Prepublication draft of Jan. 22, 2017, for *Michigan Law Review*, Vol. 115

# Digging Into the Foundations of Evidence Law

*David H. Kaye*[*]

**The Psychological Foundations of Evidence Law**. By Michael J. Saks and Barbara A. Spellman. New York and London: New York University Press. 2016. Pp xiv, 241. Cloth, $89; paper, $38.

## Introduction

Professors Michael Saks[1] and Barbara Spellman[2] have produced a gem of a book. A concise, cogent, and thoughtful introduction to the major rules of evidence, *The Psychological Foundations of Evidence Law* glitters in the light of ideas from social and cognitive psychology. *PFEL*, as I will abbreviate it, is an eminently accessible[3] book that evidence professors should assign to their students; that psychologists seeking research questions about evidence law should consult;[4] that litigators seeking to sharpen their persuasive powers should peruse; and that judges engaged in the "metacognitive" task of applying rules for screening improperly prejudicial evidence from jurors should examine. In Saks and Spellman's words, the book

> explores a number of important practices from evidence law about which psychology does, or could, have a lot to say to illuminate the underlying assumptions, and evaluates whether those assumptions are consistent with the psychological research or whether the law's goals for evidence doctrine could be achieved more successfully with a modified rule or a different rule or no rule at all. (p. 3)

This Review has three goals. Part I surveys *PFEL*'s scope and provides examples of several of its conclusions. Part II focuses on one psychological (or logical) model that the book

[*] Associate Dean for Research, Distinguished Professor of Law and Weiss Family Scholar, Penn State Law. Regents Professor Emeritus, ASU Sandra Day O'Connor College of Law. Jay Koehler and Michael Saks provided comments on a draft of this essay.

[1] Regent's Professor of Law and Psychology, Arizona State University Sandra Day O'Connor College of Law.

[2] Professor of Law, University of Virginia School of Law.

[3] It stands as a counterexample to the generalization that, in psychology, "even outstanding authorities have been known to run in circles 'describing things which everyone knows in language which no one understands.' " Raymond B. Cattell, The Scientific Analysis of Personality 18 (1965).

[4] *Cf.* Michael J. Saks, Editorial, *The Law Does Not Live by Eyewitness Testimony Alone*, 10 Law & Hum. Behav. 279 (1986) (encouraging behavioral psychologists to expand their research beyond eyewitness testimony and into other areas of evidence law).

presents for understanding two fundamental concepts in evidence: relevance and probative value. It digs into the foundations of these concepts more deeply to exhibit a slightly different conception of the probative value, or, more metaphorically, the weight of evidence.

## I. **Exposing the Foundations**

At a granular level, the sheer number of rules of evidence is vast,[5] but *PFEL* surveys the major federal rules, with emphasis on probative value, counterweights to relevant evidence, character evidence, competency of witnesses, privileges, impeachment, hearsay, limited admissibility,[6] and expert and scientific evidence.[7] Throughout these areas, *PFEL* maintains that psychology can improve on rulemakers' "assumptions, impressions, anecdotes, and reasoning about human behavior" in two ways (p. 16)—by conducting controlled experiments (simulations with mock jurors and "similar studies to take place in the setting of actual trials")[8] and by

---

[5] *See* Steuart Henderson Britt, *The Rules of Evidence—An Empirical Study in Psychology and Law*, 25 Cornell L.Q. 556, 558 (1940) ("From the entire body of principles and definitions known as the 'rules of evidence,' 154 rules were selected for intensive statistical study. These 154 rules were . . . selected by careful analysis from the 3,150 rules in Wigmore's *Code of Evidence*."). *PFEL* reproduces most of the Federal Rules of Evidence to generate 30 pages (9%) of its text. Pp. 247–77.

[6] "Limited admissibility" refers to the doctrine that evidence may be admissible for one purpose but not for another. This situation arises frequently. Obvious examples are other-crimes evidence (when introduced for so-called "non-propensity" purposes); impeachment evidence (a prior inconsistent statement that is inadmissible hearsay if offered for its truth is nevertheless admissible to impeach the witness making the current statement as an indication that the witness says contradictory things and therefore should not be believed about anything); substantively inadmissible evidence introduced solely to show the basis for an expert's opinion; and evidence of subsequent remedial measures, compromise offers and negotiations, offers to pay medical and similar expenses, and liability insurance. *PFEL* ably surveys the efficacy of methods to help jurors reason as the law would like them to under the heading of "Instructions to Disregard and to Limit." P. 85. Strangely, it refers to only the last set of rules as "the categorical exclusion rules." P. 73. A rule such as Rule 404(a) on character evidence is no less a "categorical exclusion rule" that reflects an across-the-board balancing of probative value and prejudice. Likewise for the "*Frye* rule," which categorically excludes testimony about scientific test results when the test lacks general acceptance in the relevant scientific community. What makes *PFEL*'s subset of categorical rules distinctive is that they are not simply a crystallization of "the balancing of probativeness and prejudice that judges are called upon to perform during trials under Rule 403." P. 72. As Saks and Spellman quickly recognize, those exclusionary rules seek to promote what other writers have called "extrinsic policies." David P. Leonard, The New Wigmore, A Treatise on Evidence: Selected Rules of Limited Admissibility (rev. ed. 2002).

[7] *See* pp. 322–23.

[8] *See* pp. 16–17. The authors allude to "many different research designs," p. 17, and refer to some observational studies (such as a before-and-after study of a change in a rule and a crude comparison between civil and criminal cases), pp. 224–25, but there is less discussion of this mode of causal inference. For a more detailed but elementary exposition of the relative strengths and weaknesses of

"borrow[ing] from the findings of more basic empirical research, or from theoretical knowledge about human perception, memory, and information processing" (p. 17). In particular, *PFEL* brings to bear findings about mental contamination ("how prior knowledge can bias future judgments in an unwanted manner"),[9] dual systems reasoning (System 1's fast and frugal heuristics[10] versus System 2's "slower, conscious, and reflective" processing) (p. 20), information integration theory ("how people combine bits of information into a final judgment") (p. 43), motivated reasoning (p. 208), contextual bias (p. 26), attribution theory (p. 79), "weapons of influence" (pp. 40–41), and "diffusion of responsibility" (p. 216). It also includes specific studies on such topics as the impact of gruesome pictures (pp. 64–65), observer or expectancy effects in forensic science,[11] lie detection (pp. 122–28), decision aids for expert evidence (pp. 226–27), and framing effects in the presentation of statistical evidence (pp. 221–22).

On the basis of such theory and findings, Saks and Spellman give the modern trial process high marks in some respects. As they see it, "[t]he rules came into being to rein in the inevitable excesses of lawyers in an adversarial system" (p. 11). That is, Saks and Spellman embrace the theory of several legal historians that "limits had to be placed on what . . . lawyers would be permitted to do" (p. 11) in an adversarial system as opposed to an inquisitorial one in which these one-sided lawyers played a more ancillary role. But they believe that the adversarial system alone is not the full explanation. The presence of impressionable jurors also matters. "[P]ut the adversary process together with juries, and evidence rules become a necessary device for restraining lawyers, thereby protecting jurors from being deceived or misled" (p. 11). They then find that, as a whole, the rules of evidence perform this restraining function very well:

> [T]he rules that govern the trial have the effect of blocking or dampening the use of influence techniques that do not contribute to the (relatively) rational resolution

---

different methods, and for more examples of their use in studies of the legal system, see generally Hans Zeisel & David Kaye, Prove It with Figures: Empirical Methods in Law and Litigation (1997).

