ARTICLE

ESTIMATION EVIDENCE

JONAH B. GELBACH†

This Article radically rethinks the treatment of statistical estimation evidence in civil litigation, focusing for convenience on the federal courts. It proposes an approach that harmonizes legal standards and statistical concepts, replacing the arbitrary and

† Professor of Law, University of California at Berkeley Law School. I thank Matthew Adler, Ron Allen, Andrew Baker, Bobby Bartlett, Bill Bratton, Steve Burbank, Michelle Burris, David Card, John Donohue, Ryan Doerfler, David Eil, Aaron Edlin, William Eskridge, Jill Fisch, Josh Fischman, Jean Galbraith, Michael Gilbert, Maria Glover, Jacob Goldin, Michael Green, J.B. Heaton, Dan Ho, Derek Ho, David Hoffman, William Hubbard, David Kaye, Jon Klick, Bruce Kobayashi, Justin McCrary, Robert Merges, Greg Mitchell, Michael Pardo, Andrea Roth, Frederick Schauer, Steven Davidoff Solomon, Sean Sullivan, Andrew Verstein, Tobias Barrington Wolff, and workshop participants at Berkeley, Georgetown, Penn, St. John’s, the University of Virginia, and Wake Forest for helpful conversations and comments.
elevated standards of conventional hypothesis testing with an approach that fits what we otherwise think the preponderance standard means.

Through careful argumentation using Bayesian hypothesis testing, the Article offers a fundamental default rule for statistical estimation evidence: if the evidence comes from a credibly designed and implemented study, then it is presumptively admissible and enough to withstand motions for judgment under either Rule 50 or Rule 56, whenever it points in the direction of the party offering it. In many situations, the fundamental default rule may be cast as the policy that evidence is legally sufficient and presumptively admissible whenever it yields a finding of statistical significance at the 50 percent level—equivalently, whenever its p-value is less than 0.5. Thus, the Article indicates that many courts’ current practice is far too strict with respect to statistical estimation evidence.

The Article also discusses the appropriate gatekeeping role for federal district courts under Federal Rules of Evidence 403 and 702, and it engages the question of when courts might legitimately move away from the fundamental default rule for policy reasons.

INTRODUCTION ................................................................. 551
I. BAYESIAN HYPOTHESIS TESTING APPLIED TO LEGAL SUFFICIENCY AND THE PREPONDERANCE STANDARD YIELDS
   A FUNDAMENTAL DEFAULT RULE ........................................... 560
      A. A Brief Description of Conventional Hypothesis Testing ............ 561
      B. The Bayesian Hypothesis Testing Approach ............................ 564
      C. Why Bayesian Hypothesis Testing is the Right Framework For Legal Sufficiency ............................................................... 568
      D. The Fundamental Default Rule Results from Viewing the Evidence in the Light Most Favorable to the Non-Movant ....................... 574
      E. Conventional Hypothesis Testing Leads to an Inappropriate Rule For Determining Legal Sufficiency of Statistical Estimation Evidence, Unless the Significance Level of 50% is Used ........................................... 581
      F. Summary of Legal Sufficiency and the Preponderance Standard .... 584
II. HOW TO ANALYZE ADMISSIBILITY OF STATISTICAL ESTIMATION EVIDENCE UNDER THE FEDERAL RULES AND ASSOCIATED CASE LAW ................................................................. 585
      A. Reliability Under Rule 702 and the Daubert Trilogy .................... 587
      B. Understanding the Concepts of Strength and Credibility of Estimation Evidence ............................................................... 590
         1. Technical Implementation .................................................. 592
         2. Credibility for the Proffered Purpose .................................... 592
         3. Analogy to Non-Statistical Evidence .................................... 595
C. The Proper Application of Rule 702 to Estimation Evidence Is Friendly to the Fundamental Default Rule ........................................... 596
1. Rule 702 and the Daubert Trilogy Do Not Allow
   Gatekeeping Based on Conventional Hypothesis Testing
   at Conventional Significance Levels .................................. 597
   a. Daubert ........................................................................ 599
   b. Joiner .......................................................................... 602
   c. Kumho Tire .................................................................... 607
2. The Proper Focus in Gatekeeping is on Technical
   Implementation and Credibility for the Proper Purpose......609
   a. Credibility for the Proffered Purpose: Mayor of
      Philadelphia v. Educational Equality League ............... 610
   b. Credibility for the Proffered Purpose: Merck v. Garza..... 611
   c. Cherry Picking .............................................................. 612
   d. Summary ....................................................................... 613

III. Policy Considerations Related to the Fundamental
     Default Rule.....................................................................613
     A. The Fundamental Default Rule Would Improve Administrability
        and Litigation Practice .................................................. 614
     B. The Quantity of Litigation Activity Might or Might Not Increase
        with the Fundamental Default Rule, but the Rule Likely Would
        Shift Bargaining Power to Plaintiffs .................................617

IV. FEDERAL COURTS COULD USE COMMON LAW POWERS TO
    ADJUST THE EVIDENTIARY STANDARD IN CASES BASED ON
    FEDERAL LAW .................................................................. 621
     A. In Cases Based on State Law Claims, the State’s Law as to
        Statistical Estimation Evidence Should Control in Federal Court.... 624
     B. The Supreme Court Has Common Law Powers to Alter the
        Standard of Evidence for Federal Law Claims, but It Should Use
        Them Transparently Rather Than Characterizing These Powers in
        Terms of the Federal Rules of Evidence .............................. 627

CONCLUSION ........................................................................... 630
APPENDIX: MULTIPLE STUDIES ............................................. 630

INTRODUCTION

This Article sets forth a theory of how federal courts should handle statistical estimation evidence—quantitative estimates with hypothesis testing used to quantify its strength—in civil litigation under certain
important conditions.¹ An example of statistical estimation evidence, discussed in detail below, is the use of data from a randomized controlled trial of the effects of the drug Lipitor, which thousands of plaintiffs alleged caused them to develop Type 2 diabetes.² In the litigation that ensued, this statistical estimation evidence went to the question of general causation—whether Lipitor does, in general, cause diabetes. The disposition of many cases turned on this question.

In addition to mass torts such as the Lipitor case just noted, statistical evidence plays an important role in federal litigation involving other important areas of the law, including employment discrimination cases involving race- or sex-based differences in promotion rates; securities fraud litigation involving changes in stock prices on dates of alleged corrective disclosures; and antitrust cases involving effects on prices in concentrated industries.

My central claim in this Article is that litigants and courts are applying the wrong standard when—as has often been the case—they use statistical standards conventionally used by scholars in scholarly work. I develop this argument in several pieces. First, in Part I, I address the requirements imposed on statistical estimation evidence by the dominant preponderance standard for proof in civil litigation. In this Part, I summarize the conventional hypothesis testing approach, which is based on statistical significance testing, and contrast it to the Bayesian hypothesis testing alternative. I then argue that the Bayesian hypothesis testing approach fits the preponderance standard much better than does the conventional hypothesis testing approach. And I develop what I call the “fundamental default rule of statistical estimation evidence.”³ Under this rule, when the preponderance standard applies, statistical estimation evidence should be considered legally sufficient and presumptively admissible whenever it points in the direction of the party proffering it. As I discuss in Section I.E, this result may be understood in terms of conventional hypothesis testing methodology, but with the unusually high significance level of 50 percent.⁴

¹ By statistical estimation evidence, I mean quantitative evidence whose importance to the case typically is assessed through statistical hypothesis testing (which I discuss in detail in Sections I.A–I.C). I use the term “statistical estimation evidence” so as to distinguish it from classics of “statistical evidence” like the Blue Bus problem, see, e.g., Edward K. Cheng, Reconceptualizing the Burden of Proof, 122 YALE L.J. 1254, 1273–74 (2013), for which I offer nothing new here. Scholars discussing those problems simply take it as given that the proffered probabilities are correct, whereas statistical estimation evidence involves using methods of statistical inference—hypothesis testing—to determine what to believe about probabilities of interest.
² See infra note 10.
³ It is important to understand that the fundamental default rule follows only under certain assumptions regarding a construct called the plaintiff’s most favorable juror; see infra Section I.D for details.
⁴ This result is not the result of a common fallacy involving the general confusion of p-values and posterior probabilities; see infra note 93 for details.
In Part II, I turn to issues related to admissibility of expert testimony related to statistical estimation evidence. If estimation evidence makes it into the trial record at all, it is virtually always through that channel. Litigants battle to get their experts’ testimony admitted and to exclude the other side’s. With respect to evidence deemed admissible, the parties struggle just as vigorously over summary judgment. Parties dispute the methods each other’s experts use, they argue that claimed statistical significance is illusory, and sometimes they even accuse each other’s experts of doing “junk science” in violation of the standards set forth by Federal Rule of Evidence 702 and the Daubert trilogy. Some judges display a sophisticated grasp of statistical concepts in determining which expert testimony to admit, and which will be deemed legally sufficient. Others struggle to referee the battle of the experts.

