisional Recommendation*, *supra* note 63, at 538). The term *imprecision* is defined as the "standard deviation or coefficient of variation of the results of a set of replicate measurement[s]." *Id.* So defined, the precision of a measurement method is a measure of its reliability, and imprecision is another term for random analytical error. See *id.*

Precision can be measured within the same analytical run ("within-run precision"), within different runs on the same day ("within-day precision"), and within different runs on different days ("day-to-day" or "between-day precision"), with the longer scales including random errors from additional variables, such as different operators or other conditions in the laboratory. *Id.* The reliability or precision of a method can also be assessed between laboratories, using split samples of the same test material. See, e.g., Mandel, *supra* note 76, at 256-60.

82. See, e.g., E. CARMINES & R. ZELLER, *supra* note 34, at 39-40, 50. A logically similar technique is to divide the total number of items (questions) in a survey or test instrument into halves (the "split-halves" method) and examine the degree of correlation between the two halves. Difficulties arise, however, in determining how to partition the different items into two groups: different groupings are likely to result in different reliability estimates. *Id.* at 41-43, 50.

83. See *id.* at 32-34, 47 (differences between truly parallel measurements due to purely random error); E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 192 ("reliability can be defined as the extent of unsystematic variation of one individual's scores on a series of parallel tests").

84. E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 204, 257-58.

One of the most popular of such measures is *Cronbach's coefficient alpha*, which is a function of the average correlation between all of the multiple items in the instrument (the mean inter-item correlation). E. CARMINES & R. ZELLER, *supra* note 34, at 43-47. Cronbach's alpha is a conservative estimate of an instrument's internal consistency: if the items are truly parallel, with inter-item correlations all being equal, see E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 203-04,

of such a notion is that consistency can be determined on the basis of a single administration of an instrument.⁸⁵

When measures of reliability are available, they can be used to characterize the residual uncertainty associated with reliability. In the case of standardized measures varying between zero and one, zero usually indicates totally random results (no correlation between scores in a replication experiment or on parallel tests, or no inter-item correlation) and one indicates perfect correlation with no inconsistency in results.⁸⁶ The acceptability of any given degree of reliability should be a function of the purpose for which the results are to be used.⁸⁷ Depending upon the risks, benefits, and policies involved in a given legal setting, one level of reliability might be acceptable in one proceeding, such as an administrative rulemaking, but unacceptable in another, such as a criminal trial.⁸⁸

IV. SAMPLING UNCERTAINTY

Scientists seldom classify or measure all of the things or instances they ideally would like to measure. Such incomplete observation often occurs for practical reasons: the incremental information to be gained by observing every instance often diminishes drastically in significance as additional measurements are made. There can also be theoretical

then Cronbach's alpha is equal to the reliability coefficient, *see* E. CARMINES & R. ZELLER, *supra* note 34, at 45; E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 203-04. To the extent that the average inter-item correlation decreases toward zero, Cronbach's alpha likewise decreases toward zero. *See* E. CARMINES & R. ZELLER, *supra* note 34, at 45.

The internal inconsistency of a multi-item measurement, as measured by Cronbach's alpha, can be reduced in several ways. First, as items or questions are better constructed so that the average inter-item correlation is increased, the Cronbach's alpha for the instrument will increase. *See id.* at 45-46. In addition, for any given average of inter-item correlation, the Cronbach's alpha will increase as the number of items in the measurement instrument increases. *Id.*

85. E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 253.

86. *Id.* at 193-94, 204, 205-07 (discussing reliability coefficient); E. CARMINES & R. ZELLER, *supra* note 34, at 45 (discussing Cronbach's alpha).

87. For example, lawyers are familiar with the specious precision of numbers specified to decimal places well beyond what is truly significant. Carrying a number to decimal places well beyond what can be justified given the level of precision of the measurement method appears to serve no function except to mislead. *See, e.g.,* Berry & Geisser, *Inference in Cases of Disputed Paternity*, in *STATISTICS AND THE LAW* 353, 376 (M. DeGroot, S. Fienberg & J. Kadane ed. 1986) (example of using six-digit paternity index when measurement and sampling uncertainty render calculations to more than two digits "suspect").

88. While analytical chemists might expect a high degree of precision before they would consider a measurement method sufficiently reliable, social scientists generally expect a Cronbach's alpha of only 0.8 before they regard a multi-item instrument as sufficiently reliable. E. CARMINES & R. ZELLER, *supra* note 34, at 51 ("reliabilities should not be below .80 for widely used scales").

barriers to observing every instance.⁸⁹ Much of the power of science, and much of its attractiveness to decisionmakers, lies in its historical ability to generalize beyond the scope of personal observation, whether such generalizations are important for broad policy decisions (such as environmental or economic regulation) or as part of the rationale for deciding a particular case (for example, a generic causal connection between a kind of pharmaceutical and the plaintiff's type of injury in a tort case). When scientists generalize, however, they create the potential for sampling error.⁹⁰

Statisticians refer to the set of actual measurements as the *sample*, and to the larger set of potential measurements as the *population*.⁹¹ A summary number characterizing a sample is called a *statistic* (for example, the proportion of sample instances in a certain category or the arithmetic mean of sample measurements), while the corresponding characteristic of a population is called a *parameter*.⁹² *Sampling error* is the error that can be introduced into our conclusions by the process of inferring a parameter from a statistic. The potential for such error can be thought of as *sampling uncertainty*.

Scientists have made substantial progress in their ability to determine or characterize sampling uncertainty. Within the classical theories of probability and statistics,⁹³ the two major (but related) approaches are "significance testing" and the construction of "confidence intervals."⁹⁴ These techniques, which will be discussed in turn, can be

89. See, e.g., H. LOETHER & D. MCTAVISH, *supra* note 37, at 5-6. Theoretical barriers occur when some instances will occur in the future (instances that are not yet observable), or when a set is infinite (e.g., the set of spatial points in a force field).

90. If the scientific information relevant to decisionmaking relates merely to the measurement of a particular thing, then conceptual and measurement uncertainties might be the only uncertainties associated with the information. More frequently, however, individual measurements are made in the process of ultimately drawing a conclusion about a group of things, not all of which have been measured or observed individually.

91. H. LOETHER & D. MCTAVISH, *supra* note 37, at 4-5; see W. HAYS, *STATISTICS* 190-92 (4th ed. 1988).

92. H. LOETHER & D. MCTAVISH, *supra* note 37, at 6.

93. The classical theories interpret probability statements as assertions about long-run relative frequencies or proportions. Another interpretation of probability statements will be discussed in Section VII, in connection with epistemic uncertainty. That alternative interpretation also provides an alternative approach to determining sampling uncertainty.

94. These two techniques provide adequate illustrations of sampling uncertainty, although they are both examples of ways to deal with "Type I error"—the type of error made when a correct hypothesis about a population parameter is rejected. This Article does not discuss the type of sampling error called "Type II error"—the error made when an incorrect hypothesis about the parameter is not rejected—or the "power" of a study to reject an incorrect hypothesis. See, e.g., W. HAYS, *supra* note 91, at 261-63; H. LOETHER & D. MCTAVISH, *supra* note 37, at 511-15. Cf.

of great use to decisionmakers, provided the techniques are properly understood and their limitations respected.

A. *Significance Testing*

All classical reasoning about sampling uncertainty is hypothetical in form: *if* the unknown parameter is x , what can we deduce about the probability of drawing certain samples from that population? The reasoning known as *significance testing* or *hypothesis testing* begins by positing an hypothesis about the value of the population parameter, then using that hypothetical value to reach conclusions about the probability of obtaining values for statistics in samples. Drawing a sample with a statistic whose probability is extremely low thus provides evidence against the truth of the hypothesis. We therefore may conclude that such an hypothesis is unlikely to state the true parameter value.

An example should help to clarify this reasoning. Suppose that there are 100 balls in a closed box (the population). Suppose also that each of the balls is either red or white, and that we are trying to reach a conclusion about the proportion of red balls in the box. We are allowed to draw out ten balls without replacing any, line them up, and examine them (the sample). If we draw 4 red balls and 6 white, we can reach a limited number of conclusions about the population with no sampling uncertainty. For example, any conclusion that the population contains less than 4 red balls or more than 94 red balls would be demonstrably false.⁹⁵ We cannot be certain, however, of the total number of red balls, other than knowing it is between 4 and 94.

Under certain conditions we can go further in structuring our uncertainty. For example, if we draw our sample as a *simple random sample* (that is, every possible sample of 10 balls has an equal chance of being the sample that we draw),⁹⁶ then, even if we draw our sample

Confidence in Probability, *supra* note 13, at 410-17 (explaining Type I and Type II errors); Feinberg, *Teaching the Type I and Type II Errors: The Judicial Process*, THE AM. STATISTICIAN, June 1971, at 30 (using the null hypothesis that a specific criminal defendant is innocent); *Statistical Decision Theory*, *supra* note 5, at 364 (discussing standing for complaining of Type II error).

95. If we had drawn the sample while replacing the ball in the box after each draw, we could be *certain* only that at least one ball in the population is red and that at least one ball is white.

96. See, e.g., W. HAYS, *supra* note 91, at 52-53; H. LOETHER & D. MCTAVISH, *supra* note 37, at 407. What allows us to structure our uncertainty within the range of *possible* values in the example in the text is the assumption about the equal probability of drawing any particular sample. It is not essential, however, that the sample be a simple random sample. What is essential is that the sampling be conducted in such a way that we are able to generate a *probability distribution* for the relevant statistic in all possible samples of a given size, and we are therefore able to

while replacing each ball after it has been drawn and examined, we can conclude that it is *extremely unlikely* that there are only 4 red balls in the box. The reasoning is as follows. First, consider the hypothesis that only 4 balls out of the 100 are red. If each ball has an equal chance of being drawn, then on each draw the probability of drawing a red ball is $4/100$, or 0.04 .⁹⁷ We can conceptualize our sample as a sequence of 10 events, each event being a drawing with a 0.04 chance of drawing a red ball.⁹⁸ In addition, we can calculate the probability of drawing 4 red balls in our sample of 10.⁹⁹ In this case, the probability is very, very small.¹⁰⁰ Because the likelihood of drawing a simple random sample containing 4 red balls is so small, we conclude that our hypothesis about there being only 4 red balls in the population is prob-

attach a probability to drawing any particular sample (e.g., a sample having 4 red balls out of 10). See, e.g., *id.* at 415-22 (discussing cluster and stratified random samples); W. HAYS, *supra* note 91, at 53, 209-10. Such a probability distribution is called a *sampling distribution* for the statistic. H. LOETHER & D. MCTAVISH, *supra* note 37, at 435.

97. See, e.g., W. HAYS, *supra* note 91, at 120 ("For equally probable elementary events the probability of any event A is simply the ratio of the number of members of A to the total number of elementary events."). In terms that will be used later, this formulation involves a "relative frequency" interpretation of probability statements. See *id.* at 25-28.

98. Each such event is usually referred to as a "Bernoulli trial," if there are only two outcomes (drawing a red ball or drawing a white ball). Moreover, because we are sampling independently with replacement, so that the probability of drawing a red ball remains unchanged from draw to draw, the sampling constitutes a "stationary" Bernoulli process. See W. HAYS, *supra* note 91, at 128-31.

99. The probability of any given sequence of N independent Bernoulli trials is given by

$$p^r \times q^{N-r},$$

where r is the number of "successes" (here, drawing a red ball), N-r is the number of "failures" (drawing a white ball), p is the probability of a success on each trial, and q is the probability of a failure ($q = 1 - p$). W. HAYS, *supra* note 91, at 128-29.

In addition, the number of distinct sequences by which r successes can occur within N trials is given by the *binomial coefficient*, written as

$$\binom{N}{r}$$

and defined as

$$\frac{N!}{r!(N-r)!}.$$

Id. at 125-26, 129-30. Thus, the probability of observing exactly r successes in N independent trials, regardless of the order in which the successes occur, is:

$$pr(r, N, p) = \binom{N}{r} p^r q^{N-r},$$

where p is the probability of a success in any given trial. *Id.* at 130.

100. The probability of drawing exactly 4 red balls and 6 white balls when $p = 0.04$ is (approximately):

ably false, and we can reject it with a substantial amount of confidence.¹⁰¹

As it turns out, the hypothesis yielding the greatest number of 10-ball simple random samples containing exactly 4 red balls is the hypothesis that the box contains 40 red balls.¹⁰² In other words, the hypothesis that makes a sample of 40% red balls *most* likely is the hypothesis that 40% of the balls in the population are red.¹⁰³ A population with only 39 red balls (or one with 41 red balls) would also yield a 4-red-ball sample with a rather high probability, so the hypothesis of 40% is not significantly stronger than an hypothesis of 39% (or 41%).¹⁰⁴ An hypothesis of 40 red balls *is* substantially stronger, how-

$$\binom{10}{4} (0.04)^4 (0.96)^6 = (210)(0.0000026)(0.7827578)$$

$$= (210)(0.000002)$$

$$= 0.00042.$$

101. On similar reasoning, the hypothesis that 94 out of the 100 balls are red would also be rejected. So would hypotheses that the box contains exactly 5 or 93 red balls, and so forth.