[9] P. 18. Saks and Spellman note that the phrase is not common. P. 18.

[10] P. 20 ("System I responds to information quickly, unconsciously, and by using heuristics."). The particular phrase "fast and frugal" comes from the writing of Gerd Gigerenzer and his colleagues. *E.g.*, Gerd Gigerenzer et al., *How Good Are Fast and Frugal Heuristics?*, *in* Heuristics and Biases: The Psychology of Intuitive Judgment 559, 559 (Thomas Gilovich et al. eds. 2002).

[11] Pp. 207–11 ("Once we desire to reach or hold onto a conclusion, that motivation activates a number of cognitive processes to push our reasoning in the direction of the desired conclusion. Our mind discounts contrary evidence and overweights confirmatory information, stretches or shrinks logical connections, widens or narrows categories, etc., all to help us reach the conclusion we want to be able to reach. The expert does not intentionally mislead.").

> of the disputed issues. In their effort to make trials more information-based and more rational, the rulemakers have done quite well by limiting the possibilities for using the most powerful tools of persuasion and influence. (p. 41)

Those tools, in the nomenclature of social psychologist Robert Cialdini, are the "six weapons"—reciprocation,[12] commitment and consistency,[13] authority,[14] social proof,[15] scarcity,[16] and liking[17]—and they are not fully functional for lawyers seeking to persuade jurors and judges.

*PFEL* is similarly laudatory in its assessments of the rules that channel and limit evidence of a person's character (or traits of character). In developing this doctrine (which continues to befuddle courts that must apply it), "rulemakers had latched onto some notions that are even more sound than they ever realized" (p. 29). "[T]he law's doubts about the usefulness of the concepts of character and personality are on the correct track psychologically. As to habit, from the perspective of psychological research and theory, the law is even more on target" (pp. 166–67).

*PFEL* also has much to say about jurors' abilities and performance on various tasks. The institution of multi-member juries fares well. "Twelve (or 8 or 6) heads are better than one" for

---

[12] P. 40 ("[P]eople more readily comply with requests (for favors, information, contributions, concessions, etc.) from those who have previously provided something to them. Attorneys cannot do favors for or give gifts to jurors, or treat them to lunch.").

[13] *Id*. ("[P]eople more readily change their opinion in a particular direction if they see it as consistent with an existing recent commitment. The challenge, therefore, is first to get people to make a small commitment in the desired direction; they will then be more willing to take the larger step that you want them to take. Attorneys certainly cannot obtain from jurors public behavioral commitments that are favorable to one side.").

[14] *Id*. ("[P]eople are more willing to follow the directions or recommendations of a communicator whom they perceive as having relevant authority or expertise. Jurors might believe that attorneys are authorities—but there are attorneys on both sides. Issues of following authority are more likely to come up with judges . . . or experts . . . .").

[15] Pp. 40–41 ("[P]eople are more willing to take a requested or suggested course of action if they see others, especially others similar to themselves, doing so. . . . [But] jurors typically are instructed to avoid learning what anyone else thinks about the case they are in the process of deciding. Lawyers themselves might be torn over whether to appear to jurors as authorities or as people "just like them.").

[16] P. 41 ("[P]eople find objects, activities, and information more attractive to the extent that those things are seen as scarce or declining in availability. A salesman might say, 'There are only two of this product left' or 'The sale lasts only until tomorrow.' How to apply this tactic in a trial is not apparent.").

[17] P. 41 ("[P]eople are more inclined to comply with the wishes of those they know and like than to strangers or people they know but dislike. Lawyers can't directly employ this weapon because jurors who are friends of one (or more) of the lawyers will be dismissed from the jury panel. But they might try in subtle ways to get jurors to like them.").

many reasons, "including the averaging of information, the aggregation of information, the suppression of individual jurors' biases, and the necessity of engaging in System 2 deliberative reasoning" (p. 46). And "a substantial body of empirical research has found that, for the great majority of cases, the characteristics of the jurors make only a modest difference to the verdict, whereas the evidence and arguments presented at trial have the greatest impact."[18]

Of course, jurors (and judges) have many limitations. Nobody (or at least no cross-section of people) is a very good lie detector.

> Despite decades of research effort to maximize the accuracy of deception judgments, detection rates rarely budge. Professionals' judgments, interactants' judgments, judgments of high-stakes lies, judgments of unsanctioned lies, judgments made by long-term acquaintances—all reveal detection rates within a few points of 50%. [Are we ready to] accept the conclusion implied by the first 384 research samples—that to people who must judge deception in real time with no special aids, many lies are undetectable[?][19]

Moreover, "[w]hen it comes to direct assertions of confidence (e.g., 'I'm sure') . . . the research shows that witness accuracy and witness confidence are generally not highly correlated. Jurors, however, place great stock in such assertions of confidence" (p. 133). Because "the actual value of the evidence and how jurors perceive the value of the evidence are far apart[, this] is[] an area begging for a rule" (p. 133).

In other areas, however, the authors find the research insufficient to refine or modify the folk psychology of jurists and legislators. For example,

> Many, but not all, of the studies find that the mock jurors give less weight to the same testimony when presented as hearsay than when presented by the witness with personal knowledge. . . . The studies vary on the hearsay exception that the information falls under, the instructions of the judge to the mock jurors, whether there was deliberation by the mock jurors, whether an expert testified on the problems of hearsay evidence, etc. With so many variables and so few studies, the

---

[18] P. 36. However, the authors wisely caution that "there are some important exceptions . . . . Individual factors are most likely to matter if cases are close, if they involve noncommon knowledge, or if personal attributes are an issue." P. 38. The first exception is related to the "liberation hypothesis" propounded in Harry Kalven & Hans Zeisel, The American J**ury** (1966), and studied extensively since then. *E.g.*, Dennis J. Devine et al., *Strength of Evidence, Extraevidentiary Influence, and the Liberation Hypothesis: Data from the Field*, 33 Law & Hum. Behav. 136 (2009); Amy Farrell & Daniel Givelber, *Liberation Reconsidered: Understanding Why Judges and Juries Disagree about Guilt*, 100 J. Crim. L. & Criminology 1549 (2010).

[19] P. 125 (first alteration in original) (quoting Charles F. Bond, Jr. & Bella M. DePaulo, *Accuracy of Deception Judgments*, 10 Personality & Soc. Psychol. Rev. 214, 231 (2006)).