Litigants and courts frequently focus on conventional hypothesis testing, by which I mean null hypothesis significance testing, typically at the significance level of 5%, because that is the approach many statistics-using scholars take in their scholarly activities. But as Professor Stephen Burbank once wrote: “Courtrooms are not laboratories.” Professor Frederick Schauer has elaborated on this point, explaining that sometimes “bad science makes good law.” As I shall argue, acting otherwise leads to statistical standards that

---


7 Stephen B. Burbank, The Good, The Bad, and The Ugly, 79 Judicature 318, 322 (1996). Professor Burbank went on to note that “it may be a mistake to let science furnish not only evidence with which we adjudicate controversies but the standards for deciding whether evidence can be considered.” Id. See also Susan Haack, Legal Probabilism: An Epistemological Dissent, in Evidence Matters: Science, Proof, and Truth in the Law 55 (2014) (“[A] trial is very different from an open-ended scientific or scholarly investigation sifting for as long as it takes through all the evidence that can be had.”); Richard O. Lempert, Uncovering ‘Nondiscernible’ Differences: Empirical Research and the Jury–Size Cases, 73 Mich. L. Rev. 643, 659 (1975) (“The values of social science, however, are not the values of the law.”). As Daubert itself put it:

The scientific project is advanced by broad and wide-ranging consideration of a multitude of hypotheses, for those that are incorrect will eventually be shown to be so, and that in itself is an advance. Conjectures that are probably wrong are of little use, however, in the project of reaching a quick, final, and binding legal judgment—often of great consequence—about a particular set of events in the past.

Daubert, 509 U.S. at 597.

8 Frederick Schauer, Can Bad Science Be Good Evidence? Neuroscience, Lie-Detection, and Beyond, 95 Cornell L. Rev. 1191, 1216 (2010); see also Richard A. Posner, An Economic Approach to the Law of Evidence, 51 Stan. L. Rev. 1477, 1511 (1999) (“The five percent convention is rooted in considerations that have no direct relevance to litigation, such as the need to ration pages in scientific journals.”); Frederick Schauer, Neuroscience, Lie-Detection, and the Law: A Contrarian
map poorly onto the legal standards that courts otherwise say they use for civil litigation.9

In Part III, I turn to policy considerations that adopting my analysis and standard would implicate. These relate to administrability, which would be greatly improved, and to the social costs and benefits associated with what would likely be a shift in bargaining power toward various types of plaintiffs in complex litigation.

In Part IV, I ask whether courts might prefer to apply more demanding standards than the preponderance standard, given my claims about that standard in Parts I and II. This is a classic question about the legitimate powers of courts to make substantive law. Whereas state courts generally have such powers, the federal courts on which I concentrate in this Article have such powers only in limited circumstances. I address these circumstances, differentiating between claims rooted in state law—which raise interesting but manageable Erie-related questions—from those involving only federal law. Federal courts in some instances have the power to impose the standards that I argue, in Parts I and II, fail to match the preponderance standard. But because adopting such standards involves substantive lawmaking, federal courts owe litigants and the public more transparency about that activity than they have provided. Rule of law values demand more in a system of government founded on predictable laws and observable lawmaking. Throughout this Article I assume for simplicity that the party offering the evidence is the plaintiff, as they have the burden of proof, and I consider how courts should handle either a defendant’s challenge to the legal sufficiency of estimation evidence (via Rule 50 or Rule 56 of the Federal Rules of Civil Procedure), or a defendant’s motion to exclude the plaintiff’s expert testimony about that evidence (via Rule 702 of the Federal Rules of Evidence). I explain why common practice by litigants and courts fails to satisfy the preponderance standard. And I use a combination of black-letter doctrine and mathematical statistics to justify a simple alternative—what I

9 Consider, for example, the statistical measure called the p-value. As one Reference Manual on Scientific Evidence chapter puts it, the p-value for measuring statistical significance “is the probability of extreme data given the null hypothesis. [It] is not the probability of the null hypothesis given extreme data.” David H. Kaye & David A. Freedman, Reference Guide on Statistics, in FEDERAL JUDICIAL CENTER, REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 250 (3d ed. 2011) [hereinafter “RMSE”]. Yet some courts and litigants confuse the p-value with the probability that one party or another has a stronger case, or with the probability that an observed estimate is the result of random chance rather than a causal relationship. See id. at 250-251, n.99 (collecting cases); see also David H. Kaye, Is Proof of Statistical Significance Relevant?, 61 WASH. L. REV. 1333, 1334 (1986) (arguing that “explicit hypothesis testing is poorly suited for courtroom use” and that “[s]tatements as to what results are or are not ‘statistically significant’ should be inadmissible.”).
call the fundamental default rule of estimation evidence. According to this rule, estimation evidence is legally sufficient and generally admissible when it points in the direction of the party offering the evidence.

To motivate my discussion, I use mass-tort litigation involving the prescription anti-cholesterol drug Lipitor. Thousands of women sued Pfizer, the drug’s producer, alleging that Lipitor caused them to develop type 2 diabetes.10 This was a multidistrict litigation (MDL), with all the complexities that arise in such cases. I focus here on one aspect of the MDL, which involved plaintiffs’ general causation expert, Dr. Sonal Singh. General causation in this context concerns whether Lipitor can cause type 2 diabetes in general, in the population, apart from whether it caused any particular plaintiff’s disease.11

Based on his expert report and deposition testimony, Dr. Singh was prepared to testify at trial that taking Lipitor at a 10mg dose causes onset of type 2 diabetes to at least some extent.12 After a Daubert hearing,13 the district court excluded Dr. Singh’s testimony regarding a 10mg dosage of Lipitor because Dr. Singh testified that the statistical estimation evidence underlying his testimony was not statistically significant at conventional levels of significance—i.e., a significance level of .05, which is frequently used in scholarly publication.14 With Dr. Singh excluded, the plaintiffs were left with no ability to prove general causation, leading to the grant of judgment as to many of the cases associated with the MDL.15 The Fourth Circuit recently affirmed the district court’s determination on this matter.16

These courts got this issue wrong.

When the preponderance standard applies, a juror is asked to determine whether the plaintiff’s or the defendant’s story is more probable in light of the evidence. But the analytical foundation of conventional hypothesis testing is entirely unrelated to this question. Instead, conventional hypothesis testing as conventionally practiced (i) asks how unlikely it would be to observe the evidence if the defendant’s position were right, and then (ii)

11 Id. at 638.
12 Id.
13 “Daubert hearing” is a colloquial term often used to describe a hearing to determine whether and/or to what extent an expert’s testimony should be admitted.
14 See id. at 638 (“[T]he district court excluded Dr. Singh’s opinion for each dose except 80 mg.”).
15 Id. at 631.
16 Id. at 649. For an amicus brief regarding the aspect of the appeal that involved the admissibility of Dr. Singh’s testimony with respect to the 10mg dose, see Brief for Amicus Curiae Jonah B. Gelbach in Support of Plaintiffs-Appellants, In Re Lipitor (Atorvastatin Calcium) Mktg., Sales Practice & Prod. Liab. Litig. (No. II) MDL 2502, 892 F.3d 624 (4th Cir. 2018), 2017 WL 1628475.
determines in favor of the defendant unless the answer is that the evidence would be *extremely* unlikely.

That’s a problem. Think about the preponderance standard as it is understood in cases with only non-statistical evidence. In those cases, the plaintiff can win even if jurors think the defendant is not all that unlikely to be right, given the evidence. If a juror determines that the plaintiff’s litigation position is more probable than the defendant’s, the juror is supposed to find for the plaintiff. And it is a hoary truism that “more probable” is satisfied by even very small differences in a juror’s belief. As the Vermont Supreme Court explained it recently, preponderance is satisfied even when the addition of evidence causes evenly balanced “scales [to] drop but a feather’s weight.”