102. If 40 of the 100 balls are red, then the probability *p* for any Bernoulli trial is 0.4, *q* is 0.6, and the probability of drawing exactly 4 red balls out of 10 is (approximately):

$$\binom{10}{4} (0.4)^4 (0.6)^6 = (210)(0.0256)(0.046656)$$

$$= (210)(0.0011944)$$

$$= 0.2508.$$

Thus, if there are 40 red balls in all, the probability of drawing a sample with 4 red balls is slightly greater than $\frac{1}{4}$.

103. See W. HAYS, *supra* note 91, at 195-97 (if no prior information at all about the value of the population proportion *p*, then the maximum likelihood estimate of *p* would be the sample proportion *P*, since among all possible values of *p* this value makes the actual sample *P* have the greatest a priori likelihood).

104. The probability of drawing a 4-red-ball sample out of a population containing 39 red balls is (approximately):

$$\binom{10}{4} (0.39)^4 (0.61)^6 = (210)(0.0231344)(0.0515204)$$

$$= (210)(0.0011919)$$

$$= 0.2503.$$

The probability from a 41-red-ball population is (approximately):

ever, than an hypothesis of, say, 20.¹⁰⁵

Scientists have adopted the convention that hypotheses are generally rejected, for scientific purposes, if the sample actually drawn is in that subset of least likely samples that collectively has a probability less than 0.05 (less than 1 chance in 20).¹⁰⁶ Therefore, a sample result is routinely said to be "*statistically significant*" with respect to an hypothesis if the probability of drawing a statistic at least as extreme as that actually drawn is less than 0.05; such sampling results are conventionally regarded as a proper basis for rejecting the hypothesis.¹⁰⁷ If the sampling results are not statistically significant, the hypothesis should not be rejected, and no conclusion is justified about the truth of the hypothesis on the basis of this sample (given the 0.05 convention).

There are several techniques for reducing sampling uncertainty. The obvious method, increasing the size of the sample, does so in two

$$\begin{aligned} \binom{10}{4} (0.41)^4 (0.59)^6 &= (210)(0.0282576)(0.0421805) \\ &= (210)(0.0011919) \\ &= 0.2503. \end{aligned}$$

The probability distribution for drawing 4 red balls out of 10 has its highest single value for a population that has 40% of its balls colored red.

105. Whereas drawing a simple random sample of 4 red balls has a probability of 0.2508 given a box with 40 red balls, *see supra* note 102, drawing such a sample has a probability of only 0.088 given a box with with 20 red balls:

$$\begin{aligned} \binom{10}{4} (0.2)^4 (0.8)^6 &= (210)(0.0016)(0.262144) \\ &= (210)(0.0004194) \\ &= 0.088. \end{aligned}$$

106. *E.g.*, M. BLAND, *AN INTRODUCTION TO MEDICAL STATISTICS* 152 (1987) (medical sciences); J. COHEN & P. COHEN, *supra* note 78, at 20-21 (behavioral sciences); H. LOETHER & D. MCTAVISH, *supra* note 37, at 508-09 (sociology); Cowles & Davis, *On the Origins of the .05 Level of Statistical Significance*, 37 *AM. PSYCHOLOGIST* 553, 553 (May 1982); Ware, Mosteller & Ingelfinger, *P Values*, in *MEDICAL USES OF STATISTICS*, 149, 155-158 (J. Bailar III & F. Mosteller eds. 1986) (medical literature).

There appears to be no compelling reason for making the so-called "*critical value*" 0.05, or the "*critical region*" for rejecting hypotheses the range of probabilities less than 0.05 (prob. < 0.05). For a discussion of the origin of the 0.05 convention and criticisms of its mechanical application, see references cited *infra* note 127.

107. *See generally* M. BLAND, *supra* note 106, at 148-62; W. HAYS, *supra* note 91, at 249-63; H. LOETHER & D. MCTAVISH, *supra* note 37, at 499-526.

ways. First, increasing the size of a sample drawn without replacement can decrease the range over which our uncertainty extends, at least when the population size is finite.¹⁰⁸ Observing even one more ball (for example, another red one) would tell us more about the composition of the population (namely, that at least 5 balls are red). Second, probability distributions for statistics such as proportions or arithmetic means are a function of sample size, with the result that the number of rejectable hypotheses increases with the increase in sample size.¹⁰⁹

Other techniques for reducing sampling uncertainty relate to sampling design. For instance, sampling a small population without replacement can increase the precision of the sample estimate compared to sampling with replacement.¹¹⁰ And in some situations, a *stratified random sample* can have a smaller sampling error than a simple random sample of the same size.¹¹¹ Thus, variations in sampling design can directly affect the amount of residual sampling uncertainty.

A primary function of significance testing is to provide a means of characterizing one type of residual sampling uncertainty.¹¹² The probability of drawing a certain sample, given an hypothesis about the population, is an indirect measure of the uncertainty that can be introduced by the sampling process.¹¹³ If we can calculate the probability that the statistic in our sample would be drawn from a population ac-

108. Cf. M. FINKELSTEIN & B. LEVIN, *STATISTICS FOR LAWYERS* 261 (1990) (for small populations, sampling without replacement produces somewhat more precise estimates than sampling with replacement, "because as the sample size increases to an appreciable fraction of the population, it becomes increasingly unlikely that the sample mean will vary by a given amount from the population mean"); W. HAYS, *supra* note 91, at 205-06 (variance of sampling distribution of the mean "tends to be *somewhat smaller*" for a fixed sample size N from a finite population than from an infinite population).

109. See, e.g., W. HAYS, *supra* note 91, at 237-38.

110. See *supra* note 108.

111. H. LOETHER & D. MCTAVISH, *supra* note 37, at 418. In a stratified random sample, "the population is divided into subpopulations (strata) and then simple random samples are drawn from each subpopulation." *Id.* at 418. This sampling method can reduce sampling error, when compared to a simple random sample of the same size, if the strata or subpopulations are determined using variables correlated with the variables of the study, thus producing more homogeneity in the subpopulations than exists in the population. *Id.* at 418-19.

112. See *supra* note 94.

113. This measure is "indirect" in the sense that it is not a direct measure of our uncertainty about the true value of the parameter, but rather a measure of the likelihood of drawing a particular sample (statistic) given a particular hypothesis about the population. E.g., W. HAYS, *supra* note 91, at 236 (the probability statement is not really about the population parameter, but about samples); cf. Ware, Mosteller & Ingelfinger, *supra* note 106, at 154 ("popular misconception is that the P value is the probability that the null hypothesis is true"). A subjective interpretation of the meaning of probability statements, however, might allow us to conceptualize this as a direct measure of our uncertainty about the parameter. See *infra* text accompanying note 213.

curately characterized by some particular hypothesis, then that probability is a measure of our risk of being wrong if we were to reject that hypothesis as our estimate of the population parameter. If we reject hypotheses only when our sampling results have a probability of less than 0.05, our rate of rejecting true hypotheses should be less than 1 case in 20. Once scientists had developed quantitative methods for assessing such a risk, this achievement was followed by the informal adoption of conventional quantitative thresholds for rejecting hypotheses, for deciding when such sampling uncertainty is "sufficiently low" for scientific purposes.¹¹⁴

This reasoning has been relied upon in various ways by decisionmakers in legal settings. Courts trying claims of unlawful discrimination routinely use such calculations of probability to determine whether a plaintiff has presented a *prima facie* case under either the fourteenth amendment or Title VII.¹¹⁵ In such cases, the probability of drawing a sample with a particular racial or sexual composition (for example, the proportion of minority or women candidates actually hired) in a simple random manner from the population (those eligible to be hired) is taken as an indication of whether the selection process (the hiring process) was actually unbiased with respect to race or sex.

Courts and agencies also rely upon significance testing for the same purpose as scientists: to estimate values of unknown parameters.¹¹⁶ In such cases, decisionmakers often find the generation of confidence intervals a more useful form of information than the probabilities deduced by significance testing.

114. See *supra* text accompanying notes 106-07.

115. See, e.g., *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 650-55 (1989) (statistically significant disparate impact, together with showing that specific employment practices caused the disparity, can make out *prima facie* case of employment discrimination under Title VII of Civil Rights Act of 1964); *Hazelwood School Dist. v. United States*, 433 U.S. 299, 311 n.17 (1977) (if deviation of actual number of black teachers hired from expected number is more than "two or three standard deviations," this would "undercut the hypothesis that decisions were being made randomly with respect to race"); *Castaneda v. Partida*, 430 U.S. 482, 496 & n.17 (1977) (*prima facie* case of discrimination against Mexican-Americans in grand jury selection established if difference between expected number of Mexican-Americans to be selected and actual number selected is "greater than two or three standard deviations"); *Palmer v. Shultz*, 815 F.2d 84, 90, 96 (D.C. Cir. 1987) (statistically significant deviations at 0.05 level in either direction from equality in selection rates constitute *prima facie* case of unlawful discrimination). See generally *Statistical Decision Theory*, *supra* note 5.

116. E.g., Guidelines for Carcinogen Risk Assessment, 51 Fed. Reg. 33,992, 33,997 (EPA 1986) (quantitative risk extrapolations generally performed only on data from animal studies for tumor sites showing "statistically significant elevations" in tumor incidence).

B. Confidence Intervals

The second principal method of classical statistical theory for characterizing sampling uncertainty is the construction of *confidence intervals*.¹¹⁷ This technique is closely related to significance testing in its underlying rationale. In effect, the technique takes the set of all possible hypotheses about the parameter and divides the hypotheses into those that can and cannot be rejected on the basis of the sample drawn. By reporting a confidence interval for red balls, for example, of 10 - 70, we are saying that, given our sample, hypotheses of less than 10 and greater than 70 should be rejected.

All the hypotheses about the true population value that fall *outside* the confidence interval can be rejected on the basis of the sample. In order to construct any particular interval, of course, we need to identify the threshold probability below which hypotheses should be rejected. This probability is usually the same conventional 0.05 discussed above.¹¹⁸ In other words, the sample results are statistically significant for all hypotheses outside the confidence interval.¹¹⁹ The hypotheses *within* the confidence interval are those that cannot be rejected on the basis of the sample (the sample results are not statistically significant for those hypotheses). A confidence interval, therefore, is a means of identifying which hypotheses can be rejected, given the sample and the threshold probability selected for rejecting hypotheses.

Confidence intervals are also explained in terms of "estimates," instead of hypotheses.¹²⁰ We sometimes use sample results to make a single *point estimate* of the parameter. Thus, in our example of red and white balls, we can use the percentage of red balls in our sample (40%) to estimate that 40% of the balls in the population are red.¹²¹ By so restricting our estimate to a single value, however, we incur a

117. The following are illustrations of notation for confidence intervals: " $\dots \leq x \leq \dots$ "; " $\dots \leftrightarrow \dots$ "; " $CI = x \pm \dots$ "; " $CI = \dots - \dots$ ".

118. See *supra* text accompanying notes 106-07.

119. J. COHEN & P. COHEN, *supra* note 78, at 63; W. HAYS, *supra* note 91, at 206-09, 235-39 (if hypothetical value for parameter not covered by confidence interval, then hypothesis may be rejected).

120. See generally M. BLAND, *supra* note 106, at 134-45; M. FINKELSTEIN & B. LEVIN, *supra* note 108, at 171-81, 227; W. HAYS, *supra* note 91, at 206-09, 235-39.

121. It can be demonstrated theoretically that the proportion in a sample is an unbiased, maximum-likelihood estimator of the proportion in the population, provided the sample (of size N) is drawn as a result of N independent trials and the probability of a specified outcome (here, "drawing a red ball") remains the same throughout the trials. That is, in the long run, the mean of the proportions in all possible such samples of the same size is identical to the population proportion. W. HAYS, *supra* note 91, at 128-39, 195-99, 240-41.

sizeable risk of being wrong. If, however, we use our sample statistic to generate an *interval estimate* (a range of values) for the parameter, then we can increase the likelihood that our estimate captures the parameter somewhere *within* it. In our example of sampling without replacement, we could be *certain* that the interval estimate of 4 - 94 red balls would capture the true value for the population. We can narrow our interval estimate if we are willing to accept some degree of confidence short of certainty. We can choose any specified probability for constructing a confidence interval.¹²² For example, a *95% confidence interval* is broad enough to give us at least a 0.95 probability (19/20 chance) of containing the true parameter value.