> research is not rich or systematic enough to draw conclusions about specific hearsay exceptions. . . .
>
> . . . .
>
> . . . Psychology research could do a lot more . . . . (pp. 188, 201)

Moving outside the realm of assumptions and studies on jury performance, Saks and Spellman are also guarded in their conclusions. Thus, they write that

> The policy rationale [for Rule 407 on the admissibility of proof of subsequent remedial measures] is that the law wants to encourage defendants to make repairs . . . [but critics maintain that] citizens are unlikely to be aware of Rule 407, and if they are not aware of it, it is unlikely to affect their behavior. (p. 77)

Yet the authors "are aware of no empirical research testing whether the rulemakers or the critics have the stronger behavioral argument" (p. 78). Indeed, *PFEL* suggests that direct psychological or sociological studies are unlikely to provide an answer to some of the law's empirical questions. For example, they note that

> psychologists have not done much research on privileges, most likely because they are not amenable to experimental research . . . . [T]here is some survey research . . . . but it is doubtful that survey methodology could provide good reliable answers to most of the important questions about privileges. (pp. 140–41)

I could continue describing the authors' many assessments and insights, but I would not want to steal *all* the punchlines. Suffice it to say that, in *PFEL*, two of the world's leading scholars and teachers of law and psychology[20] distill a huge body of information and present their own opinions across the wide landscape of trial rules, procedures, and players.[21]

## II. **Excavating One Cornerstone**

[20] Inasmuch as *PFEL* concerns effects of mental contamination (and full disclosure is always valuable), I should note that Professor Saks and I worked together for years as colleagues at Arizona State University and as two of the co-authors of the treatise David L. Faigman, David H. Kaye, Michael J. Saks & Joseph Sanders, Modern Scientific Evidence: The Law and Science of Expert Testimony (2002). I know Barbara Spellman only through (some of) her writing and accolades, such as her selection as a fellow of the American Association for the Advancement of Science on the basis of her " 'distinguished contributions to the field of cognitive psychology' for her innovative research on reasoning and memory, and ' for academic service to psychological journals and societies.' " Fariss Samarrai, *U. Va.'s Barbara A. Spellman, Psychology and Law Professor, Named AAAS Fellow*, UVA Today (Jan. 11, 2011), https://news.virginia.edu/content/uvas-barbara-spellman-psychology-and-law-professor-named-aaas-fellow [https://perma.cc/J6WP-X36L].

[21] Emphasis could be placed on the word "distill." *PFEL* is not the place to find complete bibliographies, formal meta-analyses, and filigreed literature reviews.

The panoramic view of *PFEL* blurs some details, and a number of its comments leave subtleties unstated.[22] This Part pursues *PFEL*'s exposition of two foundational concepts in evidence: relevance and probative value. My objective is to dig more deeply into this corner of the law by unveiling alternatives to the psychological (or logical) models mentioned in *PFEL*.[23]

Saks and Spellman rightly observe:

> Often, when rulemakers adopt a general rule of evidence, they do so by trying to compare the optimal or ideal or proper inferences to be drawn from a class of evidence to how they imagine average jurors will interpret the evidence.
>
> . . . .
>
> . . . [R]ulemakers are acting not only as amateur psychologists, but also as amateur logicians, statisticians, and scientists of various kinds. (pp. 7–8; footnotes omitted)

This is clearly true of the interlocking concepts of relevance, probative value, and the "counterweights"[24] that can block the admission of relevant evidence. In particular, psychologists, economists, and other social scientists interested in mathematical models for human inference and decisionmaking have long adapted Bayes's rule of "inverse probability"[25]—described below—to serve as a model for updating a person's assessments of probabilities. Saks and Spellman follow suit, writing that

> [t]he most prominent theory bearing on ideal inferences from evidence is Bayes' Theorem. . . . [T]he theorem deals with the situation in which decision makers update, perhaps repeatedly, their initial estimation of the probability that a certain

---

[22] For more details on many of the topics surveyed in *PFEL*, see Roger C. Park & Michael J. Saks, *Evidence Scholarship Reconsidered: Results of the Interdisciplinary Turn*, 47 B.C. L. Rev. 949 (2006).

[23] *PFEL* also points to a "holistic-story model" for integrating information, pp. 44–46, 61, but it does not present any description of degrees of probative value within a coherence-based theory. In my view, the fact that jurors and judges seek to fit information into a psychologically coherent framework does not preclude using the measures of probative value presented in this Part and in *PFEL*, pp. 59–61, to make judgments about relevance and probative value, but I will not defend this view here. For more elaborate discussion of coherence-based reasoning, see Amalia Amaya, The Tapestry of Reason: An Inquiry into the Nature of Coherence and Its Role in Legal Argument (2015). *See also* Jennifer L. Mnookin, *Atomism, Holism, and the Judicial Assessment of Evidence*, 60 UCLA L. Rev. 1524, 1543 (2013) (observing that "the Federal Rules take a strikingly atomistic approach to determining whether an item is relevant under Rule 401 and hence presumptively admissible under Rule 402").

[24] *See generally* 1 McCormick on Evidence § 185, at 994–1012 (Kenneth S. Broun ed., 7th ed. 2013).

[25] *E.g.*, Clyde H. Coombs et al., Mathematical Psychology 145–47 (1970); Raymond S. Nickerson, Cognition and Chance: The Psychology of Probabilistic Reasoning 109 –42 (2004). On the terminology and history of inverse probability, see Andrew I. Dale, A History of Inverse Probability (2d ed. 1999). As Nickerson and Dale indicate, Laplace should share substantial credit for the version of the rule that today bears Bayes's name.

> conclusion is true. That should sound something like the task of a factfinder in a trial, presented with unfolding evidence. At any given point in the trial, a decision maker has an estimation of the [probability] that a certain conclusion, such as the guilt of the defendant, is true (called the *prior probability*); the decision maker is provided with additional evidence (reflected in the theorem as a likelihood ratio), which enables a revision of that estimation, increasing or decreasing the estimate of the probability of guilt (called the *posterior probability*). Bayes' Theorem has been used to model the concept of relevance in Rule 401—which, in part, states that a fact is relevant if "it has any tendency to make a fact more or less probable than it would be without the evidence"— and to argue for solutions to evidentiary problems facing trial factfinders.[26]

The Bayesian account of relevance that *PFEL* cites is Professor Richard Lempert's [27] lucid exposition of the "likelihood ratio" and the "regret matrix" as heuristic models[28] for defining "relevant" evidence in Rule 401 and explaining the risks of "misleading the jury" and

---

[26] P. 44. They add that "[h]umans are not, however, intuitive Bayesians, and the extent to which conclusions that emerge from Bayesian models should be explicitly presented to juries is controversial." P. 44. This is a reference to the debate initiated by the proposal in Michael O. Finkelstein & William B. Fairley, *A Bayesian Approach to Identification Evidence*, 83 Harv. L. Rev. 489 (1970), to use Bayes's rule to explain to jurors how the rarity of features used to make an identification from trace evidence should affect their judgment of the probability that the defendant is the source of the trace. For discussion of cases in which experts have presented such numbers, see David H. Kaye et al., The New Wigmore: A Treatise on Evidence § 14.3 (2d ed. 2011).

[27] Richard O. Lempert, *Modeling Relevance*, 75 Mich. L. Rev. 1021 (1977).

[28] In speaking of Bayes's rule (or any other part of mathematics) as a "model," it helps to distinguish between descriptive, normative, and heuristic applications. First, whether Bayes's rule is an accurate description of the probability judgments of most people (perhaps in some circumstances but not in others) is an empirical question. For a few opinions on how well Bayesian learning models fit the results of some studies of human cognition under different conditions, see Leda Cosmides & John Tooby, *Are Humans Good Intuitive Statisticians After All? Rethinking Some Conclusions from the Literature on Judgment Under Uncertainty*, 58 Cognition 1 (1996); Joshua B. Tenenbaum et al., *Theory-Based Bayesian Models of Inductive Learning and Reasoning*, 10 Trends in Cognitive Sci. 309 (2006); *cf.* Paul Cisek, *The Currency of Guessing*, 447 Nature 1061 (2007) (research suggesting that monkeys use likelihood ratios in classifying objects); Hannes Rakoczy et al., *Apes Are Intuitive Statisticians*, 131 Cognition 60 (2014).