Further, when a defendant moves for judgment—whether it is for summary judgment or judgment as a matter of law—the court is not supposed to grant it unless it’s the case that no reasonable juror could find for the plaintiff. And in making its determination, a court ruling on a motion for judgment is supposed to view the evidence in the light most favorable to the plaintiff.

Conventional hypothesis testing fails to reflect these principles as well.

This Article develops the primary alternative to conventional hypothesis testing for assessing estimation evidence: Bayesian hypothesis testing. Beginning in the late 1960s, Bayesian theory became one of the primary approaches legal scholars have used to understand evidence as a general matter (i.e., not just estimation evidence). Over roughly the last quarter century, Professor Ronald Allen and coauthors have mounted a frontal attack on the application of Bayesian theory in the law. Allen’s relative plausibility of competing explanations approach has gained much ground in the field, and

---

17 For an argument that courts should view the admissibility of expert evidence like that of non-expert evidence, see Frederick Schauer & Barbara Spellman, *Is Expert Evidence Really Different?*, 88 NOTRE DAME L. REV. 1, 4 (2013).


19 I will use the term “motion for judgment” and the like as a generic term that encompasses both sorts of motions; what matters is that the legal sufficiency of the record’s evidence is being tested. To the extent that my arguments apply in state courts, one should think of “motion for judgment” as also embracing motions for j.n.o.v. and so on.

20 See FED. R. CIV. P. 50(A)(i) (limiting judgment as a matter of law to those circumstances in which “the court finds that a reasonable jury would not have a legally sufficient evidentiary basis to find for the party on that issue”); FED. R. CIV. P. 56(A) (invoking “entitle[ment] to judgment as a matter of law” as a necessary condition for when summary judgment may be granted).


22 See, e.g., John Kaplan, *Decision Theory and the Factfinding Process*, 20 STAN. L. REV. 1065, 1083 (1968) (noting that Bayes’ Theorem was first used to explain the probabilities of events “only a few years ago”).

23 See supra notes 56-61 and accompanying text.
he and Professor Michael Pardo have recently argued that it has reached paradigm status. Still, as I explain in Section I.C, there are good reasons why even those in the relative plausibility camp should accept my approach when estimation evidence is at issue.

Section I.D presents the analytical core of this Article, which revolves around a hypothetical person called the “plaintiff’s most favorable juror.” This juror has neutral prior beliefs in the sense that before she sees any evidence, she believes that the plaintiff and defendant litigation positions are equally probable; in Bayesian terms, this juror places equal prior probability on the event that each party is right. Whether the juror would find for the plaintiff on the question related to the estimation evidence thus depends on whether the estimation evidence causes her to update her beliefs in favor of the plaintiff or the defendant. What makes the plaintiff’s most favorable juror notable is that estimation evidence pointing in favor of the plaintiff necessarily will cause her to update her beliefs in the direction of the plaintiff’s litigation position. And because she started out neutral, this juror’s beliefs after she sees the estimation evidence must favor the plaintiff. When the evidence points in the plaintiff’s direction, the plaintiff’s most favorable juror will think the plaintiff has satisfied the preponderance standard. That’s enough for legal sufficiency.

To illustrate what it means for estimation evidence to point in the plaintiff’s direction, consider again the Lipitor litigation. At issue was the true difference in the probability of being diagnosed with type 2 diabetes. Estimation evidence on this question came from a large randomized controlled trial, in the form of the difference in the observed frequency of type 2 diabetes onset among those who were randomly assigned to take Lipitor or a placebo. To say the estimation evidence points in the direction of the plaintiff is to say simply that the frequency of new diagnoses of type 2 diabetes was greater among those assigned to Lipitor than among those assigned the placebo. One doesn’t need to know the level of statistical

26 The results from the trial used in this Article may be found in Table 3 of Peter S. Sever et al., Prevention of Coronary and Stroke Events with Atorvastatin in Hypertensive Patients who have Average or Lower-than-Average Cholesterol Concentrations, in the Anglo-Scandinavian Cardiac Outcomes Trial—Lipid Lowering Arm (ASCOT-LLA): A Multicentre Randomised Controlled Trial, 361 LANCET 1149, 1153 (2003).
significance of such an estimated difference to know that a greater frequency among Lipitor takers points toward the plaintiff’s litigation position.\footnote{That said, the discussion below shows that the criterion of pointing toward the plaintiff’s litigation position turns out to be equivalent to conventional hypothesis testing with a significance level of 50%. See infra Part I.D.}

This is a radically more forgiving standard than what courts handling estimation evidence often do. But paying attention to the plaintiff’s most favorable juror exactly captures what it means to view the evidence in the light most favorable to the plaintiff. This means estimation evidence is legally sufficient when the plaintiff’s most favorable juror would find for the plaintiff on the question addressed by that evidence. Accordingly, a defendant’s motion for judgment may not be granted on the element involving estimation evidence when the plaintiff’s most favorable juror would find for the plaintiff as to that evidence.

Part I of the Article discusses the above relationship between hypothesis testing and legal sufficiency. It shows that the estimation evidence at issue in the Lipitor litigation discussed above was strong enough to be legally sufficient for the claim that a 10mg dose of Lipitor generally causes type 2 diabetes. It also discusses the dramatic extent to which conventional hypothesis testing at conventional significance levels raises the standard of proof facing plaintiffs.

Part II turns to two issues related to the admissibility of estimation evidence. The first is evidentiary relevance: when does estimation evidence meet Rule 401’s standard of making a fact of consequence more or less probable? As one would think, legally sufficient evidence generally meets that bar. Thus estimation evidence that points in the plaintiff’s direction generally should be admissible except when it runs afoul of Rule 403’s balancing test or Rule 702’s reliability requirements.

Many courts have excluded expert testimony about estimation evidence that doesn’t meet conventional hypothesis testing significance levels such as 5% (equivalently, the 95% confidence level).\footnote{Consider toxic tort case Brock v. Merrell Dow Pharm., Inc., 874 F.2d 307, 312 (5th Cir. 1989), amended by 884 F.2d 167, 167 (5th Cir. 1989), a Bendectin case in which the Fifth Circuit overturned a jury verdict for the plaintiff and issued the per curiam statement that “[W]e do not wish this case to stand as a bar to future Bendectin cases in the event that new and conclusive studies emerge which would give a jury a firmer basis on which to determine the issue of causation.” The epidemiology chapter in the Reference Manual on Scientific Evidence cites several subsequent cases in which trial courts excluded expert testimony for lack of statistical significance. Michael D. Green, D. Michal Freedman & Leon Gordis, Reference Guide on Epidemiology, in FED. JUDICIAL CTR., REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 578-79 n.85 (3d ed. 2011) [hereinafter RMSE].

In the discrimination field, consider the pre-Daubert case of Palmer v. Shultz, in which female employees of the State Department alleged the foreign service discriminated against them with respect to various personnel policies. 815 F.2d 84, 89 (D.C. Cir. 1987). The D.C. Circuit undertook a long discussion of conventional hypothesis testing methodology and concluded that statistical
significance level and reliability. As I explain in Section II.C, the level of statistical significance is best understood not as a measure of reliability, but rather as a measure of the strength of estimation evidence that may or may not be reliable for the question to which it is directed. My argument may be summarized using a simple question: If a study's results would be “unreliable” when they are statistically significant only at the 15% level, how can these results be “reliable” if they are significant at the 5% level? It's the same study, carried out by the same people, studying the same phenomena. To be clear, I am not arguing that the level of statistical significance is irrelevant to litigation. Estimation evidence that favors a proposition can do so with varying strength. But that is true of any kind of evidence.

Adopting the fundamental default rule proposed here would increase the set of estimation evidence that is admissible and legally sufficient. This would affect both the ease with which judges can manage complex litigation and its volume. Part III engages these issues. It also discusses the allocation of error risks across plaintiffs and defendants, explaining that the fundamental default rule best represents what the Supreme Court has repeatedly said: that error risks should be equally balanced.