Under the classical theory of probability, interpreting the meaning of a confidence interval requires some care in terminology. With a 95% confidence interval, we draw the conclusion that the probability is at least 0.95 that the sample drawn and confidence interval constructed is one of those samples and intervals that includes the parameter within it.¹²³ The method is driven, therefore, by assumptions about the sampling process, the probability distribution for statistics given that process, and the selection of a "level of confidence" for constructing the intervals. Just as with statistical significance in hypothesis testing, it is a matter of convention to use a 95% confidence interval,¹²⁴ although higher levels of confidence are sometimes used by scientists.¹²⁵

Nothing in the reasoning behind hypothesis testing requires statistical significance to be identified with 0.05 probability or the level of confidence to be 95% or greater. It may simply be that once sampling uncertainty has been quantified using probability theory, the simplest answer to the question of what level of uncertainty should be acceptable is to pick a probability by consensus. This approach, however, would hardly seem defensible if it were adopted by decisionmakers in legal contexts.¹²⁶ Deciding when sampling uncertainty is sufficiently

122. See *id.* at 235-37.

123. *Id.*; J. COHEN & P. COHEN, *supra* note 78, at 62.

124. See, e.g., J. COHEN & P. COHEN, *supra* note 78, at 52, 63; W. HAYS, *supra* note 91, at 235-37, 249-63; Cowles & Davis, *supra* note 106, at 553.

125. Ninety-nine percent confidence intervals are not uncommon. See, e.g., Ware, Mosteller & Ingelfinger, *supra* note 106, at 155-57. Often a lower probability than 0.05 is reported if it is significantly lower. The usual benchmarks for reporting lower probabilities are decreasing orders of magnitude: ≤ 0.01 , ≤ 0.001 , etc. See J. COHEN & P. COHEN, *supra* note 78, at 20-21 (behavioral sciences).

126. See J. COHRSSSEN & V. COVELLO, *supra* note 27, at 92 (use of confidence intervals is a relatively common method used by analysts of environmental risks to limit problems with extreme upper- and lower-bound risk estimates, but there is an implicit policy judgment in selecting 95%

low for legal purposes should not be a matter of mere quantitative convention.¹²⁷ In the case of socially significant decisions, the decision whether to proceed on the basis of data that creates a given level of sampling uncertainty should take into account such factors as the costs and benefits of so proceeding. Such factors vary from case to case and are not taken into account at all by the scientific convention. When real errors, and real costs and benefits, hang in the balance, it might well be unreasonable for a decisionmaker to choose not to rely on study results merely because they have not quite crossed the currently conventional bright line into "statistical significance."

Decisionmakers would be better served by scientific information that includes a range of confidence intervals, covering the range of levels of confidence in which the decisionmaker might be interested. Such a range would show the decisionmaker the sensitivity of the confidence intervals to the choice of a level of confidence. Reporting only a 95% confidence interval might not tell the decisionmaker all she or he needs to know about the residual sampling uncertainty.

V. MODELING UNCERTAINTY

In addition to recording observed values for variables and generalizing from those limited observations to populations, scientists use mathematics to relate multiple variables: values for one variable are expressed as a mathematical function of values for other variables.¹²⁸

as a measure of confidence).

127. See *Confidence in Probability*, *supra* note 13, at 409-417 (appropriate level of significance requires consideration of both possibility of errors and the costs associated with error; conventional but arbitrary test of statistical significance for scientific hypotheses has been applied in legal system without critical analysis); *Statistical Decision Theory*, *supra* note 5, at 364 (appropriate level of significance in discrimination cases is a "legal issue" that depends upon what is at risk in making a Type I error).

For scientific discussions consistent with this view, see generally W. HAYS, *supra* note 91, at 246-61, 281-82 ("conventions about significant results should not be turned into canons of good scientific practice"); McCloskey, *The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests*, 75 AEA PAPERS & PROCEEDINGS 201 (1985) ("Roughly three-quarters of the contributors to the *American Economic Review* misuse the test of significance."); Silber & Kaizer, *Loss Weighting and the Human Cost of Experimentation*, 38 J. CHRONIC DISABILITY 507 (1985) (arguing that choice of level of statistical significance in medical clinical studies is a function of implicit decision concerning relative importance of future vs. present patients); Ware, Mosteller & Ingelfinger, *supra* note 106, at 155-56 (the popular scientific convention has disadvantage of suggesting "a rather mindless cutoff point, which has nothing to do with the importance of the decision to be made or with the costs and losses associated with the outcomes"). But cf. Cowles & Davis, *supra* note 106, at 553 (detailing history behind adoption of the convention, and suggesting that the choice was related to the earlier concept of "probable error").

128. See, e.g., W. HAYS, *supra* note 91, at 341-48 (providing example of linear function be-

Such mathematical functions provide the basis for prediction or for testing an explanatory theory. For example, if groups living in proximity to electric power transmission lines were to experience a higher than normal incidence of childhood leukemia, then the mathematical relationship between exposure to electromagnetic fields and developing leukemia might be evidence of causation and might provide guidance for regulatory decisions about permissible duration of exposure and field strength.¹²⁹

Whenever we express one variable as a mathematical function of another variable, we create the potential for a kind of error not yet discussed in this Article. *Modeling uncertainty* arises whenever a claim is made that variable Y has a particular mathematical relationship to variable X. In general, modeling errors are made either through choosing the wrong mathematical function or by incorrectly specifying its constants. For example, if someone were to claim that for every milligauss of alternating magnetic field to which one is exposed for some specified period of time the risk of leukemia increases by "two times," he might be in error because simple multiplication is not the correct form of the relationship (perhaps risk increases exponentially, instead of as a simple product). On the other hand, the number "two" in "two times" might be incorrect; the correct number might be "three," or perhaps "one-half." Modeling uncertainty—the potential for such modeling errors—is thus created whenever two or more variables are related to each other mathematically.

Modeling error results in descriptive error about the world, generally in the form of predictive error: the predictions of the values for a variable turn out to be false.¹³⁰ Predictive error, however, is not identi-

tween two variables). Often, mathematical functions relate statistics for one variable (such as a percentage or the mean) to statistics for other variables, instead of relating individual values to individual values. When I refer to "values of a variable" in this context, I am referring to either particular values or statistics based on those values.

129. See, e.g., Scientific Advisory Panel, New York State Power Lines Project, Final Report on Biological Effects of Power Line Fields 72-87 (July 1, 1987) (part of research program conducted under agreement between New York State Public Service Commission and New York Power Authority, and administered by New York State Department of Health).

130. Although mathematical modeling is an important step toward the establishment of causal theories, which are discussed in Section VI *infra*, prediction is still a major objective behind mathematical modeling. The traditional use of regression analysis in the behavioral sciences has been for prediction, with only incidental attention to explanation or causal analysis. J. COHEN & P. COHEN, *supra* note 78, at 41, 111. Although this emphasis has been changing, prediction still remains an important role for mathematical modeling. *Id.* at 111-15. An illustration of the usefulness of predictor variables, regardless of underlying causal networks, can be found in DeTray, *Veteran Status as a Screening Device*, 72 AM. ECON. REV. 133 (1982) (employing multiple regres-

cal to modeling error. Errors in predictions can result from conceptual error, measurement error, or sampling error, as well as from modeling error. If we select or define variables inappropriately, make errors in measurement, or generalize improperly, we can be led to false predictions, even if our mathematical model happens to be the correct one. On the other hand, even in situations where a completely valid and reliable measurement method is used to classify instances correctly, modeling error might still occur once we try to express one variable as a mathematical function of another variable. In order to focus attention on modeling error alone, I will assume that all predictive error in the following analysis is due to modeling error.

This section of the Article provides an extended example of modeling uncertainty by discussing *linear regression models*. These models are commonly used for predictive purposes in science.¹³¹ Their use by agencies and courts has grown tremendously, addressing problems as diverse as investigating the economic effects of rulemaking or determining the existence of past discrimination or voting dilution.¹³² Regression models also provide an important foundation for claims of causality.¹³³

A. *Bivariate Linear Regression*

A bivariate linear regression model expresses one variable as a "linear" mathematical function of one other variable. The variable whose values are being predicted is the *dependent variable*; the variable

sion model to test hypotheses about the predictive value of veteran status for future worker productivity); see also J. COHEN & P. COHEN, *supra* note 78, at 114-15 (discussing techniques useful only for prediction of values of the dependent variable, but not for causal analysis). An illustration of models being constructed primarily for predictive purposes is provided by W. BERRY & S. FELDMAN, *MULTIPLE REGRESSION IN PRACTICE* 59 (1985) (using polynomial transforms to achieve linearity).

131. See generally M. BLAND, *supra* note 106, at 188-215 (medicine); J. COHEN & P. COHEN, *supra* note 78 (behavioral sciences); W. HAYS, *supra* note 91, at 544-733 (experimental psychology); H. LOETHER & D. MCTAVISH, *supra* note 37, at 232-47, 306-40 (sociology); Godfrey, *Simple Linear Regression in Medical Research*, in *MEDICAL USES OF STATISTICS* 170 (J. Bailar III & F. Mosteller ed. 1986) (focusing on examples from articles published in the *New England Journal of Medicine*).

132. See generally M. FINKELSTEIN & B. LEVIN, *supra* note 108, at 323-467; *Regression Studies*, *supra* note 5; Fisher, *supra* note 7; Finkelstein, *supra* note 7. E.g., *Bazemore v. Friday*, 478 U.S. 385 (1986) (employment discrimination); *Campos v. City of Baytown, Texas*, 840 F.2d 1240 (5th Cir. 1988), *cert. denied*, 488 U.S. 1002 (1989) (voting dilution). For an illustration of using multiple regression to determine the incremental effect of regulation, see Spiller, *supra* note 7 (airline regulation). For a critique of a multiple regression model on a regulatory matter, see Graham & Garber, *Evaluating the Effects of Automobile Safety Regulation*, 3 J. POL'Y ANALYSIS & MGMT. 206 (1984).

133. See *infra* text accompanying notes 163-85.

used in making the prediction is the *independent variable*.¹³⁴ On a graph, the values of the independent variable X are conventionally laid out along the horizontal axis, increasing from 0 (at the intersection with the vertical or Y axis) toward the right. See Figure 1. The values of Y are scaled along the vertical axis. Each instance can then be located as a point on the graph by identifying its value for each of the two variables (for example, $X = 25$, $Y = 20$).

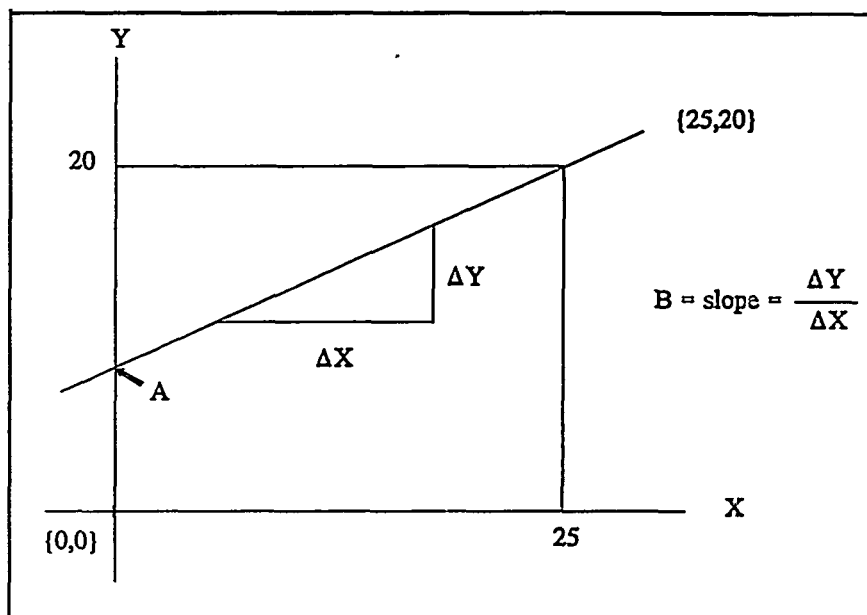


Figure 1

A model is *linear* if the form of the mathematical function defines a straight line.¹³⁵ The straight line sloping up and to the right in Figure

134. Medical literature uses the terminology *outcome variable* or *response variable* (referring to the dependent variable) and *predictor variable* (referring to the independent variable). See M. BLAND, *supra* note 106, at 190; Godfrey, *supra* note 131, at 170-71.

135. The algebraic form of the model is:

$$\hat{Y}_i = A + B(X_i),$$

where \hat{Y}_i is the predicted value of the dependent variable Y for instance i and X_i is the value of the variable X for that same instance i . A is called the *Y-intercept* or *regression constant*; B is

1 is the geometric interpretation of a positive linear function. For every value of X (for example, 25) the line relates some single value of Y (here, 20). This model is "linear" because it assigns a constant amount of change to the value of the dependent variable (Y) for each unit change in the value of the independent variable (X) over the relevant range of values of the independent variable.¹³⁶ This can be seen from Figure 1, which depicts that when $X = 0$, $Y = A$ (hence, A is called the "Y-intercept," the point at which the line crosses the Y axis), and that as the value of X increases, the value of Y increases at a constant rate B .¹³⁷

If a bivariate linear regression model is used for prediction, then the value of Y that is predicted for any value of X is provided by the line defined by the model.¹³⁸ In general, then, predictive error occurs when the true value of Y deviates from the predicted value. Geometrically, a plot of paired actual values for X and Y can be constructed, as in Figure 2; predictive error occurs whenever the plotted points do not fall precisely on the straight line that is used in making the predictions.¹³⁹

called the *regression coefficient* or *slope* for predicting Y from X . See generally J. COHEN & P. COHEN, *supra* note 78, at 11-12, 41-44; H. LOETHER & D. MCTAVISH, *supra* note 37, at 242-47; L. SCHROEDER, D. SJOQUIST & P. STEPHAN, UNDERSTANDING REGRESSION ANALYSIS: AN INTRODUCTORY GUIDE 11-17 (1986); Godfrey, *supra* note 131, at 171-81.