Second, whether people should conform their judgments to those of the model in certain circumstances is a normative matter. If the law of evidence "wants" jurors to reach results that accord with those prescribed by the model, then we might be able to construct rules that somehow lead jurors to those outcomes (by prompting them to use System 2 reasoning or by affecting inputs for System 1 reasoning).

Finally, if some of the law is structured to bring juror performance more in line with the prescriptions of the models, these models can give law students, lawyers, and judges a more precise understanding of those features of the legal doctrine. In particular, "[a]s a language, mathematics can help clarify those legal rules that involve weighing evidence in an essentially probabilistic fashion," Lempert, *supra* note 27, at 1021–22, and it can explicate the meaning of ambiguous terms. *E.g.*, David Kaye, *Probability Theory Meets Res Ipsa Loquitor*, 77 Mich. L. Rev. 1456–57 (1979).

"prejudice" that Rule 403 requires judges to balance against "probative value." *PFEL* does not delve deeply into either of these models. It does not deploy Lempert's decisionmaking model (or well-known variations of it) to show how evidence could be unfairly prejudicial. It does not explicitly recognize that Bayesian inference is hardly the sole basis for the likelihood-ratio definition of relevance (and perhaps not even the original one).[29] And it endorses a debatable definition of probative value without comparing it to the more mainstream alternative for quantification.[30] To be sure, it is unfair to ask too much of short discussions designed to illustrate broader ideas, but a brief elaboration on the last two points might further cement this part of the "foundations of evidence law." To do so, the remainder of this Review answers three questions: What is a likelihood ratio? What is Bayes's rule? And how are they related to one another and to the legal concepts of relevance and probative value? I describe a way to think about the likelihood ratio that does not depend on Bayes's rule, and I defend the use of the likelihood ratio as a measure of probative value instead of the particular change-in-probability quantity that *PFEL* presents.

### A. *Relevance Defined*

#### 1. The Likelihood Ratio as the Foundation

The "likelihood ratio" is a concept that pervades statistics.[31] As Lempert argued, it can be used to define *whether* an item of evidence is relevant.[32] For example, in the 1990s, researchers

---

[29] *See* John Maynard Keynes, A Treatise on Probability 55 (1921) (defining "irrelevant" evidence as that which does not change a conditional probability); David H. Kaye, *Likelihoodism, Bayesianism, and a Pair of Shoes*, 53 Jurimetrics J. 1, 5–8 (2012) (arguing that the likelihood-ratio definition of probative value does not require a commitment to Bayesian inference at all).

[30] There are, of course, alternatives to quantification as well as alternative ideas for quantification. *See, e.g.*, Kevin M. Clermont, *Trial by Traditional Probability, Relative Plausibility, or Belief Function?*, 66 Case W. Res. L. Rev. 353, 353 (2015) (arguing that Dempster-Shafer "belief functions nicely clarify the workings of burdens of persuasion and production"); Michael S. Pardo & Ronald J. Allen, *Juridical Proof and the Best Explanation*, 27 Law & Phil. 223, 223–25 (2008) (favoring "inference to the best explanation"); Alex Stein, *Inefficient Evidence*, 66 Ala. L. Rev. 423, 424–25 (2015) (invoking the "signal-to-noise ratio").

[31] Vic Barnett, Comparative Statistical Inference 306 (3d ed. 1999) ("The principles of maximum likelihood and of likelihood ratio tests occupy a central place in statistical methodology."); *see, e.g., id.* at 178–80 (describing likelihood ratio tests in frequentist hypothesis testing); N. Reid, *Likelihood*, *in* Statistics in the 21st Century 419 (Adrian E. Raftery et al. eds., 2002).

[32] Lempert, *supra* note 27, at 1025–27.

developed a prostate cancer test based on the level of prostate-specific antigen ("PSA").[33] The test, they said, was far from definitive but still had diagnostic value.[34] Should anyone have believed them? A straightforward method for validation is to run the test on subjects known to have the disease and on other subjects known to be disease-free. The PSA test was shown to give a positive result (to indicate that the cancer is present) about 70% of the time when the cancer is, in fact, present, and about 10% of the time when the cancer is not actually present.[35] Thus, the test has diagnostic value. The doctor and patient can understand that positive results arise more often among patients with the disease than among those without it.[36]

But *why* should we say that the greater probability of the evidence (a positive test result) among cancer patients than among cancer-free patients makes the test diagnostic of prostate cancer? There are three answers. One is that if we use it to sort patients into the two categories, we will (in the long run) do a better job than if we use some totally bogus procedure (such as flipping a coin). This is a frequentist interpretation of diagnostic value.

A second justification takes the notion of "support" for a hypothesis as fundamental.[37] Results that are more probable under a hypothesis $H_1$ about the true state of affairs are stronger evidence for $H_1$ than for any alternative ($H_2$) under which they are less probable. If the evidence were to occur with equal probability under both states, however, the evidence would lend equal support to both possibilities. In this example, such evidence would provide no basis for distinguishing between cancer-free and cancer-afflicted patients. It would have no diagnostic

[33] *See* Nat'l Cancer Inst., Prostate-Specific Antigen (PSA) Test, July 24, 2012, http://www.cancer.gov/types/prostate/psa-fact-sheet [https://perma.cc/PR9ZM4T7].

[34] *Id.*

[35] Carvell T. Nguyen & Michael W. Kattan, *Prediction Models in Prostate Cancer Diagnosis*, *in* Prostate Cancer Diagnosis: PSA, Biopsy and Beyond 85, 86 (J. Stephen Jones ed., 2013). In the nomenclature of clinical medicine, the PSA test is 70% sensitive and 90% specific.

[36] Likewise, negative results occur more often among patients without the disease than among those with it.

[37] A "support function" can be required to have several appealing, formal properties, such as transitivity and additivity. *E.g.*, A. W. F. Edwards, Likelihood 28–32 (Johns Hopkins Univ. Press, expanded ed. 1992) (1972). It also can be derived, in simple cases, from other, arguably more fundamental, principles. *E.g.*, Barnett, *supra* note 31, at 310–11.

value,[38] and the test should be kept off the market. The coin-flipping test is like this. A head is no more or less probable when the cancer is present than when it is absent.

A difference between the frequentist, long-run justification and the likelihoodist, support-based understanding is that the latter applies even when we do not perform or imagine a long series of tests. If it really is more probable to observe the data under one state of affairs than another, it would seem perverse to conclude that the data somehow support the latter over the former. The data are "more consistent" with the state of affairs that makes their appearance on a single occasion more probable (even without the possibility of replication).

The same thing is true of circumstantial evidence in law. Circumstantial evidence $E$ that is just as probable when one party's account is true as it is when that account is false has no value as proof that the account is true or false. It supports both states of nature equally and is logically irrelevant. To condense these observations into a formula, we can write:

$E$ is irrelevant (to choosing between $H_1$ and $H_2$) if $P(E|H_1) = P(E|H_2)$,

where $P(E|H_1)$ and $P(E|H_2)$ are the probabilities of the evidence conditional on ("given the truth of," or just "given") the hypotheses. The conditional probabilities (or quantities that are directly proportional to them) have a special name: likelihoods. So a mathematically equivalent statement is that

$E$ is irrelevant if the likelihood ratio $L = P(E|H_1) / P(E|H_2) = 1$.

A fancier way to express it is that $E$ is irrelevant if the logarithm of $L$ is 0. Such evidence $E$ has zero "weight" when placed on a metaphorical balance scale that aggregates the weight of the evidence in favor of one hypothesis or the other.[39] In this case, the likelihood ratio for a positive test result is 70% ÷ 10% = 7, which is greater than 1. Thus, the test is relevant evidence in deciding whether the patient has cancer.