That said, I do not claim that the preponderance standard is actually optimal from a policy point of view. Consider securities fraud litigation. Each day thousands of stocks trade on various exchanges in the U.S. So it is also true that thousands of stocks see positive returns and thousands see negative returns each day.29 Applying the properly understood preponderance of the evidence standard to securities litigation would probably induce a massive increase in the amount of such litigation, sometimes in some very marginal cases. But this is an argument in favor of developing and sensibly applying a higher standard that better reflects policy goals30—not an argument for pretending the law means something other than what courts otherwise say it does.

evidence would have to meet the standard for significance at least at the 5% level. Id. at 96; see also Dicker v. Allstate Life Ins. Co., No. 89-4984., 1997 WL 182290, at *42 (N.D. Ill. Apr. 9, 1997) (finding that because "Plaintiffs have not met the burden of showing a statistically significant difference in promotion rates . . . . Their claim . . . is dismissed." (emphasis added)).

In the securities litigation arena, consider Erica P. John Fund, Inc. v. Halliburton Co., 309 F.R.D. 251 (N.D. Tex. 2015). Following remand by the Fifth Circuit, the district court conducted a searching Daubert hearing and delivered a lengthy opinion whose conclusions as to class certification were mostly driven by its determinations as to the statistical significance of event study results at a confidence level of 95%. Id. at 262.

To be precise, what matters is not the daily stock return, but the daily excess return after accounting for other factors likely to be associated with stock-price movements. See Jill E. Fisch, Jonah B. Gelbach & Jonathan Klick, The Logic and Limits of Event Studies in Securities Fraud Litigation, 96 TEX. L. REV. 553, 574-75 (2018).

The EEOC's four-fifths rule, which federal courts have approved, could be seen in this way. Applying the fundamental default rule to employment discrimination cases would create some situations in which estimation evidence would virtually always be legally sufficient to establish
Part IV addresses some questions related to the substantive lawmaking powers of federal courts. As noted above, in this Part I argue that federal courts could legitimately adopt elevated proof standards for statistical estimation evidence where those courts have common law powers. But where they do that, they should do it more openly, and with more attention to whether Congress, or agencies, would be the better source for such decisions.

In sum, this Article radically rethinks the treatment of statistical estimation evidence in federal civil litigation. It proposes an approach that harmonizes legal standards and statistical concepts, replacing the arbitrary and elevated standards of conventional hypothesis testing with an approach that fits what the preponderance standard means when non-statistical evidence is at issue. And it shows how courts might in some cases legitimately move away from that standard where doing so makes policy sense, taking seriously the limited but real common law powers of federal courts.  

I. BAYESIAN HYPOTHESIS TESTING APPLIED TO LEGAL SUFFICIENCY AND THE PREPONDERANCE STANDARD YIELDS A FUNDAMENTAL DEFAULT RULE

This Part begins in Section I.A with a quick discussion of the basic principles of conventional hypothesis testing, as applied to the Lipitor MDL. Section I.B then introduces the basic framework of Bayesian hypothesis testing. Section I.C engages the general critiques of Bayesian approaches that evidence scholars have lodged, arguing that in the estimation evidence context, Bayesian hypothesis testing is an acceptable framework. Section I.D then introduces the hypothetical plaintiff’s most favorable juror and explains how seeing estimation evidence through this juror’s eyes corresponds to viewing the evidence in the light most favorable to the plaintiff. The fundamental default rule follows directly. Section I.E then revisits the conventional hypothesis testing approach, explaining that its use effectively raises the legal standard far above the preponderance level.

disparate impact for at least one group. Although the four-fifths rule is often seen as plaintiff-friendly, it would protect employers from that unreasonable outcome. Another setting in which agency rulemaking could be used to set the standard of proof for litigation is securities fraud litigation; for an argument in favor, see Jill E. Fisch & Jonah B. Gelbach, Power and Statistical Significance in Securities Fraud Litigation, HARV. BUS. L. REV. (forthcoming 2021) (draft on file with author).

31 Much of what I argue applies in state courts as well, except insofar as states have chosen different evidentiary standards of proof. I focus on federal courts in this paper only for length reasons.
A. A Brief Description of Conventional Hypothesis Testing

Attention in conventional hypothesis testing centers on the question of how unlikely it would be to observe the actually-observed data in the counterfactual event that the variable of interest had a particular value. This particular value is known as the “null value,” and the hypothesis that the variable has that value is known as the “null hypothesis.” In the Lipitor litigation, the null hypothesis of interest is that Lipitor has no effect on type 2 diabetes incidence, so the null value is zero.

The relevant data from the Lipitor litigation come from the ASCOT-LLA randomized trial. The data show that type 2 diabetes incidence was 3.0% among those who received Lipitor in 10mg doses and was 2.6% for the control group that received a placebo. The difference in incidence was thus 0.4 percentage points — roughly a 15% increase over the no-Lipitor incidence. On its face, this evidence indicates that Lipitor is associated with an increase in type 2 diabetes — not a good thing, and one that might bring tort liability.

An expert using conventional hypothesis testing won't stop with what I just described as the facial conclusion, though: she will ask whether this difference is enough greater than zero to be "statistically significantly different from zero"; this is sometimes shortened to just “statistically significant.” To answer that question, the difference in type 2 diabetes incidence can be converted into a t-statistic. When the null hypothesis is that the true incidence is zero, the t-statistic is the ratio of the difference in incidence rates — here, 0.4 percentage points — to the estimated standard error of this difference. In the Lipitor case, the value of the resulting t-statistic is 1.2.

\[ t = \frac{\text{Difference in Incidence}}{\text{Estimated Standard Error}} \]

For a discussion of t-statistics, see Kaye & Freedman, supra note 9, at 282 (explaining that a t-statistic is [an] estimated value divided by its standard error and discussing the distribution of t-statistics under the null hypothesis).

If the null hypothesis were that the true incidence is zero, then an expert using conventional hypothesis testing would first subtract that null value from the estimated difference in incidence, and only then divide by the estimated standard error.

This calculation follows from the following facts. First, a consistent estimator of the variance of an estimated proportion, $\hat{q}$, is given by $\hat{q}(1 - \hat{q})/n$, where $n$ is the number of observations from which the estimate is computed. Second, a consistent estimator of the variance of the difference between one estimated proportion and another is the sum of their estimated variances (when they are independent, as is the case here). Third, a consistent estimator of the standard error of the difference between two estimated proportions is the square-root of any consistent estimator of the variance. Fourth, in the ASCOT-LLA trial, there were 5,168 subjects assigned to the Lipitor treatment, of whom 154 were determined to have developed type 2 diabetes thereafter. Sever et al., supra note 26, at 1151, 1153. Fifth, there were 5,137 subjects assigned to receive the placebo, of whom 134 were determined to have developed type 2 diabetes thereafter. Id. Plugging these numbers into...
The ASCOT-LLA trial had more than 10,000 subjects in the treatment and placebo groups combined. That is helpful for various reasons, but for our purposes the key one is that when the sample size is large, the \( t \)-statistic has an approximately normal distribution with variance equal to 1. If the null hypothesis were true, then the true type 2 diabetes incidence would be the same in the treatment and placebo groups. Thus, the \( t \)-statistic has an average value of zero under the null hypothesis.

So when the null hypothesis is true, the \( t \)-statistic has a normal distribution with mean 0 and variance 1. The percentiles of this distribution are known. Therefore, one can determine the “critical value” such that a random draw from this distribution will exceed that critical value only 5% of the time; this value is roughly 1.65. An expert using conventional hypothesis testing at the 5% significance level will “reject” the null hypothesis whenever the observed \( t \)-statistic exceeds 1.65, and not otherwise. By design, this test procedure will cause an expert to reject the null hypothesis when it is actually true only 5% of the time.

When the statistical estimation evidence leads the expert to reject the null hypothesis at a given significance level, one says that an estimate is statistically significant at that level, e.g., “statistically significant at the 5% (or 0.05) level.”

the formulas just described and dividing the difference in type 2 diabetes incidence by the result yields a \( t \)-statistic of 1.2.