Although regression models are most often applied in practice to sample data, the models as such are applicable to population data as well. Thus, when sample-based estimates are made of parameters of the regression model for the population, significance testing and the construction of confidence intervals can be used to characterize sampling uncertainty. See, e.g., J. COHEN & P. COHEN, *supra* note 78, at 62-65; W. HAYS, *supra* note 91, at 571-73, 588-93. The discussion in this section assumes that the models are being applied to sample data or to completely enumerated population values, in order to emphasize that modeling uncertainty is logically distinct from the potential for sampling error.

136. J. COHEN & P. COHEN, *supra* note 78, at 27; Finkelstein, *supra* note 7, at 1448; Godfrey, *supra* note 131, at 171.

137. A straight line is defined by point $[0, \alpha]$ and slope β . See, e.g., M. BLAND, *supra* note 106, at 188-189; W. HAYS, *supra* note 91, at 545-48.

138. In practice, of course, the actual prediction of the value of Y for a given value of X will be made by using the algebraic definition of the line, see *supra* note 135, not the geometric depiction as such. The predicted value of Y would lie precisely on the line defined by the model.

139. As mentioned earlier, *supra* text accompanying note 130, such predictive error can also occur when measurements of Y are in error. The true value of Y *might* lie on the prediction line, and no modeling error is present, but an error in measurement might hide this fact by giving an observed value for Y that does not fall on the prediction line. Similarly, when measurements of independent variables contain errors, the modeling results can be expected to be affected. E.g., Graham, *supra* note 132, at 212-13.

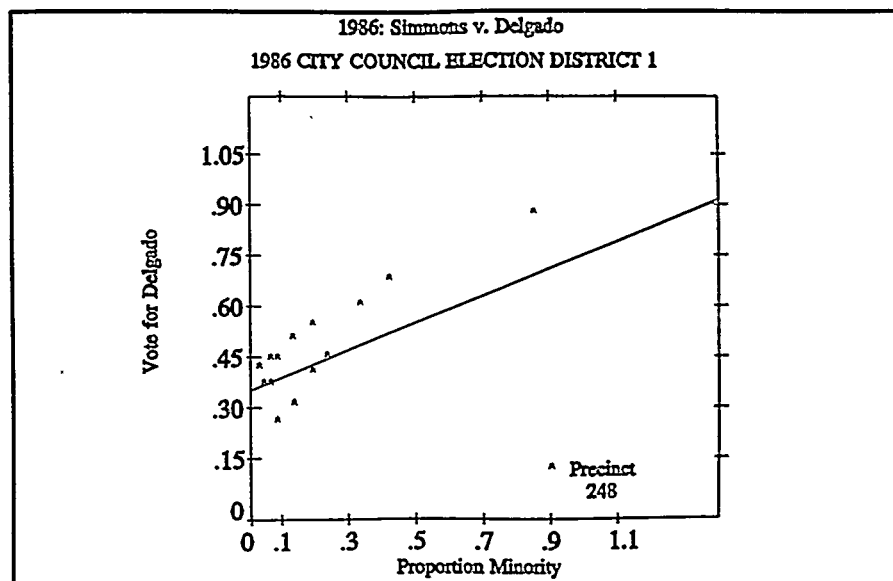


Figure 2

Source: Campos v. City of Baytown, Texas, 840 F.2d 1240, 1251 (5th Cir. 1988) (regressing proportion of vote for minority candidate on proportion minority in population, by precinct; Y-intercept (A) = 0.37, slope (B) = 0.46).

There may be an infinite number of straight lines that could be used as predictive models for any given set of data. Selecting a *particular* straight line occurs when particular values are chosen for the constants A and B. If predictive error shows up as the difference between the predicted value (a point on the line) and the observed value (a point determined by actual data, which may or may not be on the line),¹⁴⁰ the amount of modeling error that results is clearly dependent upon which of many possible lines is chosen as the predictive model.

The most commonly used measure of *average error* associated with a model is the average of the squared differences between predicted values for Y and observed values of Y.¹⁴¹ Scientists reduce mod-

140. The observed or measured value is identical to the true value if there is no measurement error.

141. Squaring the difference between the predicted and the observed values has distinct advantages over using either the signed differences themselves or their absolute values. Squaring the differences, however, weights larger errors more than smaller errors. See generally M. FINKELSTEIN & B. LEVIN, *supra* note 108, at 332-36.

eling uncertainty by using a technique called *least squares* to select the straight line that produces the *least* amount of predictive error, measured by squared differences.¹⁴² Such a "best-fitting" linear model is called the *linear regression equation* of Y on X.

Identifying the linear regression model for any data set selects the "best" of all possible linear models, but does not ensure that there is no residual modeling uncertainty. This residual modeling uncertainty is reflected in the remaining scatter of the observed values of Y around the best-fitting regression line. Figure 2 provides an illustration, in which the proportion of vote for a minority candidate (by precinct) in a city council election was regressed on the proportion of minority population.

Scientists characterize the residual modeling uncertainty in different ways, depending upon the uses to which the modeling will be put.¹⁴³ Probably the most frequently encountered approach is to compare the residual error using the regression model to the error that would result if the mean of Y alone were used as a predictor. This approach tells us how much better off we are using the regression line to predict Y, instead of simply predicting Y using the mean of all the values of Y. Such a comparison is a means of gauging whether it is better to use the values of X and the regression line to predict Y, or whether we should simply ignore X and base our prediction on Y values alone.

Scientists are usually concerned with whether their regression models *improve* predictive power. They therefore assess how good a model is by determining the amount of predictive error that is *eliminated* by using the model. The measures generally used are the *coefficient of determination* (symbolized as r^2) and its square root, called Pearson's *correlation coefficient* (r).¹⁴⁴ The coefficient of determination

142. The mathematical technique of "least squares" is so named because it picks out the linear model (prediction line) that has the least average squared error for the data presented. See generally M. BLAND, *supra* note 106, at 191-95; J. COHEN & P. COHEN, *supra* note 78, at 42-43, 50, 77; H. LOETHER & D. MCTAVISH, *supra* note 37, at 247-50; L. SCHROEDER, D. SJOQUIST & P. STEPHAN, *supra* note 135, at 17-23.

143. Two direct measures of this residual modeling uncertainty are the *variance of residual error* and the *standard deviation of residual error*. The variance of residual error is sometimes referred to as the *variance of residuals*, and the standard deviation of residual error as the *standard deviation of residuals*. J. COHEN & P. COHEN, *supra* note 78, at 47-48. The variance of residual error is the *average* of the squared differences between observed scores and predicted scores, and the standard deviation of residual error is the square root of that average. See *id.*

144. W. HAYS, *supra* note 91, at 554-60; H. LOETHER & D. MCTAVISH, *supra* note 37, at 250-55; L. SCHROEDER, D. SJOQUIST & P. STEPHAN, *supra* note 135, at 26; Finkelstein, *supra* note 7, at 1448-53. The convention of the small roman letter r (instead of the Greek letter rho) is used in this section to emphasize that this is a measure of residual error in *observed* data (such as a sample or a completely enumerated population). Sampling error is logically distinct from model-

is the proportion of average predictive error that is eliminated by using the regression line instead of the mean of Y as a predictive rule.¹⁴⁵ When $r^2 = 1$ (and hence $r = 1$), the dependent variable is *perfectly* predicted by the independent variable: all the actual values of Y fall squarely on the regression line, and there is no residual modeling error. When $r^2 = 0$ (and hence $r = 0$), there is no linear relationship at all between the variables, and the regression line is no better as a predictive rule than is the mean of Y . In such a situation, the linear regression model has no predictive value at all. As r^2 and r take on values from 0 to 1, they indicate increased correlation between the variables and decreased predictive error.¹⁴⁶ Thus, r^2 and r are commonly used to characterize the *strength of the linear association* between two variables.¹⁴⁷

ing error, see *supra* note 135, and is being ignored in this discussion of modeling error.

The correlation coefficient is different from, but related to, the regression coefficient. The correlation coefficient is a symmetrical measure of association between the two variables, while the regression coefficient is asymmetrical because it predicts the dependent variable using the independent variable. The value of the correlation coefficient would be identical to that of a regression coefficient calculated for standardized scores of the two variables. See, e.g., W. HAYS, *supra* note 91, at 554-58; H. LOETHER & D. MCTAVISH, *supra* note 37, at 255-59; L. SCHROEDER, D. SJOQUIST & P. STEPHAN, *supra* note 135, at 28-29; Godfrey, *supra* note 131, at 181-88.

145. The average amount of predictive error that is associated with using the mean of Y as the predictor for any Y_i is the variance of Y . The variance of Y is the average of the squared differences between the actual score for each instance (Y_i) and the mean of Y . This method of characterizing residual error for the mean of Y as a predictor uses the same measure of error as the regression line—namely, average squared difference. See *supra* text accompanying notes 141-42; see, e.g., H. LOETHER & D. MCTAVISH, *supra* note 37, at 247-49.

The ratio of the residual modeling error for the regression line to the variance of Y , therefore, is a useful measure of the predictive success of the model, for that ratio provides the proportion of error that is *not eliminated* by using the regression line instead of the mean as a predictor. See J. COHEN & P. COHEN, *supra* note 78, at 47-48 (referring to the square root of this ratio as the "coefficient of alienation"; for standardized scores, this can be thought of as the "coefficient of noncorrelation").

The coefficient of determination, r^2 , is therefore an indirect measure of residual modeling uncertainty in the sense that, instead of measuring residual error directly, it measures the proportion of Y 's variance that is *eliminated* using the regression equation, or the proportion of Y 's variance that is linearly associated with X . See generally J. COHEN & P. COHEN, *supra* note 78, at 34-36, 46-48; H. LOETHER & D. MCTAVISH, *supra* note 37, at 250-55; L. SCHROEDER, D. SJOQUIST & P. STEPHAN, *supra* note 135, at 23-29.

146. A correlation coefficient r that is derived for a *sample* data set should be subjected to significance testing or should have a confidence interval constructed if the conclusion to be reached is about the degree of correlation in the *population*. See, e.g., J. COHEN & P. COHEN, *supra* note 78, at 51-59, 62-65 (significance testing and confidence intervals for r^2 and r); Finkelstein, *supra* note 7, at 1449-53; Fisher, *supra* note 7, at 716-20. This fact emphasizes the logically distinct natures of sampling and modeling uncertainties.

147. The correlation coefficient has already been encountered in earlier discussions in this Article, although we were not then in a position to draw attention to it. In the case of criterion

At the beginning of this section, I suggested two fundamental ways to generate modeling error: the wrong choice of function form and the wrong choice of constants. Identifying the best-fitting linear regression line minimizes only the latter source of error, by selecting the best constants (A and B) for the linear equation.¹⁴⁸ But uncertainty is also created by the decision to use a *straight line* algebraic form as a predictive rule instead of using a nonlinear model. Perhaps the selection of a curved line would result in even less predictive error than the use of even the best-fitting straight line.¹⁴⁹ Thus, although the least squares technique can be used to select the *linear* model with the least residual modeling error, some *nonlinear* mathematical model might produce an even better fit for the data.¹⁵⁰ Of course, each mathematical model, whether linear or nonlinear, will contain its own mathematical assumptions and constants and will thus create its own potential for modeling error.¹⁵¹

B. Multiple Linear Regression

Agencies and courts have been increasingly confronted with multiple regression models, in which the dependent variable is related mathematically to more than one independent variable.¹⁵² These models introduce modeling uncertainty in a way similar to bivariate models,¹⁵³ but are more difficult to depict geometrically because the addi-

validity of measurement, *supra* text accompanying notes 67-70, the *validity coefficient* is the correlation coefficient for the criterion measurements and the measurements taken using the method being evaluated. See E. CARMINES & R. ZELLER, *supra* note 34, at 17-18; E. GHISELLI, J. CAMPBELL & S. ZEDICK, *supra* note 37, at 269; *supra* note 67. For an example of the use of regression analysis in comparing a test method of chemical analysis to a reference method, see Peters & Westgard, *supra* note 61, at 421-22.