---

[38] *See* Steven McGee, *Simplifying Likelihood Ratios*, 17 J. Gen. Internal Med. 647, 647 (2002) ("Findings whose [likelihood ratios] equal 1 lack diagnostic value.").

[39] *See generally* I. J. Good, *Weight of Evidence and the Bayesian Likelihood Ratio*, in The Use of Statistics in Forensic Science 85 (C. G. G. Aitken & D. A. Stoney eds., 1991); I. J. Good, *Weight of Evidence: A Brief Survey*, *in* 2 Bayesian Statistics 249 (J.M. Bernardo et al. eds., 1985) (providing background information regarding the use of Bayesian statistics in evaluating weight of evidence). Good's conception of the weight of evidence, see infra note 58, which is used here in spirit, if not in great detail, should not be confused with "Keynesian weight" discussed recently in Dale A. Nance, The Burdens of Proof: Discriminatory Power, Weight of Evidence, and Tenacity of Belief (2016).

## 2. Bayes's Rule as the Foundation

Nothing that I have said so far involves Bayes's rule.[40] "Likelihood" and "support" are the primitive concepts. Lempert argued for a likelihood ratio of 1 as the defining characteristic of relevance by relying on a third justification—the Bayesian model of learning.[41] How does this work? Think of probability as a pile of poker chips. Being 100% certain that a particular hypothesis about the world is correct means that all of the chips sit on top of that hypothesis. Twenty-five percent certainty means that 25% of the chips sit on the same hypothesis, and the remaining 75% are allocated to the other hypotheses.[42] To keep things as simple as possible, let's assume there are only two hypotheses that could be true.[43] To be concrete, let's say that $H_1$ asserts that the individual has cancer and that $H_2$ asserts that he does not. Assume that doctors know that men with this patient's symptoms have a 25% probability of having prostate cancer. We start with 25% of the chips on hypothesis 1 ($H_1$: cancer) and 75% on the alternative ($H_2$: some other cause of the symptoms). Learning that the PSA test is positive for cancer requires us to take some of the chips from $H_2$ and put them on $H_1$. Bayes's rule dictates just how many chips we transfer. The exact amount generally depends on two things: the percentage of chips that were on $H_1$ (the prior probability) and the likelihood ratio $L$ in this simple situation. The top panel of Figure 1 illustrates the reallocation of probability (visualized as the height of piles of chips) in light of new evidence E (the elevated PSA level that occurs 7 times as often when the cancer is present than when it is absent). The bottom panel shows the reallocation when $L = 1$.

---

[40] *Cf.* D. H. Kaye, *Quantifying Probative Value*, 66 B.U. L. Rev. 761, 763–66 (1986) (discussing the likelihood ratio and log-*L* as a measure of the degree to which evidence has probative value).

[41] Lempert, *supra* note 27, at 1025.

[42] If the individual were to keep some of the chips in reserve, the analogy between the fraction of them on a hypothesis and the kind of probability that pertains to random events such as games of chance would break down.

[43] These hypotheses might be quite specific—for example, either the defendant forged the signature of his deceased brother on the suicide note ($H_1$), or the brother wrote the signature ($H_2$) in a case in which no one else could have written the note—or they might be a conjunction of propositions—for example, the prosecution's account of the alleged crime ($H_1$) and the only other account ($H_2$) that the jury considers even vaguely plausible.

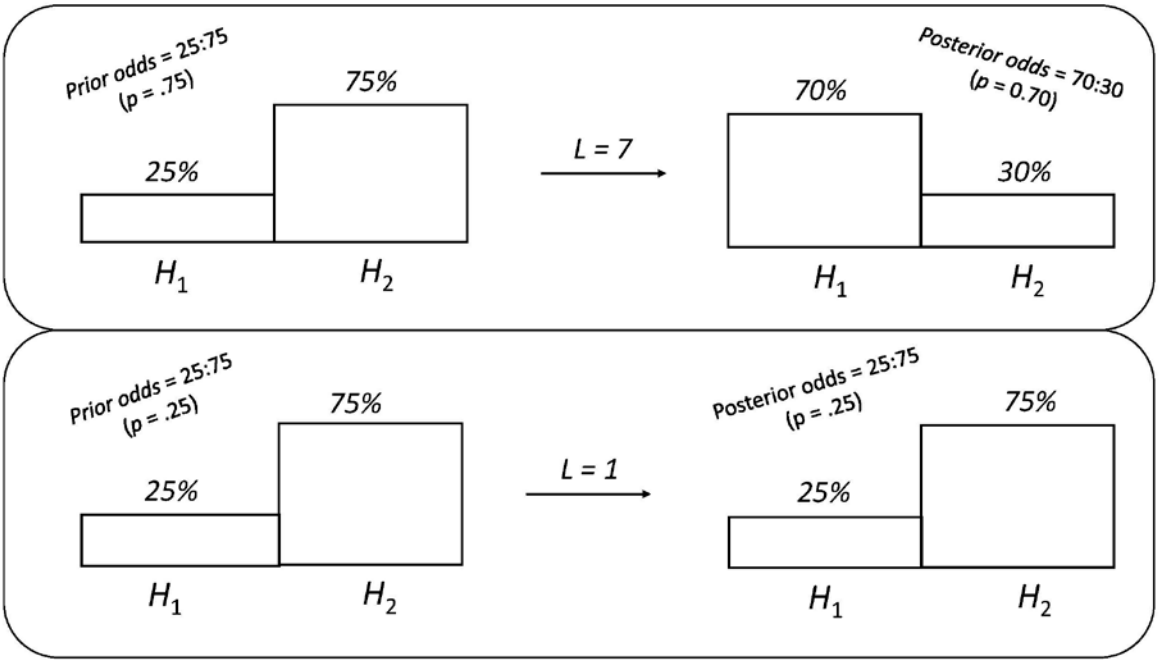


FIGURE 1. Bayes's rule applied to two mutually exclusive, collectively exhaustive hypotheses with prior odds of 25:75 (1:3) and likelihood ratios of 7 (top panel) and 1 (bottom panel).

This picture follows from the very simple structure of Bayes' rule in this case:

$$Odds(H_1) \cdot L = Odds(H_1|E).$$

The rule requires updating the "prior odds" (on the left-hand side) by multiplying by the Bayes factor (which also is the likelihood ratio $L$) to arrive at the "posterior odds" (on the right-hand side). Each set of odds is just the ratio of the percentage of chips given to each hypothesis before, and then after, receiving the evidence. These odds are implicit in the relative heights of the piles of chips. The initial pile of 25% of the chips on $H_1$ corresponds to the prior odds of 25:75 = 1:3. There is one chip on $H_1$ to every three chips on $H_2$.

The crucial point is that multiplication by $L = 1$ never changes the prior odds. Evidence that is equally probable under each hypothesis produces no change in the allocation of the chips—*no matter what the initial distribution*. Prior odds of 1:3 become posterior odds of 1:3. Prior odds of 10,000:1 become posterior odds of 10,000:1. The evidence is never worth considering. Again, we can get fancy and place the odds and the likelihood ratio on a logarithmic scale. Then the posterior log-odds are the prior log-odds plus the weight of the evidence (*WOE* = log-*L*):

*New LO = Prior LO + WOE*.[44]

Evidence that has zero weight ($L = 1$, log-$L = 0$) leaves us where we started. Evidence $E$ that does not change the odds (and, hence, the corresponding probability) is uninformative—it is irrelevant. Inversely, evidence that does change the probability is relevant—as Rule 401 states in near-identical terms. This, in a nutshell, is the Bayesian explanation of the rule as it applies to circumstantial evidence. It tracks the text of the rule better than the likelihoodist, support-based analysis, but both lead to the conclusion that relevance vel non turns on whether the likelihood ratio departs from 1.