To be more precise, this is how an expert using a one-sided conventional hypothesis test would behave. The other type of composite hypothesis test is the two-sided test. With a two-sided test, the expert rejects the null hypothesis whenever the observed \( t \)-statistic is far from zero in either direction—i.e., large negative values are treated as evidence against the null hypothesis (this approach can be implemented by rejecting the null hypothesis whenever the absolute value of the \( t \)-statistic exceeds the critical value). At least one court has expressed the belief that two-sided tests are preferable, even though that may make no sense in the litigation in question. See, e.g., E.E.O.C. v. Fed. Reserve Bank of Richmond, 698 F.2d 633, 660 (4th Cir. 1983), rev’d sub nom, Cooper v. Fed. Reserve Bank of Richmond, 467 U.S. 867, 104 S. Ct. 2794, 81 L. Ed. 2d 718 (1984) (“[W]e are not persuaded that it is at all proper to use a test such as the ‘one-tail’ test which all opinion finds to be skewed in favor of plaintiffs in discrimination cases ...”); cf. In re Novatel Wireless Sec. Litig., 910 F. Supp. 2d 1209, 1216 (S.D. Cal. 2012), order vacated on reconsideration, No. 08CV1689 AJB RBB, 2013 WL 494361 (S.D. Cal. Feb. 7, 2013), on reconsideration in part, No. 08CV1689 AJB (RBB), 2013 WL 12144449 (S.D. Cal. Mar. 6, 2013), vacated, No. 08CV1689 AJB (RBB), 2013 WL 12144450 (S.D. Cal. Oct. 25, 2013) (stating that because experts in the case disagreed on the question, the one-sided/two-sided issue "is not an issue of admissibility, but rather of probative value to be addressed at trial").

A large negative value of the observed \( t \)-statistic in the Lipitor litigation, for example, would suggest that Lipitor actually reduced type 2 diabetes incidence, in which case the plaintiffs would have no entitlement to any remedy. My argument works either way, although if two-sided testing is used then the precise numbers I discuss below would change somewhat.

Another way to say this is that the probability of a false positive—rejecting the null hypothesis when it is actually true, also known as the Type I error rate—is the same as the significance level. A significance level greater than 5%, such as 10%, is more forgiving and will entail a lower critical value; a more demanding significance level, such as 1%, will have a greater critical value.
An equivalent statement for current purposes is that the estimate is “statistically significant at the 95% confidence level,” because the “confidence level” is defined to be 100% times one minus the significance level.

An equivalent way to understand conventional hypothesis testing involves the concept of $p$-values. For present purposes, the $p$-value is the probability that we would observe a value at least as extreme as the actually observed $t$-statistic’s value, if we took a new random draw from the normal distribution with mean 0 and variance 1 (because this is the distribution of the $t$-statistic that holds under the null hypothesis). It can be shown that a policy of rejecting the null hypothesis whenever the observed $t$-statistic exceeds the critical value described above is equivalent to a policy of rejecting the null hypothesis whenever the $p$-value is less than the significance level. Thus, statistical significance at a given significance level or confidence level and the $p$-value are intimately related concepts.

As applied to civil litigation, the justification for using conventional hypothesis testing must be based on the claim that in a legal sufficiency challenge, we should default to the defendant’s position unless the observed data provide evidence that would be very unlikely to occur if the defendant’s position were true.

In scholarly and other research contexts, failure to reject the null hypothesis does not generally mean that one accepts the truth of the null hypothesis. It could be that the null hypothesis is false, but the observed $t$-statistic fell short of the critical value due to random variation in the sample that was selected for study. Scholars and other researchers typically have the luxury of reserving judgment until more evidence trots along.

But in the litigation context, this is problematic. For one thing, the approach considers only the behavior of the $t$-statistic when the defendant’s position is correct; it pays no heed to the comparative likelihood of the observed data when the plaintiff’s position is correct. A value of the $t$-statistic could be not-all-that-unlikely when the defendant is right, but extremely likely otherwise. It is inconsistent with the nature of legal factfinding to ignore half of the picture in this way. For example, in a tort case that did not involve statistical evidence, fact finders would be expected to pay attention not only to how likely the observed evidence would be if the defendant had taken due care, but also to the likelihood of the evidence if the plaintiff’s allegation of carelessness were true.

A second problem with conventional hypothesis testing is that the choice of significance level both (i) determines the result of the test for any given $t$-statistic and (ii) is entirely arbitrary. The significance level is determinative because for any value of the observed $t$-statistic, there exist sufficiently

---

41 Such a false negative outcome is known as a Type II error.
relaxed significance levels so that the expert would reject the null hypothesis and sufficiently demanding ones so that the expert would fail to reject. It is arbitrary because no reason but decades of conventional practice explains applied statisticians' typical use of the 5% significance level. Indeed, that standard seems to have taken hold merely because of an offhand remark by influential statistician R. A. Fisher nearly a century ago:

The value for which [the significance level is] .05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant.

Moreover, as Judge Posner put it, “The five percent test is a convention employed in academic research,” but it is “not one to which the research community adheres rigidly.”

Thus, conventional hypothesis testing bears little resemblance to the preponderance standard as courts describe that standard; it is an apple to the preponderant orange. Still, as I discuss in Section I.E, it is possible to say some useful things about the standard of evidence implied by various choices of conventional hypothesis testing’s significance level. Before we get there, though, we must discuss Bayesian hypothesis testing.

B. The Bayesian Hypothesis Testing Approach

Bayesian theory is about how new information should cause a person to update her beliefs about the probability that a proposition is true. Suppose the proposition is that the US soccer team will win a particular soccer game. New information comes along, in the form of the score at halftime. Before the game starts, our observer thinks the US has a 20 percent chance to win. In Bayesian terms, this is the prior probability in favor of a U.S. win. The prior probability of an event is also known as the event’s unconditional probability.

It is often helpful to work with odds rather than probabilities. The odds in favor of a U.S. win are the ratio of the probability that the U.S. wins to

42 Another way to put this is that as the significance level varies from 0 to 1, the range of critical values varies from $-\infty$ to $\infty$.


45 Readers interested in more about Bayesian theory, situated in the legal context, might consult MICHAEL O. FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS 75-81 (2001) or Kaye & Freedman, supra note 9, at 273-75.
the probability that the U.S. does not win. Because our observer thinks the prior probability is 20 percent that the U.S. will win and 80 percent that the U.S. will not win, her odds in favor of a U.S. win are one-fourth (20 divided by 80). Notice that a probability below 50 percent is the same as odds below 1, and a probability above 50 percent is the same as odds above 1.

Now suppose our observer finds out the U.S. has a lead at halftime. What should our observer think about the odds in favor of a U.S. win now? To ask this question is to ask what should be the odds in favor of a U.S. win given the information that the U.S. leads at the half. These odds are known as the posterior odds, or, equivalently, the conditional odds.

Bayes’s Theorem tells us how the observer can update her beliefs in a way that is consistent with the conventional laws of probability. Suppose for a moment there is only one score by which the U.S. could win, and only one score by which it could fail to win. In statistical lingo, this means that both outcomes involve simple hypotheses. Then a foundational mathematical fact known as Bayes’s Theorem says that the conditional odds equal the product of the unconditional odds and the likelihood ratio:

\[
\text{Conditional Odds} = \text{Unconditional Odds} \times \text{Likelihood Ratio} \quad (1)
\]

The likelihood ratio term is the key to understanding how information causes Bayesian decisionmakers to change their beliefs by updating from unconditional to conditional odds. The numerator of this ratio is the probability that the U.S. team would be expected to have a halftime lead in those games the team will go on to win. This probability is also the likelihood of a U.S. win given that the team has a halftime lead. The denominator of the likelihood ratio in our example is the probability the U.S. will have a halftime lead in those games the team goes on to lose; it is also known as the likelihood that the U.S. will not win, given the halftime lead.

---

46 This representation of the theorem, using “posterior” and “prior” in place of “conditional” and “unconditional,” appears in, e.g., David H. Kaye & George Sensabaugh, Reference Guide on DNA Identification Evidence, in REFERENCE MANUAL ON SCIENTIFIC EVIDENCE, supra note 9, at 173. The discussion in the text above glosses over some important details in which there are multiple ways in which F could occur. I discuss these issues below.

47 The probability in question is the probability of an outcome (the U.S. has a lead at the half) given the probability distribution induced by a condition (that the U.S. will go on to win). This fits the definition of likelihood given by IAN HACKING, LOGIC OF STATISTICAL INFERENCE 54 (Cambridge Univ. Press 2016).