148. See *supra* text accompanying notes 140-42.

149. See, e.g., J. COHEN & P. COHEN, *supra* note 78, at 63; H. LOETHER & D. MCTAVISH, *supra* note 37, at 255.

150. If the values in a population are not linearly related, the value of r will be near zero. This means only that the best-fitting *linear* model is not a very good predictor, compared to the mean of the dependent variable. A low r value, however, supplies no information about the goodness of fit of nonlinear models. See H. LOETHER & D. MCTAVISH, *supra* note 37, at 255.

151. Sometimes an independent variable is not itself linearly related to the dependent variable, but a mathematical transformation of the independent variable is. See, e.g., J. COHEN & P. COHEN, *supra* note 78, at 76, 253-71. In such a case, a linear model could use as a variable the transform of the original variable, and this might reduce residual modeling error.

152. See sources cited *supra* note 132; see also *McCleskey v. Zant*, 580 F. Supp. 338, 352-80, 403 (N.D. Ga. 1984), *rev'd sub nom. McCleskey v. Kemp*, 753 F.2d 877 (11th Cir. 1985), *aff'd*, 481 U.S. 279 (1987).

153. Other assumptions associated with multiple linear regression models, such as additivity and perfect collinearity, see W. BERRY & S. FELDMAN, *supra* note 130, at 37-38, 51-53, will not

tion of each independent variable amounts to the addition of a new dimension and a new axis to the geometric model. Thus, while a bivariate model can be depicted by a two-dimensional figure (such as Figure 1 *supra*), a model with three variables (1 dependent, 2 independent) requires three-dimensional geometry.¹⁵⁴

In a multiple regression model, each independent variable has its own regression coefficient relating its effect on the predicted value of Y. Such *partial regression coefficients* are measures of their associated independent variable's direct effect on the predicted value of Y, after the effects of all the other independent variables have been taken into account. The partial regression coefficient for an independent variable gives the incremental effect of that variable, after the other independent variables in the model have been statistically "held constant."¹⁵⁵

Similar to the situation with bivariate linear regression models, the amount of residual predictive uncertainty associated with the choice of a multiple linear model can be minimized through the use of the least squares technique. The multiple regression equation is that particular linear model that minimizes the overall modeling error.¹⁵⁶ The resulting set of predicted values for Y will be as close as possible to the real values of Y, given the restriction that the model must be linear--that is, that the independent variables are each to be given a single, constant weight.

be discussed here. The purpose of this Article is not to explore the intricacies of regression models, but rather to explain and illustrate the nature of modeling uncertainty.

154. J. NETER, W. WASSERMAN & M. KUTNER, *APPLIED LINEAR STATISTICAL MODELS* 226-28 (3d ed. 1990). The general form for a multiple linear regression model is:

$$\hat{Y}_i = A + B_1(X_{i1}) + \dots + B_k(X_{ik}),$$

where \hat{Y}_i is the predicted value of Y for instance i, X_1 through X_k are k independent variables and X_{ki} is the value of variable X_k for the instance i; A is the Y-intercept or regression constant, and B_1 through B_k are called the "partial regression coefficients" for the independent variables. J. COHEN & P. COHEN, *supra* note 78, at 81-83; W. HAYS, *supra* note 91, at 608-15, 621-23. The meaning of the regression constant here is similar to that in the bivariate model: A is the value of \hat{Y} when the value of each independent variable is zero. A partial regression coefficient, B_j , is the (constant) increase in \hat{Y} for a unit increase in the independent variable X_j , when the values of all the other independent variables are held constant. J. COHEN & P. COHEN, *supra* note 78, at 83-84, 98-100; W. HAYS, *supra* note 91, at 608-09. Thus, the assumption of linearity is still a condition of the multiple regression model in the sense that each independent variable is still linearly related to the dependent variable. See W. BERRY & S. FELDMAN, *supra* note 130, at 51.

155. J. COHEN & P. COHEN, *supra* note 78, at 82-85, 91-92. Thus, where the hypothesis under investigation is the incremental effect of regulation on highway fatalities, the primary interest is in the estimated partial regression coefficient of the regulation variable in the regression model. Graham & Garber, *supra* note 132, at 217.

156. See J. COHEN & P. COHEN, *supra* note 78, at 83; W. HAYS, *supra* note 91, at 621-26.

The *multiple correlation coefficient*, R , is a common measure of the linear association between the dependent variable Y and the multiple independent variables *taken as a group*.¹⁵⁷ When $R = 0$, there is no linear relationship between Y and the group of independent variables; when $R = 1$, there is a perfect linear relationship.¹⁵⁸ R therefore provides an indirect measure of the residual modeling error for the multiple linear regression model as a whole.¹⁵⁹ Finally, just as with bivariate linear regression, the modeling error that is generated using a multiple linear regression equation might be reduced or avoided if non-linear functions were used instead.

VI. CAUSAL UNCERTAINTY

Scientific information concerning causation is of fundamental importance to most social decisions. Prospective regulatory decisions are effective and efficient to the extent that we correctly anticipate their effects.¹⁶⁰ Retrospective tort adjudications, whether sounding in negligence or strict liability, depend upon proof of causation. *Prima facie* cases establishing employment discrimination under Title VII of the Civil Rights Act of 1964 must include a showing of causation between employment practices and statistical imbalances in the composition of work forces.¹⁶¹ The conceptual problems associated with causation in the law, however, are notoriously difficult and persistent.¹⁶² While this Article does not undertake a philosophical analysis of causality, some appreciation of the nature of the causal relationship is necessary in order to understand causal uncertainty and how specifying a causal system differs from positing a mathematical model. My objective is to illustrate how an assertion of causation goes beyond mathematical

157. J. COHEN & P. COHEN, *supra* note 78, at 86-88.

158. For $R = 0$, the model predicts Y no better than the mean of Y does; when $R = 1$, there is no residual modeling error ($Y = \hat{Y}$). *See id.*; H. LOETHER & D. MCTAVISH, *supra* note 37, at 344-45.

159. Also in parallel with bivariate regression, R^2 is the proportion of the variance of Y that is eliminated through the use of the regression equation as a predictive rule, while $1 - R^2$ is the proportion of modeling uncertainty remaining with use of the model. *See* J. COHEN & P. COHEN, *supra* note 78, at 86-88, 100; W. HAYS, *supra* note 91, at 630-33; H. LOETHER & D. MCTAVISH, *supra* note 37, at 344-45.

160. *See, e.g.*, Graham & Garber, *supra* note 132 (critiquing a major study that suggested that automobile safety standards had failed to save lives).

161. *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 645-46 (1989).

162. *See, e.g.*, W. KEETON, D. DOBBS, R. KEETON & D. OWEN, PROSSER AND KEETON ON THE LAW OF TORTS § 41, at 263 (5th ed. 1984) (perhaps nothing "in the entire field of law" has generated more disagreement than the nature of the "reasonable connection" that must be found in torts between the act or omission of the defendant and the damage suffered by the plaintiff).

modeling, and how such an assertion creates causal uncertainty, which is distinct from and incremental to the kinds of uncertainty so far considered.

Like the mathematical models discussed in the preceding section, causal analyses connect variables to each other. In a causal analysis, however, the relationship between variables is not (merely) a mathematical function, but rather a causal relationship. Whereas mathematical models allow us to *predict* how variables will behave, only causal analyses provide an *explanation* for how a system of variables works, *why* a system works the way it does, or why it makes sense to think of certain variables as "a system" at all. To understand how causal analyses differ from mere mathematical modeling, we must identify those distinctive characteristics of the causal relationship that are commonly accepted by scientists.

First, there is probably universal agreement among scientists that mere statistical correlation between two variables does not entail causation between the variables.¹⁶³ In the terminology of regression analysis, we can say that even if the dependent variable Y and the independent variable X are highly correlated, it does not necessarily follow that X causes Y.¹⁶⁴ Put simply, correlation does not *entail* causation.

There is, however, a logical relationship between statistical correlation and causation.¹⁶⁵ One of the principal reasons we are interested in causal analysis is so that we can predict how groups of variables will continue to function beyond the sample studied—beyond the number of

163. E.g., J. DAVIS, *THE LOGIC OF CAUSAL ORDER* 10 (1985); D. KENNY, *CORRELATION AND CAUSALITY* 1-4 (1979); *Statistical Decision Theory*, *supra* note 5, at 375. Juries are sometimes instructed that the fact that one event follows another, standing alone, is not evidence of causation. See *In re Richardson-Merrell, Inc. "Bendectin" Products Liability Litigation*, 624 F. Supp. 1212, 1267 (S.D. Ohio 1985) (Appendix D, Jury Instructions).

164. Talking about causation between *variables* might appear odd to decisionmakers, who are more accustomed to speaking of causation between *events*, not variables. For example, being exposed to a virus might be the event that makes a person become ill. But it is a short step from that locution to saying that some key feature of the first event (such as the virus) "causes" some feature of the second event. For example, we say that the virus causes the particular kind of illness. In the terminology of variables, the proposition about causation could be recast using qualitative variables (for example, being exposed to the virus or not, having the particular kind of illness or not), or quantitative variables when possible (such as duration or intensity of exposure or illness). Using such variables, scientists can study different aspects of the causal relationship between exposure to the virus and becoming ill. There is, therefore, a derivative but useful sense in which one variable "causes" another variable.

165. See J. COHEN & P. COHEN, *supra* note 78, at 15 (analysis of causation can only proceed through analysis of correlation and regression). For a discussion of the Henle-Koch-Evans Postulates used in epidemiology to make an inference from statistical associations to biological causation, see Black & Lilienfeld, *supra* note 10, at 762-64.

instances studied, beyond the measurement range studied, or beyond the time frame for which we have data. So a first condition of a causal relationship is normal concomitant change between cause and effect.¹⁶⁶ That is, a change in the level of the causing variable (usually the mean or proportion for that variable) is normally associated with a concomitant change in the level of the effect variable, unless a causal explanation is available to account for the lack of concomitant change.¹⁶⁷ For example, if exposure to alternating magnetic fields causes childhood leukemia, we would expect to find some association between exposure and leukemia (such as an increase in the number of leukemia cases as the duration of exposure increases) unless a causal explanation is available for why, in particular cases, such an association was not observed.

A second characteristic of a causal relationship is that the cause must precede the effect in time.¹⁶⁸ The move from mere prediction to causation, therefore, requires the specification of causal direction between the two variables, as well as the specification of whether the causal action occurs directly between the variables or is wholly or partially mediated by one or more other variables.¹⁶⁹ One current technique for analyzing a causal system is *path analysis*, in which causal action is symbolized by arrows between variables, and causal "paths" can be traced involving multiple variables.¹⁷⁰ For a schematic example, see Figure 3.

166. See, e.g., D. KENNY, *supra* note 163, at 2-4; J.S. MILL, A SYSTEM OF LOGIC, Book III, ch. VIII, § 6, at 466-70 (1843) (method of concomitant variation); Finkelstein, *supra* note 7, at 1449 n.27.

167. J. DAVIS, *supra* note 163, at 22.

168. J. DAVIS, *supra* note 163, at 11; D. KENNY, *supra* note 163, at 2-3. See also D. HUME, A TREATISE OF HUMAN NATURE, Book I, Part III, Section XV (1738) (Rule 2 for determining cause-and-effect: "The cause must be prior to the effect.").

169. See J. COHEN & P. COHEN, *supra* note 78, at 92-97; J. DAVIS, *supra* note 163, at 9-24.

170. J. COHEN & P. COHEN, *supra* note 78, at 356-60. See generally *id.* at 353-78; J. DAVIS, *supra* note 163; D. KENNY, *supra* note 163.

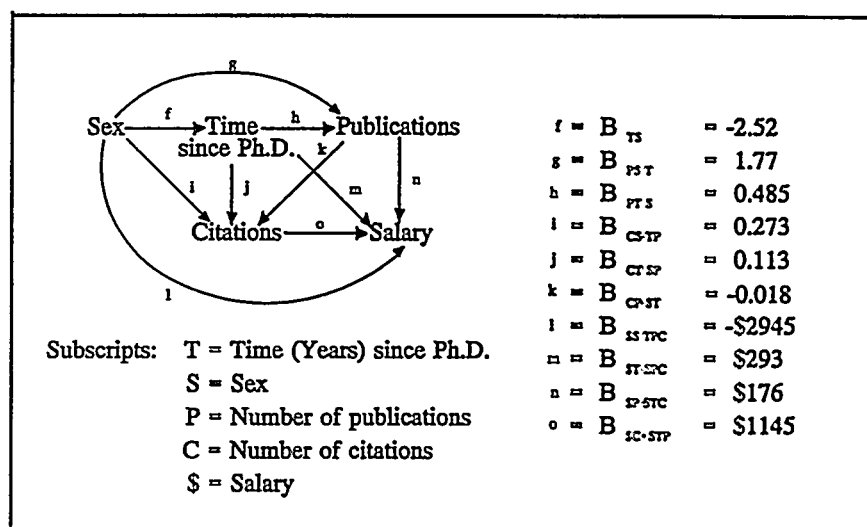


Figure 3

Source: J. COHEN & P. COHEN, *supra* note 78, at 357.