### B. *Probative Value Quantified*

I have expended some effort to give separate accounts of Rule 401 partly to correct the widespread perception among legal scholars (and in forensic science as well) that likelihood ratios are uniquely Bayesian. The effort is also useful as preparation for investigating *PFEL*'s quantification of probative value. Attempts at quantification might seem purely theoretical, but they have important practical implications. Like pregnancy, relevance is strictly binary. Evidence either is or is not relevant to a material fact. In contrast, probative value comes in an infinite number of gradations, and Rule 403 asks judges to balance it against a set of counterweights.[45] But if several divergent ways to quantify the value are equally plausible, advocates can pick the one they want to argue that an item of evidence is or is not very probative.[46] Moreover, expert witnesses can, and do, describe the implications and probative value of their data in various ways. Forensic scientists and statisticians still debate how best to express the numerical "weight of evidence"—their term for "probative value"—for scientific test results.[47] The simple likelihood ratio is the basic measure that dominates the forensic science literature on evaluative conclusions.[48] However, most writers in this area construe the likelihood ratio as the ratio of

---

[44] A deeper motivation for using logarithms may lie in information theory, but, if so, it is not important here. *See* Solomon Kullback, Information Theory and Statistics (1959).

[45] *See* Fed. R. Evid. 403 advisory committee's note.

[46] *See* Adversarial Forum, *Further Comments on the Probative Value of Evidence*, 27 Law & Hum. Behav. 623 (2003).

[47] *See, e.g.*, Nat'l Inst. of Standards and Tech., *Technical Colloquium on Quantifying the Weight of Forensic Evidence: Online Proceedings*, May 5–6, 2016, http://www.nist.gov/itl/iad/ig/evidential_value_proceedings.cfm [http://perma.cc/X8XC-8NRX].

[48] *E.g.*, John S. Buckleton et al., *An Extended Likelihood Ratio Framework for Interpreting Evidence*, 46 Sci. & Just. 69, 70 (2006) ("The idea of assessing the weight of evidence using a relative measure (known

posterior odds to prior odds and base its use on that purely Bayesian interpretation.[49] Greater clarity would come from using the related term "Bayes factor"[50] when this is the motivation for the ratio.[51]

How, then, do Saks and Spellman think probative value should be quantified? The measure they describe (and apparently endorse) is neither the likelihood ratio nor the Bayes factor. It is the arithmetic difference between the prior and posterior probabilities—that is, $P(H_1|E) - P(H_1)$. This quantity is the marginal probability $\Delta P$ that results from $E$. They present and use this $\Delta P$ rule as follows:

> One way of quantifying probative value is by considering the probability of the fact given the piece of evidence (the "conditional probability") and subtracting the probability of the fact when the evidence is unknown. Suppose the fact we want to know is whether the driver-defendant was driving the getaway car. Before the first witness's testimony that the driver-defendant owns that type of car, we have little reason to believe that he was the driver. The witness's testimony that he owns such a car makes us somewhat more likely to believe he was the driver; that difference between our before and after beliefs in his guilt is the probative value of that

---

as the likelihood ratio) . . . dominates the literature as the method of choice for interpreting forensic evidence across evidence types.") (citations omitted); Cedric Neumann et al., *Presenting Quantitative and Qualitative Information on Forensic Science Evidence in the Courtroom*, 29 Chance 37, 37 (2016) (referring to "the abundant literature published over the past 30 years advocating . . . [that] forensic scientists should report the relative support that forensic evidence provides to each side of the legal argument using a Bayes factor (also sometimes referred to as a likelihood ratio . . . )"); Anders Nordgaard et al., *Scale of Conclusions for the Value of Evidence*, 11 Law, Probability & Risk 1, 8 (2012) ("The state-of-art in forensic interpretation is to evaluate forensic evidence with the use of a likelihood ratio.").

[49] *E.g.*, Nordgaard, *supra* note 48, at 8.

[50] Robert E. Kass & Adrian E. Raftery, *Bayes Factors*, 90 J. Am. Stat. Ass'n 773, 776 (1995); *see* Peter M. Lee, Bayesian Statistics 140 (4th ed. 2012).

[51] The choice of words is not merely a labeling issue. In simple situations, the Bayes factor and the likelihood ratio are numerically equivalent, but more generally, there are conceptual and operational differences. For instance, simple likelihood ratios can be used to gauge relative support within any pair of hypotheses, even when the pair is not exhaustive. But when there are many hypotheses, the Bayes factor is not so simple. *See* Lee, *supra* note 50, at 141–42. It becomes the usual numerator divided by a weighted sum of the likelihoods for each hypothesis. The weights are the probabilities (conditional on the falsity of the hypothesis in the numerator). For an example, see Tacha Hicks et al., *A Framework for Interpreting Evidence*, *in* Forensic DNA Evidence Interpretation 37, 63 (John S. Buckleton et al. eds., 2d ed. 2016). Furthermore, there is disagreement over the use of a likelihood ratio for highly multidimensional data (such as fingerprint patterns and bullet striations) and whether and how to express uncertainty with respect to the likelihood ratio itself. *Compare* Franco Taroni et al., *Dismissal of the Illusion of Uncertainty in the Assessment of a Likelihood Ratio*, 15 Law, Probability & Risk 1, 2 (2016), with M. J. Sjerps et al., *Uncertainty and LR: To Integrate or Not to Integrate, That's the Question*, 15 Law, Probability & Risk 23, 23–26 (2016). My explanation of the Bayes factor in 1 McCormick, *supra* note 24, § 185, at 997 n.14, does not mention these subtleties.

> evidence. Now we have some belief that he was the driver and next we learn from the second witness that the two defendants are good friends. Our belief in his guilt rises once again. Finally, we learn that the second witness might have reason to lie, so we devalue the testimony of the second witness, and our belief in the robber-defendant's guilt decreases. Thus, relevant evidence can make a fact more or less probable, and it is the size of that difference in the probability of the fact with and without the evidence that represents its probative value. (Pp. 60–61; footnote omitted)

Notice that "the size of that difference in the probability of the fact" is the percentage of all the chips transferred according to Bayes's rule. It is not the relative change in the odds. It conflates that change (the Bayes factor) with the particular prior odds. When the prior odds on $H_1$ are large, more of the probability chips are on $H_2$, and equally powerful evidence moves fewer chips from $H_2$ to $H_1$ than when the prior odds are small (and there are many more chips to move). Saks and Spellman are well aware of this. They continue:

> Note that the probative value of a piece of evidence depends on what other evidence exists in the case. So, for example, when a first witness identifies the robber-defendant as the bank robber, that evidence has a lot of probative value. . . . When a second testifies to the same thing, that testimony has less probative value. And so, too with the next witness who has the same thing to say. The same testimony from a tenth witness almost certainly adds nothing to any juror's belief in the likelihood that the defendant committed the robbery; thus, it has no probative value even if, had the tenth witness been the first to testify, his testimony would have been quite probative. . . . This feature, that additional similar testimony decreases in probative value, is related to some of the factors on the "negative" side of the balancing scale regarding trial efficiency . . . . (p. 60)

This is indeed one way to conceive of the probative value of each item of evidence, but using the probabilities of the hypothesis in question to define probative value generates some anomalies. Rule 403 calls for balancing probative value against the trial-efficiency concerns of "undue delay, wasting time, or needlessly presenting cumulative evidence." But if the redundant "testimony from a tenth witness . . . has no probative value," how can it even be relevant and therefore subject to Rule 403 balancing? Certainly, such balancing occurs in the context of what has gone before, but the drafters spoke of "probative value" as a separate factor from the "need for the evidence."[52] Their notes suggest that the two factors must be combined and then balanced

[52] Fed. R. Evid. 403 advisory committee's note ("Situations in this area call for balancing the probative value of and need for the evidence against the harm likely to result from its admission.").