48 Notice that the likelihood of the true state of the world given the observed data equals the probability of observing data given the true state of the world. This is a definitional property of likelihood.
The likelihood ratio exceeds 1 when the likelihood of a U.S. win given the information that the U.S. has a halftime lead exceeds the likelihood of a U.S. loss given that information. Then equation (1) implies that the conditional odds in favor of a U.S. win are greater than the unconditional odds. If we knew the specific probabilities in question, we could use them to arrive at the numeric value of the conditional odds in favor of a U.S. win. But we don't need to know particular numeric values to grasp the main point of this discussion, which is that new information causes a Bayesian observer to increase her odds when the likelihood ratio exceeds 1.

So far I've assumed issues related to the fact that the likelihood of different winning scores might vary. In the real world, there are multiple scores by which a team can win or lose a soccer game. An observer may place different prior probabilities on the event that a team wins by different scores. For example, before the game starts, anyone who understands soccer will believe that the U.S. team is more likely to win by one goal than by forty. This means we must take account of what statisticians call composite hypotheses—overall hypotheses that are composed of multiple simple hypotheses. To accommodate composite hypotheses, we modify equation (1) as follows:

\[
\text{Conditional Odds} = \frac{\text{Overall Unconditional Odds}}{\text{Average Likelihood of } H_i}{\text{Average Likelihood of } H_o} \times \frac{\text{Average Likelihood of } H_i}{\text{Average Likelihood of } H_o}\quad (2)
\]

The qualifier “Overall” indicates that the prior, or unconditional, beliefs in question account for the probability the observer places on all the ways the U.S. team could win, and the probability she places on all the ways the team could lose.

Another difference between equations (1) and (2) is that instead of a likelihood ratio, the final term in the latter equation is the ratio of average likelihoods. The numerator is the average likelihood in favor of a U.S. team win, which is denoted “\(H_i\)”—short for the hypothesis that the U.S. team will win. This average likelihood is the average value of the likelihood function across all ways the U.S. team might win. In computing the average, the Bayesian will use her prior beliefs about each possible score to determine how much weight each score should get in the average likelihood. The

---

49 Relatedly, according to Hacking’s law of likelihood, evidence supports one hypothesis over another if the likelihood of the first hypothesis is greater than the second. *Id.*
denominator is constructed analogously—it’s the average of the likelihood that the team fails to win in each possible way.

The ratio of average likelihoods is known as the Bayes factor in favor of the numerator hypothesis. It plays the same role in composite-hypothesis problems that the likelihood ratio plays in the simple-hypothesis setting: to get to conditional odds, we multiply overall unconditional odds by the Bayes factor. So an observer’s beliefs move toward a hypothesis when the Bayes factor exceeds 1. Thus, a U.S. lead at halftime increases our Bayesian observer’s odds in favor of a U.S. win if the average likelihood of a U.S. win, given a halftime lead, exceeds the average likelihood of a U.S. non-win.

These ideas have a direct application to litigation involving estimation evidence, as the Lipitor litigation example nicely illustrates. Rather than “the U.S. team will win,” think of the numerator hypothesis as “Lipitor increases type 2 diabetes incidence.” And think of the denominator hypothesis as “Lipitor does not increase type 2 diabetes incidence.” Then the conditional odds become the odds in favor of the plaintiff’s litigation position once the jury finds out about the estimation evidence. Because odds exceeding 1 and the numerator probability exceeding the denominator probability are the same thing, a juror will favor the plaintiff’s litigation position when her conditional odds exceed 1.

Take the unconditional odds first. Because these involve nothing but jurors’ prior beliefs, unconditional odds are inherently subjective. That seems to make it intractable to use Bayesian hypothesis testing. Some legal scholars contend with this aspect of the problem by arguing that whatever jurors’ actual prior beliefs, for fairness reasons we should act as if jurors have equal prior beliefs in federal civil litigation. Alternatively, prior odds of 1 naturally arise as the result of applying the so-called minimax criterion in evidence models. Without rejecting either of these positions, I take a weaker position: in general, a reasonable juror could have overall unconditional odds equal to 1. This is all that is needed for purposes of analyzing motions for judgment, which are tied not to any particular juror’s actual beliefs, but rather to the objective reasonable juror standard. It is difficult to imagine a court excluding

50 See, e.g., Edward K. Cheng, Reconceptualizing the Burden of Proof, 122 YALE L.J. 1254, 1267 n. 24, 1268, 1275 (2013) (arguing that, although it may be inaccurate “as a matter of inference” to set prior odds to 1, it is reflective of what courts do in practice); Posner, supra note 8, at 1514 (“Ideally we want the trier of fact to work from prior odds of 1 to 1 that the plaintiff or prosecutor has a meritorious case. A substantial departure from this position, in either direction, marks the trier of fact as biased.”).

51 Edward K. Cheng & Michael S. Pardo, Accuracy, Optimality and the Preponderance Standard, 14 L. PROB. & RISK 193, 207 (2015) (describing result from a particular Gaussian model of noise that the minimax solution in that model yields the functional equivalent of prior odds of 1).
for cause a juror who walks into the courtroom with no lean in favor of either side; imagine how the eventual appeal would go if a judge excused a juror for not leaning heavily enough toward one side.

Accordingly, except where explicitly noted, I will assume a reasonable juror could have overall unconditional odds equal to 1 for the balance of this Article. This means that a reasonable juror could have conditional odds equal to her Bayes factor. Thus, for purposes of motions for judgment, the Bayesian hypothesis testing problem reduces to one of analyzing the Bayes factor. I do that analysis in Section I.D in connection with the plaintiff’s most favorable juror. First, though, I address the question of whether the Bayesian hypothesis testing framework is appropriate to begin with.

C. Why Bayesian Hypothesis Testing is the Right Framework
For Legal Sufficiency

Preponderance is the most common standard of proof in civil litigation. This standard is often expressed as the requirement that it be “more probable than not” that the plaintiff has proved her case. As the most recent version of McCormick’s treatise elaborates:

The most acceptable meaning to be given to the expression, proof by a preponderance, seems to be proof which leads the jury to find that the existence of the contested fact is more probable than its nonexistence. Thus the preponderance of evidence becomes the trier’s belief in the preponderance of probability.

This language maps directly onto notions of Bayesian posterior probability, and much theory in evidence law over the several decades following the late 1960s was founded on that idea. The Bayesian approach proved controversial, and numerous scholars reject the idea that the preponderance standard can or should be conceived in terms of probabilities. The leading alternative is what Professor Ronald Allen and

---

52 I do discuss the possibility of cases in which the court believes such neutral priors are unreasonable; see infra Section I.E.
53 KENNETH S. BROUN ET AL., MCCORMICK ON EVIDENCE 661 (Thomas Reuters 7th ed. 2013) (footnotes omitted). See also id. at 661 n.13. (citing cases).
54 According to Michael Risinger, the opening salvo in the “Bayes Wars” was fired by John Kaplan in a 1968 law review article. See Michael Risinger, Introduction to Roger C. Park et al., Bayes Wars Redivivus—An Exchange, 8 INT’L COMMENT. ON EVIDENCE 1, 1 (2010) (referring to the “watershed article” by John Kaplan, Decision Theory and the Factfinding Process, 20 STAN. L. REV. 1065 (1968)).
coauthors describe as relative plausibility of competing explanations.\textsuperscript{56} Professor Allen and Professor Michael Pardo write:

The two primary differences between our account and the more conventional [Bayesian] probabilistic accounts are, first, the criteria that [are] central to the fact-finding process (explanatory vs. probabilistic), and, second, whether the proof process is characterized as comparative or not. Unlike the conventional probabilistic accounts, the explanatory account is inherently comparative—whether an explanation satisfies the standard will depend on the strength of the possible explanations supporting each side. Under the “preponderance of the evidence” standard, fact-finders determine whether the best of the available explanations favors the plaintiff or the defendant. The best available explanation will favor the plaintiff if it includes all of the legal elements of the plaintiff’s claim; it will favor the defendant when it fails to include one [or] more elements.\textsuperscript{57}

As this quotation hints, one basis for Professor Allen’s critique of the Bayesian approach is the difficulty of knowing how to place probabilistic numbers on propositions related to the elements plaintiffs must prove.\textsuperscript{58} The comparative alternative that Professor Allen proposes is what is known in the philosophy of science literature as “inference to the best.”\textsuperscript{59} According to this approach:

\begin{footnotesize}
\begin{enumerate}


\item See PETER LIPTON, INFERENCE TO THE BEST EXPLANATION (1991) for a treatment in the philosophy of science literature. As to Professor Allen’s application of the concept in evidence
\end{enumerate}
\end{footnotesize}
At trial, the parties . . . offer competing versions of events that, if true, would explain the evidence presented at trial. Parties with the burdens of proof . . . offer versions of events that include the formal elements that make up the particular claims or defenses; opposing parties offer versions of events that fail to include one or more of the formal elements . . . .