The type of causal model in Figure 3 might be relevant in an employment discrimination suit against a university over the effect of a faculty member's sex upon his or her salary. The lower-case letters "f" through "o" denote the direct causal effects of one variable upon another (in the direction of the arrow). In this system of variables, the estimate of the direct incremental effect of sex upon salary (l) is the partial regression coefficient for sex as an independent variable, when salary is regressed on the variables sex, time since Ph.D., number of publications, and number of citations to the faculty member's work in the scientific literature in the previous year. The partial regression coefficient estimates the incremental effect of sex on salary after these other independent variables have been accounted for. Of course, the total effect of sex upon salary is the sum of its direct effect and its indirect effects on salary via the other variables (e.g., sex \rightarrow publications \rightarrow salary, or sex \rightarrow time since Ph.D. \rightarrow publications \rightarrow salary). Thus, path analysis uses regression analysis to clarify the causal relationships behind the regression model.

A third characteristic of causal assertions is that they are by nature universal: they must be about a relationship obtaining in the popu-

lation, not merely in samples. If it is true that one variable has a causal effect on another variable, then this relationship must hold between any instances of those two variables, regardless of where or when those instances occur. It does not make sense to assert that a causal relationship has been observed, but that it might be the result of random sampling itself.¹⁷¹ Causal statements are essentially universal statements, and while sampling can make correlations actually appear in a given sample when no corresponding correlation exists in the population, sampling cannot be said to create causal relationships within the data.

These three related characteristics of causal assertions virtually ensure that assertions of causation can never be conclusively demonstrated to be true. Causal assertions entail the potential for causal error in a radical and ineliminable way. Whenever scientists assert that A *causes* B, that assertion necessarily has associated with it causal uncertainty. Of course, any evidence of a statistical correlation upon which the causal assertion may be based *also* carries with it sampling, modeling, measurement, and conceptual uncertainty. The move to draw causal conclusions, however, adds causal uncertainty to these other four kinds of uncertainty.¹⁷²

Causal uncertainty is a kind of uncertainty peculiar to assertions about causal relationships, as opposed to merely predictive assertions. Thus, when we reach conclusions not just about how a system of variables happens to behave, but also about why it behaves the way that it does, we create causal uncertainty. Causal uncertainty is the incremental potential for error about the existence, direction, or strength of the causal relationship itself. In this section, I first examine several common sources of causal error, and then discuss several techniques that scientists have developed to reduce the amount of causal error.

171. Statistical associations, of course, can be observed in a random sample alone, as a result of chance, and such a result would be due to the sampling process itself. That is, the sample that happened to be drawn contains a correlation, although the population from which the sample was drawn does not. Thus, it makes perfectly good sense to say that a correlation is observed in a sample that might turn out not to exist in the population. By contrast, we seem to be saying something meaningless or contradictory if we assert that a *causal* relationship in fact exists by chance in a random sample, but not in the population. If the causal assertion is not true universally, in the population of all instances, then it is simply not true at all.

172. Cf. W. BERRY & S. FELDMAN, *supra* note 130, at 26-37 (discussing effect of measurement error upon well-specified regression models).

A. *Sources of Causal Uncertainty*

A common source of causal error is drawing causal inferences on the basis of *causally spurious correlations*.¹⁷³ This occurs when we observe a statistically significant correlation between variables A and B and conclude incorrectly that A causes B.¹⁷⁴ This conclusion might be incorrect for several reasons. First, because correlation coefficients between two variables are bidirectional or symmetrical,¹⁷⁵ a causal inference might be wrong because we have determined the direction of the causal action incorrectly. We may have mistaken the effect for the cause.¹⁷⁶

A correlation between two variables might also be causally spurious because it is the result of a third variable acting on each of two correlated variables, rather than causal action between those two variables.¹⁷⁷ For example, in Figure 3 above, a high *bivariate* correlation between number of publications and salary might be largely spurious, because time since Ph.D. might have a substantial effect upon both publications and salary.¹⁷⁸ In such a case, the observed correlation is actually brought about by the third variable (or by multiple other variables). The correlation is causally spurious to the extent that it does not manifest a causal relationship between the correlated variables.

We can also incur causal error by concluding incorrectly that no causal relationship exists between two variables because we observe no correlation between them. Even when we see no statistical correlation between two variables, or only a weak correlation, the two variables might still be causally related. This situation can result when a *suppressor variable* reduces or eliminates the observable effect of the causal variable, thus masking its causal action.¹⁷⁹ If two variables work in opposite directions on a third variable, one tending to increase the third variable's value and the other tending to decrease it, then the observed value of the affected variable might show little or no correla-

173. J. DAVIS, *supra* note 163, at 25; D. KENNY, *supra* note 163, at 4.

174. Of course, an observed correlation might be merely a sampling effect. I assume here, for sake of analysis, that sampling uncertainty is not an issue. Similarly, it is assumed in this section that there is no modeling uncertainty.

175. See *supra* note 144.

176. See J. DAVIS, *supra* note 163, at 25.

177. See *id.* at 25-27.

178. J. COHEN & P. COHEN, *supra* note 78, at 359. For an illustrative example of investigating alternative hypotheses to determine whether observed correlations are spurious, see Graham & Garber, *supra* note 132.

179. J. DAVIS, *supra* note 163, at 33.

tion with either one or both of the causing variables. Yet a conclusion that no causal relationship exists between the variables would be causal error.¹⁸⁰

A third important source of causal error is *premature closure*: the decision not to consider more variables than those included in the analysis or model.¹⁸¹ We cannot simply assume that the inclusion of additional variables would not affect the correlations we are observing. An additional variable might show that some already observed correlation is causally spurious; the addition of a suppressor variable might change our interpretation of an otherwise weak relationship.¹⁸² In a multiple regression model, the values of the partial correlation coefficients are always relative to the set of independent variables actually included in the model, with the result that those values might change whenever new independent variables are added.¹⁸³ Thus, whenever we close the set of variables that are under consideration as causal factors, it is possible that we have left out some causally relevant variable.¹⁸⁴ Our principal justification for considering only certain variables and not others must always be theoretical—that is, based upon some theory about the

180. Cf. *id.* at 57-59 (path analysis works equally well with positive or negative coefficients).

181. See *id.* at 35, 65-66 (absent variable "might do anything" if added to the model).

Too often researchers examine the simple, or raw, correlation coefficient as an indication of causal effects. The naive logic is that if X causes Y, then X and Y should be correlated, and if X does not cause Y, they should be uncorrelated. Neither statement is true. After controlling for other exogenous [i.e., causal or independent] variables, a strong relationship can vanish and a zero relationship can become strong.

D. KENNY, *supra* note 163, at 62.

The results of deleting relevant variables are potentially serious to the model. See, e.g., W. BERRY & S. FELDMAN, *supra* note 130, at 20-25.

182. See J. DAVIS, *supra* note 163, at 66.

183. See *supra* note 181.

184. In multiple regression models, error due to premature closure is referred to as "*specification error*." See W. BERRY & S. FELDMAN, *supra* note 130, at 20-25; Fisher, *supra* note 7, at 708-09; Graham & Garber, *supra* note 132, at 211-12. See also *Bazemore v. Friday*, 478 U.S. 385, 397-404 (1986) (while omission of variables from a regression analysis may render analysis less probative, analysis which accounts for "major factors" normally admissible in Title VII pattern and practice case).

Another kind of specification error is putting irrelevant variables into the multiple regression model, see W. BERRY & S. FELDMAN, *supra* note 130, at 18-20; this second kind of specification error can be a source of causal error as well. While the predictive error that results might be minimal, the variable's presence can be more misleading in a causal sense if, through sampling error, we happen to draw a sample in which the causally irrelevant variable generates a statistically significant coefficient and we therefore conclude that we have a causally relevant variable. This kind of specification error can thus combine with sampling error to produce a causally spurious correlation within the model.

causal structure of the system under study.¹⁸⁵

B. *Techniques for Reducing Causal Uncertainty*

Scientists routinely use methods that assist them in reducing the risk of causal error. One of the most traditional techniques is to *control* the values of causally relevant variables that are not under study.¹⁸⁶ The most basic kind of control consists of keeping the values of the controlled variables constant while the effects of some other variable are being studied. This kind of control can sometimes be accomplished in a laboratory, where such variables as air temperature or diet can be manipulated. Another approach, especially in a laboratory setting, is to employ a "control group": a group of animals, for example, that are as nearly identical to the test animals as possible and are subjected to all the same conditions as the test animals with the exception of the intervention or treatment whose effects are being studied.¹⁸⁷ In principle, if a perfect control group were achievable, any changes observed in the test group that did not occur in the control group could be safely attributed to the difference between the groups—namely, to the variable under study.¹⁸⁸

When a variable or group cannot be manipulated directly, however, *statistical control* of a variable can be achieved by monitoring the variable throughout the study and then taking it into account within the mathematical model used. In multiple regression modeling, independent variables can be regarded as statistical controls relative to the other independent variables.¹⁸⁹ Various statistical techniques are used to take control data into account, such as standardization of data¹⁹⁰ or replication within control categories.¹⁹¹ Regardless of the technique,

185. In a particular case, the only way to test for misspecification might be to evaluate the predictive model by testing alternative models, with more or fewer variables. See W. BERRY & S. FELDMAN, *supra* note 130, at 25-26.

186. See, e.g., J.S. MILL, *supra* note 166, Book III, ch. VIII, §§ 2-3, at 454-60; P. SPECTOR, *RESEARCH DESIGNS* 15-16 (1981).

187. For a discussion of the benefits of a control group, see D. CAMPBELL & J. STANLEY, *EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR RESEARCH* 13-16 (1963).

188. Error can be introduced in this experimental design not only by control failure (which could lead directly to causal error), but also by errors in measurement, sampling, or modeling. The present discussion is focused only on causal error.

189. See Fisher, *supra* note 7, at 708-15; *supra* text accompanying notes 154-55.

190. See, e.g., M. BLAND, *supra* note 106, at 299-302 (giving example of direct and indirect methods of age standardization of mortality rates to eliminate effects of different age structures in populations to be compared).

191. E.g., J. DAVIS, *supra* note 163, at 36 (repeating analyses or duplicating experiments separately among men and among women to eliminate effects due to sex).

however, the objective is the same: to avoid causally spurious correlations, or to quantify the causally spurious portions of correlations between variables so that what remains has a greater chance of having causal significance.¹⁹²

A second traditional technique, *randomization*, has an objective similar to that of control: it is intended to eliminate causally spurious correlations.¹⁹³ When we randomly sort cases or subjects into two or more groups, we are trying to eliminate all but chance differences between the groups.¹⁹⁴ In randomizing the cases, we are trying to eliminate all causal influences on the sorting itself, and thus eliminate all causally significant differences between the groups. After true randomization, the only differences that should be observable between the groups are those arising by chance in the sorting process and those due to differences to which the groups are subjected after sorting. Therefore, control and randomization are powerfully combined in a laboratory study design in which subjects are randomly assigned to the control and test groups.

The concept of randomization is already familiar from the notion of a simple random sample.¹⁹⁵ When a true random sample is drawn from a population, the members of the population are randomly divided into two groups: the sample group and the rest of the population. To the extent that we achieve a truly random sample, there should be no correlations within the sample that differ from those in the population except for those that have arisen by chance in the process of sam-

192. See J. DAVIS, *supra* note 163, at 36. Multiple regression models are used to achieve this result by generating *partial* regression coefficients for the independent variables. See *supra* text accompanying notes 154-55. A partial regression coefficient characterizes the change in the predicted value of the dependent variable as a function of an independent variable, but only after taking into account the functional effects on the dependent variable of all the other independent variables in the model.

193. See, e.g., J. COHEN & P. COHEN, *supra* note 78, at 13; J. DAVIS, *supra* note 163, at 34-35; D. KENNY, *supra* note 163, at 184-87.

194. See, e.g., M. BLAND, *supra* note 106, at 8-13, 22-23 (randomization is only generally satisfactory method of allocating subjects to treatment groups so that characteristics of subjects do not affect chance of being put into any particular group); D. CAMPBELL & J. STANLEY, *supra* note 187, at 6 (random assignment to separate treatment groups is "all-purpose procedure for achieving pretreatment equality of groups, within known statistical limits"); Black & Lilienfeld, *supra* note 10, at 756 (in epidemiologic studies known as "clinical trials," purpose of randomized assignment of individuals to exposed and nonexposed groups is to "ensure" that only difference between groups is in the exposure); Lavori, Louis, Bailar & Polansky, *Designs for Experiments—Parallel Comparisons of Treatment*, in MEDICAL USES OF STATISTICS, *supra* note 106, at 41, 43-50 (discussing benefits of randomization in clinical trials of medical treatments).