"against the harm likely to result from . . . admission."[53] The Supreme Court has described these two factors as necessitating the use of "discounted probative value,"[54] and the *McCormick* treatise uses the phrase "marginal probative value."[55] Such phrases seem to presume an "intrinsic probative value"[56] that does not depend on other, independent items of evidence. That kind of probative value is more like the purchasing power of money. The need for additional income may decline as one's wealth grows, but a dollar is a dollar, notwithstanding the diminishing marginal utility of additional dollars. Moreover, by making probative value depend on unrelated as well as related evidence, the marginal probability $\Delta\underline{P}$ can be essentially the same for probative and nonprobative evidence, making it a peculiar choice as a metric of probative value.[57]

The obvious Bayesian measure of probative value is the Bayes factor (*B*). In the examples used here, *B* is equal to the likelihood ratio $\underline{L}$, and therefore the statisticians' "weight

---

[53] *Id.*

[54] Old Chief v. United States, 519 U.S. 172, 182–83 (1997) ("If an alternative were found to have substantially the same or greater probative value but a lower danger of unfair prejudice, sound judicial discretion would discount the value of the item first offered and exclude it if its discounted probative value were substantially outweighed by unfairly prejudicial risk.").

[55] 1 McCormick, *supra* note 24, § 185, at 1009 n.66, § 206, at 1234; *see also* Dan M. Kahan, Essay, *The Economics—Conventional, Behavioral, and Political—of "Subsequent Remedial Measures" Evidence*, 110 Colum. L. Rev. 1616, 1639 (2010). Although I introduced this phrase into the third edition of *McCormick*, I now have some qualms about the term. For cumulative evidence, marginal probative value is the additional weight that *E* supplies compared to the already-introduced evidence on the same point. If *L* is conditionally independent of the other evidence, then *E*'s probative value is *WOE* = log-*L*, and all of this weight is added to the prior weight. As such, it appears that the marginal probative value can be large even when the evidence is cumulative. This seems paradoxical, because evidence that is merely cumulative does not supply much more information to the jury, and log-*L* is a measure of information.

I think the source of this confusion is the meaning of "cumulative." It could mean evidence that provides a good deal of information, but that jurors do not need because they have so much prior information that their minds already are made up. This situation is analogous to the diminishing marginal utility of constant sums of money. Alternatively, "cumulative" could mean evidence that is just redundant—much like having a single witness repeat himself over and over. The tail end of that kind of cumulative evidence has almost no probative value, because the $\underline{n}$th repetition will probably be the same as the rest, both when $H_1$ is true and when $H_2$ is true. Therefore, the likelihood ratio associated with the *n*th repetition is very close to 1. Formally, $L_{\underline{n}} = P(E_n|H_1, E_1, \ldots, E_{\underline{n}\text{-}1}) / P(E_{10}|H_2, E_1, \ldots, E_{\underline{n}\text{-}1}) \approx 1$ even if $L_1 = P(E_1|H_1) / P(E_1|H_2)$ is not close to 1, so the terms in the series of likelihood ratios for such cumulative evidence are not constant (unless the evidentiary items are conditionally independent). Consequently, the likelihood-ratio definition of probative value may respond to cumulative evidence without invoking the notion of marginal, instead of total, probative value.

[56] Kaye, *supra* note 40, at 766.

[57] When the prior probability is already near 1, neither strong nor weak evidence will make much of a difference. With the $\Delta P$ measure, testimony from saints and sinners alike has largely the same probative value in this range.

of evidence" is *WOE* = log-*B* = log-*L*.[58] The value of *L* in these cases tells us just how much more the evidence supports one theory than another and hence—this is the Bayesian part—just how much we should adjust our belief (expressed as odds) for any starting point. For the PSA test for cancer, *L* = 7 is "the change in odds favoring disease."[59] A test with greater diagnostic value would have a larger likelihood ratio and induce a stronger shift toward that conclusion. Linking diagnostic or probative value to the prior odds makes such comparisons between the diagnostic values of the different tests difficult, since the value for each changes according to other information about the patient. Instead, the likelihood-ratio measure (or variations on it),[60] which keeps prior probabilities out of the picture, is more typically used to describe the value of test results as evidence of disease or other conditions in medicine and psychology.[61]

Using the same measure in law has significant advantages.[62] Identifying probative value with the likelihood ratio (or simple mathematical transformations of it) clarifies statements that the evidence should be excluded because the jury will misjudge its probative value and give the

[58] The logarithm of *B* has been called "weight of evidence" since 1878. I. J. Good, *A. M. Turing's Statistical Work in World War II*, 66 Biometrika 393, 393 (1979); *accord* Lee, *supra* note 50, at 127. While working in the town of Banbury to decipher German codes, Alan Turing famously (in cryptanalysis and statistics, at least) coined the term "ban" to designate a power of 10 for this metaphorical weight. Good, *supra* at 394. Thus, a *B* of 10 is 1 ban, 100 is 2 ban, and so on.

[59] David L. Simel et al., *Likelihood Ratios with Confidence: Sample Size Estimation for Diagnostic Test Studies*, 44 J. Clinical Epidemiology 763, 763 (1991).

[60] *See, e.g.*, Afina S. Glas et al., *The Diagnostic Odds Ratio: A Single Indicator of Test Performance*, 56 J. Clinical Epidemiology 1129 (2003).

[61] *E.g.*, Jonathan J. Deeks & Douglas G. Altman, *Diagnostic Tests 4: Likelihood Ratios*, 329 Brit. Med. J. 168 (2004) ("Likelihood ratios [summarize] diagnostic accuracy [and] have several particularly powerful properties that make them more useful clinically than other statistics."); D. H. Kaye & Jonathan J. Koehler, *The Misquantification of Probative Value*, 27 Law & Hum. Behav. 645, 649 (2003) (citing authorities).