At the decision stage in civil cases where the burden of persuasion is a preponderance of the evidence, proof depends on whether the best explanation of the evidence favors the plaintiff or the defendant. Fact finders decide based on the relative plausibility of the versions of events put forth by the parties, and possibly additional ones constructed by themselves. Fact finders infer the most plausible explanation as the actual explanation and find for the party that the substantive law supports based on this accepted version.60

The inference-to-the-best approach as advocated by Professor Allen rejects the idea that individual jurors compute the conditional probability that one side or the other is right, by averaging likelihood values over their prior beliefs. Instead, inference to the best assumes that each juror determines relative plausibility by “constructing narrative versions of events to account for the evidence presented at trial based on criteria such as coherence, completeness and uniqueness.”61 Then each juror compares the best account in favor of each side and chooses the one that seems more plausible.

However convincing one finds that argument, there are good reasons why my analysis in Part I, founded on Bayesian reasoning, still applies. First, even if the claim is right that most jurors don’t actually act as Bayesians, that is quite a different thing from establishing that a juror who did would be unreasonable. After all, where it is feasible to use, Bayes’s theorem provides a coherent, consistent, and rational way to update beliefs given new information.

Second, Bayes’s theorem is feasible to use when quantitative estimation evidence is at issue. As I demonstrate in Section I.D, the assessment of statistical estimation evidence is well suited to the use of Bayesian reasoning. Professors Allen and Leiter surely are right that in most cases it is so difficult to calculate the necessary likelihood values that as a general matter, the method is at best heuristic in its value. But by construction, in the estimation evidence context, we have quantitative evidence and need only evaluate it in line with Bayesian reasoning. So Allen and Leiter’s argument that “ought implies can”62 is undaunting in this important special case.

60 Allen, Burdens of Proof, supra note 58, at 216.
61 Id.
Third, Bayesian approaches can just as well be defined in terms of odds as probability levels, as the discussion earlier in this Section shows. Professor Edward Cheng not long ago made the same point, although he did not focus on the problem of composite hypotheses. Professor Sean Sullivan has argued compellingly that the inference-to-the-best approach can be accommodated in a likelihood-based framework with composite hypotheses. Indeed, my analysis of the plaintiff’s most favorable juror in Section I.D bears important similarities to Professor Sullivan’s. Thus Professor Sullivan’s work implies that likelihood-based reasoning is not an alternative, but actually a way to implement the comparative and inference-to-the-best aspects of Professor Allen’s relative plausibility approach.

Finally, the language that courts use in discussing the preponderance standard is so tied to probability talk that it would be odd to suggest there is something wrong with using probabilistic reasoning when probability can be meaningfully quantified. Courts and litigants regularly use the words “probability” and “probable” in discussing preponderance. When they don’t, they typically use “likelihood” or “likely” instead.

Susan Haack argues to the contrary, pointing out that this language has both an epistemic and a mathematical sense. For example, if I were to say, “law review editors will probably hate my Article,” few people would puzzle

---


65 The numerator likelihood value for the plaintiff’s most favorable juror is the best possible value for the plaintiff, which is what Sullivan proposes to use. Sullivan’s denominator is the best possible value for the defendant. My analysis does not require that the plaintiff’s most favorable juror use this value, because my approach requires only that I bound the Bayes factor for the plaintiff’s most favorable juror below rather than determine its exact value. Sullivan points out that because his proposed ratio compares the best likelihood value for the plaintiff’s case to the best likelihood value for the defendant’s case, his ratio can be viewed as implementing the inference-to-the-best approach in a likelihood framework. In the statistics literature, Sullivan’s proposed ratio is known as the generalized likelihood ratio, and recent work has demonstrated that it can usefully be viewed as capturing key epistemic features of the inference-to-the-best philosophical approach. See David R. Bickel, The Strength of Statistical Evidence For Composite Hypotheses: Inference to the Best Explanation, 22 STATISTICA SINICA 1147, 1156 (2012) (noting that the “inference to the best explanation stipulates that the simple hypothesis of highest explanatory power be inferred”); see also Zhiwei Zhang & Bo Zhang, A Likelihood Paradigm for Clinical Trials, 7 J. STAT. THEORY & PRAC. 157, 160 (2013) (providing an alternative axiomatization that yields the generalized likelihood ratio as a metric of the strength of evidence).

66 Here it useful to point out the equivalence of adopting the likelihoodist view of the preponderance standard given my approach to the plaintiff’s most favorable juror below, and adopting the Bayesian view, given overall prior odds equal to 1. The two views yield the same inferences for any data.

67 Haack, supra note 7, at 47-48.
over where exactly in the interval \([0,1]\) my belief about the implied probability lies; people would just think I had low confidence, in the colloquial sense of “confidence,” in my Article’s congeniality to student law review editors.\(^68\) On the other hand, if I were to say “it is more probable that a heads-biased coin will come up heads than tails,” most people would think I meant that the mathematical chance of a heads exceeds 0.5. Haack’s epistemic notion of probability is that it conveys degrees of “warrant,” whereas mathematical probability conveys something potentially different.\(^69\) But the preponderance standard is virtually never expressed in terms of “warrant,” “credibility,” “coherence,” or words other than “probability” or “likelihood.” At least when quantifiable evidence is involved, we should be open to understanding the preponderance standard in terms of mathematical probability concepts.

Indeed, I suspect that much of the opposition to Bayesian decision theory in evidence law is traceable to the fact that it is difficult to take mathematical probability seriously as a practical means of summarizing evidence that cannot be quantified. To be sure, that is most evidence. No one really knows how to calculate the actual quantitative likelihood that there would be ice on a sidewalk—either in the absence or the presence of due care taken by the defendant.\(^70\) With non-statistical evidence, the Bayesian approach is thus at most a useful heuristic model. But the present context is limited to elements of a case involving statistical estimation evidence. In this Article, I develop a theory specific to such evidence, and only as to those issues in a case at which such evidence is directed.

For these purposes, Bayesian posterior probability works quite well. Estimation evidence by its very nature fits perfectly within the Bayesian paradigm, because its associated likelihoods are naturally quantifiable. Indeed, much of probability theory was developed to make sense of the behavior of (approximately) normal statistical estimators on which I focus in this Article. So it is quite natural to conceive of statistical estimates in terms

---

\(^{68}\) Cf. SARAH MOSS, PROBABILISTIC KNOWLEDGE 50-53 (2018) (assuming that “credences” of non-quantitatively describable events can be quantitatively expressed). I note that Moss addresses legal proof in one chapter of her book in a way that may be consistent with my Bayesian approach; a detailed treatment of her discussion is beyond the scope of the present Article.

\(^{69}\) Id.

\(^{70}\) As Leiter and Allen write:

The first worry [about using the Bayesian model] is computational complexity, which raises the specter of violating “ought implies can.” A huge and complicated data set is involved at most trials, even most “simple” trials. No computer, let alone any human, has the computational capacity to do the calculations necessary for the operation of Bayes’ Theorem in a reasonable amount of time.

Leiter & Allen, supra note 56, at 1507.
of probability theory—and thus in terms of Bayesian probability theory—in the litigation context.

None of that proves mathematical probability, rather than Haack’s notion of “warrant,” better captures the law’s object in phrases like “more probable than not.” To be clear, Haack’s position is that nothing could prove such a proposition. But it is difficult to conceive of a conceptual framework that would better capture the spirit of the law’s language when estimation evidence is at issue.

To put it differently, when the evidence itself is appropriately described in mathematically probabilistic terms, what could provide a better understanding of “warrant” than mathematical probability itself? To disagree would require accepting that there are warranted propositions whose mathematical odds are less than 1, and also that there are unwarranted propositions whose mathematical odds exceed 1. That seems like a strange notion of “warrant.” My position that we should use Bayesian posterior probability to understand what “probable” means in legal standards when using statistical estimation evidence is aptly captured by the epigram opening Haack’s Legal Probabilism chapter—a quote attributed to Bertrand Russell: “[T]he rational [person], who attaches to each proposition the right degree of credibility, will be guided by the mathematical theory of probability when it is applicable.”