195. See *supra* text accompanying note 96.

pling.¹⁹⁶ Random sampling highlights the additional power of randomization: if we are able to randomize our cases, we do not need to control (or even know the identity of) those variables being prevented from producing spurious correlations.¹⁹⁷ If we can randomize, we hope to sever *all* causally significant differences between the groups and thus eliminate the potential for an observed correlation to be causally spurious except by chance.¹⁹⁸

While these traditional techniques help to reduce the potential for causal error, the problem of premature closure guarantees that causal conclusions are never free of causal uncertainty.¹⁹⁹ In the end, despite the use of helpful statistical techniques and favorable research designs, scientists must rely upon adequate scientific theory and comprehensive empirical study to minimize causal uncertainty.

The task of characterizing residual causal uncertainty has not been addressed systematically and comprehensively. Scientists can help decisionmakers to appreciate the extent of residual causal uncertainty by explaining the theoretical weaknesses behind key causal conclusions and by reporting on empirical investigations that have led scientists to conclude that additional variables are causally irrelevant. In addition, regulatory agencies sometimes attempt to develop a scale or index for evaluating the strength of evidence for causal conclusions.²⁰⁰ It seems,

196. M. BLAND, *supra* note 106, at 8-14, 33-34.

197. J. DAVIS, *supra* note 163, at 35.

198. Besides the traditional techniques of control and randomization, various other techniques have been developed that can help to reduce causal uncertainty. Such techniques include elaboration and effects analysis. For a discussion of these techniques, see J. DAVIS, *supra* note 163, at 39-48. Path analysis is also a powerful means of identifying potential causal error. See *supra* text accompanying notes 169-70. While the first two techniques are applicable to any measure of statistical association (not just correlation), path analysis is limited to such specific statistical techniques as multiple regression and weighted regression with proportion or percentage differences. J. DAVIS, *supra* note 163, at 48, 59.

199. Conceptual uncertainty is also a significant factor, for to the extent that scientists are able to reconceptualize phenomena and redefine their variables, see *supra* text accompanying notes 47-49, they have the potential to reconceptualize causal theories as well. If scientists are dissatisfied with the extent of residual causal uncertainty, they may be led to redesign the concepts and variables employed in their theories. Thus, although conceptual uncertainty is logically distinct from causal uncertainty, there is a logical relation, and sometimes an important dynamic interplay, between these two kinds of uncertainty. Cf. Finkelstein, *supra* note 7, at 1446 (discussing example of determining cost of equity by means of a regression model, and the need for an underlying theory to support the argument that variability of earnings is an appropriate measure of risk and that risk is a relevant factor for adjusting the cost of equity).

200. For example, the EPA has developed a classification system for categorizing the weight of evidence for an environmental agent's capability of causing cancer in humans. See Guidelines for Carcinogen Risk Assessment, 51 Fed. Reg. 33,992, 33,999-34,000 (EPA 1986).

however, that no systematic, comprehensive, and standard scheme has been developed that can be used to help the decisionmaker appreciate the true potential for causal error in the information being relied upon.

VII. EPISTEMIC UNCERTAINTY

I have discussed how descriptive scientific information has associated with it the potential for error in the selection of concepts and variables, in measurement, sampling, modeling, and the drawing of causal conclusions. Such uncertainties attach to the truth of even the simplest causal assertions made by scientists. The final kind of uncertainty that I will discuss, which I call "epistemic uncertainty," is in some ways even more fundamental than these five kinds of uncertainty, because it involves the potential for error in our thinking about the nature or structure of empirical knowledge itself. Epistemic uncertainty can arise from our selection of theories in such fundamental areas as deductive logic²⁰¹ or basic mathematics. Such error in our epistemological theories can infect all of our descriptive information about the world.

An extended example should help to clarify what I mean by epistemic uncertainty. The example is the controversy surrounding the meaning of probability assertions about events in the world: assertions that some event is "probable" or that the probability of some event's happening is $1/x$. Such assertions play critical roles in social decision-making, whether the task is tort adjudication²⁰² or administrative rulemaking.²⁰³ Scientists are divided, however, on the meaning of such

201. For example, such uncertainties include whether the deductive logic used should model valid inference on a propositional level (propositional logic) or predication level (predicate or quantified logic), or whether modal operators (representing, for example, logical necessity or logical impossibility) should be available within the logic. See generally I. COPI, *SUPRA* note 30; G. HUGHES & M. CRESSWELL, *AN INTRODUCTION TO MODAL LOGIC* (1968). Today, such decisions are reaching the practical level of decisionmaking through the construction of "expert systems," for which a deductive logic must be chosen as part of the "inference engine." See R. SUSSKIND, *EXPERT SYSTEMS IN LAW* 10 (1987) ("inference engine" is the mechanism by which the knowledge base interacts and reasons with the data relating to any problem at hand). The choice of which deductive logic to use can in turn affect the kinds of inferences derivable within the expert system. See *id.* at 163-69 (inference engine of expert system necessarily implements some formal logical system, whether "traditional" syllogistic logic, "classical" propositional and predicate logics, deontic logic, modal logic, fuzzy logic, or many-valued logic).

202. The classic formula inserting probability into issues of negligence is the statement by Judge Learned Hand that due care is a function of the probability that resulting injury will occur. *United States v. Carroll Towing Co.*, 159 F.2d 169, 173 (2d Cir. 1947); see RESTATEMENT (SECOND) OF TORTS § 293 (b) & Comment b (1965) (in determining magnitude of risk for purposes of determining negligence, extent of chance that conduct will cause harm is important factor).

203. See, e.g., Guidelines for Carcinogen Risk Assessment, 51 Fed. Reg. 33,992 (1986) (ad-

statements, and, as will be seen, the choice of which theory of meaning to adopt can have significant implications for decisionmakers.

The classical interpretation of a probability assertion is that it is a statement about the relative frequency of occurrence of some type of event in the long run.²⁰⁴ When we say, for example, that the probability of throwing a "2" with a fair die is 1/6, we mean (according to the classical theory) that in the long run, the most likely frequency for throwing a "2" is 1 time per 6 throws. This classical interpretation of probability statements was used explicitly in the discussion of sampling uncertainty above,²⁰⁵ in the context of both significance testing and confidence interval construction. In classical significance testing, we use probability theory to derive the probability of drawing a particular type of sample from the (hypothetical) population. When we construct confidence intervals, we use probability theory to determine the size of the interval estimate that has the desired probability of including the true population value. In either case, a statement of probability is critical to the reasoning.

When we explain significance testing and confidence intervals using the classical interpretation of probability, we say that, in a probability distribution of sample proportions, the sample proportion with the highest probability is the sample we expect to be the most frequently drawn in the long run.²⁰⁶ Under this classical interpretation, the probability being asserted is really the expected frequency of drawing a particular type of *sample*, but the probability statement is not about the *identity* of the true population proportion (the parameter being estimated).²⁰⁷ The true proportion is unknown, but it is some definite number. The probability statement is not about the likelihood that the population has a given proportion, but rather about the likelihood of drawing possible samples from a population.²⁰⁸

dressing generically the importance of such issues as determining how likely it is that an environmental agent is a human carcinogen, estimating the likely range of excess cancer risk associated with given levels of exposure, and estimating the exposures to which populations are likely to be subject); see also J. COHRSEN & V. COVELLO, *supra* note 27, at 90-94 (discussing use of probability distributions, confidence intervals, and other techniques to characterize uncertainty in such variables as failure rates for equipment or census data on exposed populations, or in the use of models for estimating dose-response or groundwater contamination).

204. See, e.g., G. IVERSEN, *BAYESIAN STATISTICAL INFERENCE* 8 (1984); *supra* notes 93, 97.

205. See *supra* text accompanying notes 93-127.

206. The long-run, relative frequencies for other possible samples of the same size are distributed according to the binomial distribution. See *supra* note 99.

207. E.g., *supra* note 113; see also G. IVERSEN, *supra* note 204, at 76.

208. See G. IVERSEN, *supra* note 204, at 76.

Many scientists, however, believe that the classical interpretation of a probability statement is of limited usefulness. First, they contend that many ordinary probability assertions are not about the relative frequency of objectively observable events. Rather, the uncertainty expressed in our probability statements is a lack of confidence in the truth of our beliefs, and probability statements are really measures of subjective uncertainty.²⁰⁹ When we say that it will *probably* rain tomorrow, we are conveying a measure of our confidence in the proposition that it will rain tomorrow. We are not ordinarily asserting that more often than not rain occurs (or would occur) on the day following meteorological circumstances sufficiently similar to those of today. The latter classical interpretation of our probability statement, therefore, does violence to our ordinary meaning.

It is also argued that the relative frequency interpretation of probability cannot deal adequately with unique events. Probability statements should be meaningful when applied to unique events, without our having to conceptualize the unique event as somehow recurring "in the long run."²¹⁰ The fact that lawsuits typically deal with unique events may explain why probability theory (in the classical sense) has had such limited usefulness in the law.²¹¹ The classical interpretation of probability is counterintuitive to many lawyers because it seems to require us to assess the relative frequency of outcomes in the complicated circumstances of even the most bizarre accident.²¹²

A further argument against the classical interpretation is made in the context of statistical sampling theory. What we know with a great deal of confidence is the sample statistic (the data); what we are uncertain about is the population parameter, about which we are trying to reach a justified belief based upon the sample. This leads to a result precisely opposite to that of the classical viewpoint: probabilities should be computed for the value of the *parameter*, not for the sample. The sample itself has no probability attached to it.²¹³

209. *See id.* at 9, 17, 66. Arbitrary and subjective elements are present within classical theory as well, such as the choice of the level of significance for rejecting an hypothesis or the choice of a level of confidence. *See id.* at 66-67; S. PRESS, *BAYESIAN STATISTICS: PRINCIPLES, MODELS, AND APPLICATIONS* 45-46 (1989).

210. G. IVERSEN, *supra* note 204, at 9.

211. *See Confidence in Probability*, *supra* note 13, at 386-87, 391.

212. It seems counterintuitive that we decide that the accident of the famous Mrs. Palsgraf, *Palsgraf v. Long Island Railroad Co.*, 248 N.Y. 339, 162 N.E. 99 (1928), was improbable by conducting a thought experiment and tallying the number of possible scenarios containing her injury.

213. G. IVERSEN, *supra* note 204, at 17, 76. For the classical interpretation, see *supra* note

Finally, it is argued that classical statistical inference, based on the classical interpretation of probability, is inadequate to the task of measuring the change in our confidence in a conclusion as a result of taking new information into account.²¹⁴ Scientists engage in ongoing discovery and continually revise their tentative conclusions (for example, about the risks associated with exposure to alternating magnetic fields) on the basis of new data. Therefore, an adequate theory of statistical inference should be able to accommodate our initial ("prior") assessments of probability and should provide a means of adjusting that probability assessment as empirical data accumulate.²¹⁵ A subjective interpretation of probability leads us to want to develop such a cumulative rule for probability. Classical significance testing, in contrast, tends to treat each scientific study independently of prior studies and does not normally take prior probabilities into account when subsequently acquired data are tested for significance.

For these reasons, a substantial number of scientists favor using an alternative theory of statistical inference that is based upon a subjective interpretation of probability.²¹⁶ The leading alternative to the frequency-based classical theory is the theory of statistical inference based upon Bayes' Theorem.²¹⁷ Bayes' Theorem provides a rule for combining

113.

214. Another argument made against the classical interpretation is that a subjective interpretation of probability is better able to avoid certain paradoxes of classical hypothesis testing or confidence interval generation, such as the paradox of "stopping rules." It is argued that under classical theory the statistical significance of a sampling result should be a function of the researcher's *intention* in deciding how to draw the sample. See G. IVERSEN, *supra* note 204, at 73-74. This argument is technical in nature and does not add weight to the premise of this section, which is simply that there are alternatives to the classical view of the meaning of probability that have substantial adherence in the scientific community.

215. *Id.* at 11, 16. Legal commentators have also emphasized the need to determine how to combine "hard" scientific information (such as the statistical results of a carefully conducted study) with "soft" information (such as eyewitness testimony). See, e.g., Koehler & Shapiro, *supra* note 11, at 265-78; *Laws of Probability*, *supra* note 11; *Gatecrasher Paradox*, *supra* note 11, at 106-08; *Comment*, *supra* note 11; *Trial by Mathematics*, *supra* note 11; *Identification Evidence*, *supra* note 11; Kaplan, *supra* note 12, at 1083-91. Cf. Finkelstein, *supra* note 7, at 1455-75 (suggesting protocols for the use of regression models about economic effects in administrative proceedings, with the objective of moving subjective administrative considerations "back from the ultimate conclusion into preliminary questions involved in constructing a model").

216. For legal commentators arguing the merits of subjective versus frequentist interpretations of probability statements, see Brilmayer & Kornhauser, *supra* note 13, at 137-48; *Confidence in Probability*, *supra* note 13, at 390-93; *Identification Evidence*, *supra* note 11, at 504-05; Kaplan, *supra* note 12, at 1066-67; *Paradoxes*, *supra* note 13, at 641-45; *Gatecrasher Paradox*, *supra* note 11, at 104-06; Koehler & Shaviro, *supra* note 11, at 252-53.