[62] *See* Kaye & Koehler, *supra* note 61, at 645–47 (arguing against a related arithmetic-difference measure used in court by two psychologists). Most legal commentators seem to agree. *See, e.g.*, Nance, *supra* note 39, at 70 n. 184, 95–96; Christopher Slobogin, Proving the Unprovable: The Role of Law, Science, and Speculation in Adjudicating Culpability and Dangerousness 45 (Am. Psychology-Law Soc'y Series, 2007) ("[T]he ratio is an eminently sensible way of evaluating the probative value of evidence."); Kahan, *supra* note 55; Lempert, *supra* note 27. A few authors use the term "relevance ratio" for the likelihood ratio. *See* Slobogin, *supra*, at 45; Thomas D. Lyon & Jonathan J. Koehler, *The Relevance Ratio: Evaluating the Probative Value of Expert Testimony in Child Sexual Abuse Cases*, 82 Cornell L. Rev. 43 (1996).

evidence more weight than it "logically" deserves.[63] It is a natural and productive way to interpret the phrase "misleading the jury" in Rule 403.[64] The disparity between the estimate of $\underline{L}$ for ideal and actual jurors is an "estimation problem"[65] that can militate against admission of evidence. For example, DNA experts sometimes find it difficult to compute likelihood ratios for complex DNA mixtures discovered at crime scenes. If the State will not provide a statistical assessment, an advocate might oppose the admission of evidence that a defendant's DNA is such that he is "included" as a possible contributor on the theory that jurors will think that inclusion is extremely improbable if the defendant is not a contributor—making the denominator of the likelihood ratio quite small, when, in reality, it could be quite high. Defining misestimation of weight where weight is given by $\Delta\underline{P}$ is also possible, but this conception of weight makes such arguments more complex.

In fact, Saks and Spellman themselves revert to either the likelihood-ratio or Bayes-factor conceptions of probative value when presenting an example of circumstantial evidence that raises an estimation problem. They characterize the evidence as "quite relevant" simply because it has a high likelihood ratio. The example involves battering as evidence of the identity of an alleged murderer:[66]

> For example, if the question is whether a murdered woman was killed by her husband/boyfriend, is it useful to know that the man physically abused the woman during their years together? Suppose we know that of every 100,000 battered women, 45 end up being murdered (by anyone) and 99,955 are never murdered. This information would not help you predict that a battered woman will end up being a murdered woman, since it so rarely happens. But the question before a jury is different: given that a woman *has been* murdered, is it helpful to know that she had been a battered woman and that her partner was the batterer? Suppose the data show that of the 45 out of 100,000 battered women who are murdered, 40 are murdered by their partners and 5 by someone else. Now we know that battered women who are murdered are eight times more likely to have been murdered by their partners than by someone else. This clearly is relevant information by the law's definition. (p. 161)

---

[63] For an example of such phrasing, see Bruton v. United States, 391 U.S. 123, 138 (1968) (Stewart, J., concurring) (referring to "certain kinds of hearsay . . . at once so damaging, so suspect, and yet so difficult to discount, that jurors cannot be trusted to give such evidence the minimal weight it logically deserves").

[64] Fed. R. Evid. 403.

[65] Lempert, *supra* note 27, at 1028–29.

[66] Pp. 161–62. Conceivably, it is inspired by the improperly framed conditional probabilities for murder and battering in the 1994 trial of O.J. Simpson. *See* Steven Strogatz, The Joy of $\underline{x}$: A Guided Tour of Math, from One to Infinity 287–89 (2012).

"The information provided above," they write, is "known as a likelihood ratio" (p. 161), and it creates an estimation problem because the jurors might "take the evidence of prior battering to suggest that it makes the defendant 800 times as likely to have committed the murder, rather than only 8 times as likely" (p. 162). There is some confusion here, because statements about how "likely [battered and murdered women are] to have been murdered by their partners" (uxoricide) or "by someone else" (nonuxoricide murder) are statements about posterior probabilities for uxoricide and nonuxoricide of the murdered partner.[67] A frequency-based likelihood ratio depends on how much more often battering occurs among cases of uxoricide than among the cases of nonuxoricide murder. The key point, however, is that the problem of probative but misleading evidence can be conceptualized fairly well with the likelihood-ratio or Bayes-factor definition of probative value. If the $\Delta P$ definition is as helpful, Saks and Spellman's description of this example does not show it.

This is not to argue against the importance of prior odds for jurors in reaching decisions and for "metacognitive" judicial consideration of those odds in applying Rule 403 (p. 27). Suppose that a man's PSA test is positive and that the prevalence of prostate cancer in similar patients is 1 in 11. This statistic suggests the prior odds are 1 to 10. Bayes's rule instructs us to multiply these odds by 7. The high PSA level in this patient has raised the odds from 1/10 to 7/10. Odds of 7 to 10 are the same as a probability of 7/(10 + 7) = 0.41. With a 41% probability of cancer, a needle biopsy might well be advisable. But if the prevalence in similar patients were only 1 in 1001, the prior odds of 1/1000 would become 7/1000, for a posterior probability of only 7/1001, or 0.7%. The need for the same action is less clear. But the difference does not reflect a reduction in the power of the PSA test as a classifier of men with and without cancer.

Analogously, the probative value of legal evidence—for example, a scrupulously conducted DNA test showing an unambiguous match between a crime-scene sample and a defendant at many loci—is always highly probative with respect to a theory of the case that includes the hypothesis $H_1$ that the defendant's DNA was left at the crime scene. The likelihood

[67] P. 161. In the symbolism we have been using, the proof of battering by the partner is $E$; uxoricide is $H_1$; and non-uxoricide murder is $H_2$. $L$ is $P(E|H_1) / P(E|H_2)$ rather than $P(H_1|E) / P(H_2|E)$. The last quantity is the posterior odds on $H_1$. *PFEL* generally respects this distinction. The confusion in the wording (or in my reading of the words) speaks to the difficulty of describing conditional probabilities without using words that transpose (or seem to transpose) the propositions in conditional probabilities. Pictures, like the one on page 228, can be clearer.

ratio might be 1,000,000. Weighing in at six bans,[68] this evidence is quite powerful. But $\Delta P$ does not cleanly separate the power of the evidence to alter a prior belief from the strength of the prior belief itself. Using $\Delta P$, we would have to say that the same DNA evidence is not very probative if a security camera videotape and multiple fingerprints also placed the defendant at the scene of the crime. It seems clearer to say that each type of evidence is highly probative—because each carries considerable weight—but that any one or two of them would be ample to link the defendant's DNA to the crime scene. We should distinguish between *sufficiency*, which relates to the posterior probability after all the evidence is weighed (on atomic or holistic bases), and *probativity*, which relates to the support that each item of evidence (or a subset of the evidence) contributes.[69] Defining probative value as the Bayes factor captures the idea of probativity and distinguishes it from sufficiency. The $\Delta \underline{P}$ definition lacks this analytical clarity and implies that additional evidence that is very accurate in discriminating between two states of nature is not probative when it is not necessary.

In short, the Bayesian model of inference that *PFEL* invokes is one way to think about relevance, probative value, and the counterweights. But this possible foundation for Rules 401 and 403 is not the only justification for the likelihood-ratio conception of the relevance and probative value of circumstantial evidence. In addition, even within the Bayesian framework, it may be clearer to define probative value not as the difference in the before-and-after probabilities for the hypothesis, but as the Bayes factor (or its logarithm).

## Conclusion

Having questioned some of the statements in *PFEL* about relevance and probative value, I want to return to *PFEL*'s broader message that these concepts are amenable to clarification from the perspectives of psychology and related disciplines. I hope that the second part of this Review has reinforced this message by further excavating one corner of the foundation. As I stated at the outset, *PFEL* is a gem. To the extent I have criticized some small points, my remarks simply indicate that—like all gems—it can benefit from gentle polishing. Integrating and applying the variegated body of work in psychology, logic, philosophy, probability, statistics, and forensic science that bears on the theoretical and practical facets of the law of

[68] *See supra* note 58.

[69] *See* Kaye & Koehler, *supra* note 61, at 647, 655–56.

evidence is almost too much to ask of one or two scholars. Saks and Spellman's willingness to dig into the psychological assumptions and foundations of evidence law is a boon to both psychology and law.