217. Bayes' Theorem is a fundamental theorem of the mathematical theory of probability. See, e.g., *Laws of Probability*, *supra* note 11, at 41-53; *Paradoxes*, *supra* note 13, at 637-38

a "prior" probability distribution for a parameter with new data to generate a "posterior" probability distribution, which in turn could become the new prior probability distribution for subsequent studies. Bayes' Theorem, therefore, is used to take into account old estimates of a population value plus new information, in order to calculate revised estimates for that population value.²¹⁸

As an example, consider the estimation of a proportion p within a population. A Bayesian approach requires generating a prior probability distribution for the possible values of p .²¹⁹ This probability

(Bayes' Theorem does not presuppose subjective interpretation of probability, and holds for other interpretations such as relative frequency or degree of confirmation). The theorem is often formulated as follows:

$$pr(X|E) = \left[\frac{pr(E|X)}{pr(E)} \right] pr(X) ,$$

where $pr(X)$ is the prior probability that some proposition X is true, $pr(E|X)$ is the conditional probability that some evidentiary proposition E is true given that X is true, $pr(E)$ is the probability that E is true regardless of whether X is true or not, and $pr(X|E)$ is the conditional probability that X is true given the truth of E . The unconditional probability that the evidence is true, $pr(E)$, can be expanded in terms of conditional probabilities:

$$pr(E) = [pr(E|X) pr(X)] + [pr(E|non-X) pr(non-X)] .$$

For derivations of the theorem, see *Res Ipsa Loquitur*, *supra* note 13, at 1468-71; *Trial by Mathematics*, *supra* note 11, at 1351-53; *Identification Evidence*, *supra* note 11, at 498-99.

218. In recent years, there has been a vigorous debate among legal commentators over the usefulness of either Bayes' Theorem or Bayesian statistics in legal decisionmaking. See *Laws of Probability*, *supra* note 11; Kaplan, *supra* note 12. Those arguing for the usefulness of Bayesian techniques include: Finkelstein & Fairley (in *Comment*, *supra* note 11, and *Identification Evidence*, *supra* note 11); Kaye (in *Res Ipsa Loquitur*, *supra* note 13); Koehler & Shaviro (in *Verdictal Verdicts*, *supra* note 11); Kornstein (in *A Bayesian Model of Harmless Error*, *supra* note 13); and Lempert (in *Modeling Relevance*, *supra* note 13).

Those arguing against at least certain uses of Bayesian techniques in a legal context include: Brilmayer & Kornhauser (in *Review: Quantitative Methods and Legal Decisions*, *supra* note 13, at 130-48); Callen (in *Notes on a Grand Illusion*, *supra* note 11); L. Cohen (in *Subjective Probability and Logic of Proof*, *supra* note 13); and Tribe (in *Mathematical Proof and Trial by Mathematics*, *supra* note 11).

219. For use in the context of statistical inference, Bayes' Theorem could be used to relate the probability that the population proportion for some variable has the value p to the fact that the sample data has a proportion P , as follows:

$$pr(p|dataP) = \frac{pr(dataP|p) pr(p)}{[pr(dataP|p) pr(p)] + [pr(dataP|non-p) pr(non-p)]} ,$$

where $pr(p)$ is the "prior" probability that the proportion in the population (the parameter) has the value p (a probability derived prior to taking the subsequently acquired data with observed proportion P into account); "dataP" is the proposition that the data has proportion P ; $pr(dataP|p)$ is the probability of drawing a sample with the observed proportion P given a population proportion p ; $non-p$ is the complement to p — that is, p and $non-p$ are mutually exclusive and exhaustive hypotheses about the population proportion; and $pr(p|dataP)$ is the probability of the population

distribution for p might be based in a particular case upon empirically rigorous data (such as the results of a large simple random sample), less rigorous empirical data, anecdotal information, or theoretical considerations.²²⁰ The probability distribution could be generated on the basis of opinions of experts familiar with the relevant scientific field.²²¹ Whatever the appropriate method for generating such probability distributions (a matter of much dispute among even Bayesian theorists),²²² the objective is to model, using a probability distribution, our best estimates of the value of p prior to our taking the new data into account. Once the prior probability distribution has been constructed, Bayes' Theorem can be used to derive a posterior probability distribution on the basis of the prior distribution and the new data. Thus, the use of prior probabilities allows us to take prior subjective knowledge (from whatever source) into account, and the use of Bayesian methods of calculation provides a means of combining that "captured" subjective knowledge with a probability assessment based on objective, quantitative data from empirical studies.²²³

Bayesian techniques can yield different results in significance testing and different confidence intervals than those generated with classical techniques. The introduction of prior probabilities into the calculation can cause the numerical limits of Bayesian confidence intervals, for example, to diverge from classically generated ones; the degree of divergence depends upon the extent to which the prior distribution affects the outcome.²²⁴ In situations in which the prior probability has little effect, the confidence intervals generated by Bayesian methods can be numerically identical to the confidence intervals calculated using classi-

proportion's being p if the observed sample proportion is P . See G. IVERSEN, *supra* note 204, at 12-16; *supra* note 217. This expression of Bayes' Theorem illustrates its direct application to the problem of significance testing.

The appropriate use of Bayesian techniques in statistical inference is a matter of controversy even among scientists. See, e.g., Brilmayer & Kornhauser, *supra* note 13, at 135 n.68; *Laws of Probability*, *supra* note 11, at 51 n.57.

220. See G. IVERSEN, *supra* note 204, at 62.

221. See, e.g., *id.* at 64-68; H. RAIFFA, *DECISION ANALYSIS: INTRODUCTORY LECTURES ON CHOICES UNDER UNCERTAINTY* 104-28 (1968).

222. Discussions concern such issues as how best to model complete ignorance: whether, for example, such states of "knowledge" should be modeled by using equal probabilities over the range of possible values. See G. IVERSEN, *supra* note 204, at 61-62.

223. See *supra* note 215. For a detailed illustration of Bayesian techniques in combining blood test data with other kinds of evidence to decide paternity cases, see Berry & Geisser, *supra* note 87.

224. G. IVERSEN, *supra* note 204, at 30-31, 68-70.

cal methods.²²⁵ The use, therefore, of Bayesian statistics as an alternative to classical statistics can have importance at the level of practical decisionmaking to the extent that the decision rests (for example) upon the size of confidence intervals.

This example of confidence interval size also illustrates the fundamental and pervasive nature of epistemic uncertainty. The choice of either classical or Bayesian theories of statistical inference can affect the size of confidence intervals, which are useful measures of sampling uncertainty. The selection of one theory of statistical inference or interpretation of probability over the other, therefore, is itself a source of epistemic uncertainty. The extent of epistemic uncertainty, as opposed to sampling uncertainty, might be gauged in a suitable case by constructing different confidence intervals using the competing interpretations of probability and statistical inference, and by noting the difference in the derived confidence intervals.²²⁶

This brief discussion of Bayesian inferential statistics is intended merely as an illustration of epistemic uncertainty. Because the nature of epistemic uncertainty is likely to vary depending upon the source of the uncertainty, both the means for reducing epistemic uncertainty and the techniques for characterizing the extent of it also vary greatly. In the case of classical versus Bayesian inferential statistics, for example, the numerical difference in confidence intervals might be taken as a rough measure of the epistemic uncertainty introduced by the choice of one method over the other. Other instances of epistemic uncertainty might not lend themselves to such neat resolution.

225. For example, in analyzing the mean from a normal distribution, if the new data are from a large sample, then the prior probability distribution typically has little or no effect on the posterior probability distribution. *Id.* at 69.

226. A Bayesian theorist who bases prior probabilities upon purely subjective or intuitive bases, as opposed to empirically generated data, may introduce additional types of uncertainty beyond those discussed in this Article. The nature of such uncertainties depends upon the nature of the information employed and upon the extent to which the taxonomy discussed here can be generalized beyond purely scientific information. I suspect, although I do not argue here, that this taxonomy also captures most kinds of uncertainty present in any descriptive information, whether "scientific" or not.

Perhaps one reason that non-Bayesians resist Bayesian analyses as "unscientific," *see, e.g.*, S. PRESS, *supra* note 209, at 45, is precisely because such analyses seem to have the potential for contaminating scientific information (with its at least *identifiable* types of uncertainty) with sources of potential error that are elusive, ill-defined, and resistant to meaningful measurement, evaluation, or reduction.

VIII. CONCLUSION

This Article presents a taxonomic scheme for cataloging the distinct kinds of uncertainty associated with descriptive scientific information.²²⁷ I propose that these six kinds of uncertainty capture those principal sources of potential error in descriptive scientific information that pose conceptual challenges to decisionmakers.²²⁸ I know of no way to *demonstrate* that all important sources of error have been included, and I put forward this taxonomy as a first approximation of a complete understanding of the structure of scientific uncertainty.

This taxonomy, however, even if it proves to be incomplete, should assist decisionmakers in understanding many important ways in which descriptive scientific information can "go wrong." It should provide a

227. The six kinds of uncertainty discussed in the previous sections of this Article are, it is hoped, the major kinds of potential error associated with descriptive scientific information. In addition to describing the world, however, some scientists have addressed decisionmaking itself and have developed theories about how to make better decisions. Such theories include risk/benefit analysis, normative economic analysis, decision analysis, and game theory. See generally R. LUCE & H. RAIFFA, *GAMES AND DECISIONS: INTRODUCTION AND CRITICAL SURVEY* (1957); H. RAIFFA, *supra* note 221; Schotter & Schwödiauer, *Economics and the Theory of Games: A Survey*, 18 J. ECON. LITERATURE 479 (1980). A more recent area of growing regulatory importance is "risk management," a part of the process of "risk assessment." See, e.g., J. COHRSEN & V. COVELLO, *supra* note 27, at 8; NATIONAL RESEARCH COUNCIL, *supra* note 43, at 3, 17-50. It remains to be seen whether this latter area of study will produce any distinctive scientific theories about decisionmaking.

What such theories have in common is that their conclusions are often cast in the language of prescription, with normative overtones. See, e.g., *Gatecrasher Paradox*, *supra* note 11, at 106 (use of Bayes' Theorem as normative). Such theories arrive at reasoned conclusions about what we *ought* to do, *given our residual descriptive uncertainties*. They lead to conclusions about what factors ought to be taken into account in decisionmaking, what conditions or endstates ought to be preferred, and what actions ought to be taken. They produce advice about how best to conduct our social affairs in the face of our considerable uncertainty about how things work.

This Article does not analyze prescriptive scientific conclusions or the kinds of unique uncertainty that might be associated with such conclusions. One example of a source of prescriptive uncertainty might be the assignment of utilities, preferences, or losses to various outcomes. See, e.g., H. RAIFFA, *supra* note 221, at 51-103. Legal commentators have also been concerned with such problems. See, e.g., *Laws of Probability*, *supra* note 11, at 53-56; *Trial by Mathematics*, *supra* note 11, at 1378-93; Kaplan, *supra* note 12. If such scientific theories about decisionmaking reduce to descriptive theories about the conditional consequences of human action, then the uncertainties associated with scientific prescriptions should reduce to one or more of the categories of descriptive uncertainty discussed in this Article. I do not in this Article attempt such a reduction or speculate about the outcome of such an attempt.

228. I have noted in passing the additional potential for error from mistaken computation, human error, and fraud, *supra* note 20. In addition, with respect to Bayesian inferential statistics, there seems to be an unexplored possibility for infusing error of a purely subjective sort, *supra* note 226. In fact, any method of generating information or knowledge that is subjective in a radical sense could have such error associated with it, and such error would not be peculiar to science.

useful framework for analyzing the kinds of uncertainty inherent in scientific information. Those responsible for adjudicating guilt or liability, or for making regulatory decisions that affect public health or safety or the environment, need to appreciate precisely how the scientific information supporting their decisions might be in error. Armed with such an understanding, decisionmakers should be better able to identify critical questions for their scientific witnesses and advisors, reduce critical uncertainties to an acceptable level, appreciate the residual risks associated with decisions, and properly allocate to the parties the burden of producing evidence.²²⁹ As the planet becomes more populated, human demands upon dwindling resources grow, and the relationships between decisions and their effects become more complex, it is increasingly important to make decisions based upon the least amount of scientific uncertainty that the society as a whole can afford.

229. As Judge Bazelon noted: "Courts frequently grant the agency wide discretion in allocating the burden of uncertainty. This circumstance makes full disclosure of uncertainties as important as full disclosure of known risks." Bazelon, *supra* note 14, at 213. For an interesting step in the direction of allocating burdens of proof based on sampling uncertainty, see *Confidence in Probability*, *supra* note 13. There are more kinds of scientific uncertainty, however, than sampling uncertainty.