Finally, a conceptual strength of my approach is that although it is built on the Bayesian foundation, its account of black-letter law is sufficient to eliminate prior beliefs from consideration with respect to the procedural milestones of litigation on which I focus here. My argument is not a result of claims about the relative costs of Type I and Type II errors (other than insofar as those claims have led courts to adopt the preponderance standard in the first place). Nor does it require all or most jurors to be self-conscious.

71 Suppose one collected together all the warranted and unwarranted positions of this type. If the number is small, then there is little harm in ignoring the few propositions that have this type. If instead the number is large, then applying an appropriate law of large numbers would yield the result that a greater share of warranted propositions are false than true, and a greater share of unwarranted propositions are true than false. And nothing in this argument requires anyone to know which propositions are true or false; its result follows from its premises regardless of what is observable. The only way out is to assume that probabilities of truth simply don’t exist.

72 Haack, supra note 7, at 47 (quoting BERTRAND RUSSELL, HUMAN KNOWLEDGE, ITS SCOPE AND LIMITS 381 (1948)).

73 In the present context, a Type I error entails concluding that the plaintiff’s claim is true when in fact it is the defendant’s position that’s correct; a Type II error entails switching the parties in that statement—so that the error is concluding the defendant is right when in fact the plaintiff is.

74 See, e.g., Addington v. Texas, 441 U.S. 418, 423 (1979) (noting that “[s]ince society has a minimal concern with the outcome of such private suits, plaintiff’s burden of proof is a mere preponderance of the evidence. The litigants thus share the risk of error in roughly equal fashion”); see also Herman & MacLean v. Huddleston, 459 U.S. 375, 390 (1983) (citing Addington and noting
Bayesians, or to be instructed on the intricacies of Bayesian reasoning. Rather, it is the result of reasoning anchored in black-letter procedural law—specifically, about what beliefs *could* be held by a particular set of juries that I have argued would be reasonable in general.

D. The Fundamental Default Rule Results from Viewing the Evidence in the Light Most Favorable to the Non-Movant

In this Section I make the case for the fundamental default rule that statistical estimation evidence is legally sufficient under the preponderance standard whenever it points in favor of the party proffering it. For example, consider a randomized controlled trial of a drug’s effects on incidence of a disease. If the disease is more frequently observed in the treatment group than the control group, then according to the fundamental default rule the evidence is legally sufficient to establish that the drug generally causes the disease. In this Section, I develop the argument for this rule, which is much more lenient than using statistical significance at conventional significance levels.

My argument turns on what it means to view the evidence in the light most favorable to the plaintiff: to imagine a juror whose view of the evidence leads the juror to place the highest reasonable posterior probability on the plaintiff’s litigation position. As I develop the argument, it will be useful to refer to the data from the ASCOT-LLA study discussed above. The discussion in Section I.B showed that the likelihood ratio—the ratio of values of the likelihood function for two hypotheses of interest—determines how a Bayesian person updates beliefs from priors to posteriors. For the ASCOT-LLA data, Figure 1 plots the likelihood function under the assumption that the *-statistic has a normal distribution. Under that assumption, the likelihood function is the same as a normal probability distribution; *In re Winship*, 397 U.S. 358, 371–72 (1970) (Harlan, J., concurring) (describing how the preponderance standard is appropriate because “we view it as no more serious in general for there to be an erroneous verdict in the defendant’s favor than for there to be an erroneous verdict in the plaintiff’s favor”).

Cf. Michael O. Finkelstein & William B. Fairley, *A Bayesian Approach to Identification Evidence*, 83 HARV. L. REV. 489, 502 (1970) (“An expert witness could explain to jurors that their view of the statistical evidence should depend on their view of the other evidence. He might then suggest a range of hypothetical unconditional probabilities, specifying the posterior probability associated with each unconditional. Each juror could then pick the unconditional estimate that most closely matched his own view of the evidence.”).

It is common knowledge among statisticians that in a large sample, this distributional assumption can be justified by appeal to central limit theory. That theory establishes that in a large enough sample, the distribution of a sample proportion (such as type 2 diabetes incidence) is practically indistinguishable from the normal distribution. See Kaye & Freedman, *supra* note 9, at 276-278 and citations therein for more on the central limit theorem.
density function (familiarly known as the bell curve) with mean equal to 1.2 (the observed value of the \( t \)-statistic) and variance equal to 1.77 This is useful because that function’s precise form is known, which allows us to investigate its properties with particularity using Figure 1.

The figure’s horizontal axis measures different possible values for the true mean of the \( t \)-statistic’s distribution. The vertical axis tells us the likelihood of the x-axis value for the \( t \)-statistic’s mean. Recall that the \( t \)-statistic is the difference in type 2 diabetes incidence divided by the estimated standard error of this difference. Thus each value on the x-axis corresponds to a particular true impact of Lipitor on type 2 diabetes incidence, expressed in units of standard error.

77 The likelihood function is formally defined in terms of the conditional density of the full set of observed data, not just some estimator of interest. However, often in litigation the only feature of the data that is of interest is an estimator, or its associated \( t \)-statistic, as with the ASCOT-LLA study and the Lipitor litigation. When the estimator in question is the maximum likelihood estimator (MLE), as is also true with the ASCOT-LLA data, it has been demonstrated that large-sample results related to the likelihood are functionally equivalent to those that would be obtained if the full-data likelihood function were used. Lemma 3.1 of Tsung-Shan Tsou & Richard M. Royall, *Robust Likelihoods*, 90 J. AM. STAT. ASS’N. 316, 319 (1995).

In some cases, it’s necessary to use an estimator other than the MLE. For example, antitrust litigation often turns on the value of a multiple regression coefficient. Experts estimating such models are often unwilling to assume complete knowledge of the conditional density of their unobserved components. In important circumstances, a feasible “robust” estimator will be available to consistently estimate the coefficient of interest. Such estimators are usually asymptotically normal, so that their large-sample distribution differs from the MLE only in terms of variance. Working with the \( t \)-statistic eliminates this difference, so that Tsou and Royall’s Lemma 3.1, id., applies.
Figure 1: The Likelihood Function Associated with the ASCOT-LLA Data

Consider two jurors who both place prior probability one-half on the plaintiff’s litigation position as a whole (i.e., they have overall prior odds equal to 1).

- Juror A puts all prior probability, within the plaintiff’s litigation position, on the simple hypothesis that the true mean of the $t$-statistic equals the observed value of 1.2. Juror A’s Average Likelihood of $H_t$ from equation (2) thus equals the maximum of the likelihood function, which is roughly 0.4. This juror’s belief is represented by the point labeled “Max” in Figure 1’s plot of the likelihood function for the ASCOT-LLA data.
- Juror B puts equal probability on the simple hypothesis that the $t$-statistic’s true mean is 1.2 and the simple hypothesis that Lipitor increases type 2 diabetes incidence by enough to cause the $t$-statistic’s
true mean to equal 1.6. The latter point is indicated by “P” in Figure 1. Thus Juror B’s Average Likelihood of $H_1$ is the average of 0.4 and 0.37, which works out to 0.385.

Obviously Juror A’s view of the evidence is more favorable to the plaintiff than is Juror B’s. Indeed, in this instance Juror A is the most favorable juror the plaintiff could get, in terms of the juror’s prior beliefs about pro-plaintiff values of the $t$-statistic’s mean. For this reason, Juror A is the plaintiff’s most favorable juror. At the summary judgment or judgment as a matter of law stages of litigation, a non-moving plaintiff is entitled to have the court view the evidence in the light most favorable to the plaintiff. This means the plaintiff is entitled to have the court assume that the jury would be composed of twelve copies of the plaintiff’s most favorable juror. It follows that the numerator of the Bayes factor equals the maximum possible likelihood, when the evidence is viewed in the light most favorable to the plaintiff. When the $t$-statistic is normally distributed, this will always be the value of the normal density at its maximum, as depicted by point “Max” in Figure 1.

Now consider the denominator of the Bayes factor—the Average Likelihood of $H_0$. Figure 1 shows that the greatest value of the likelihood function among all defendant-favoring hypotheses occurs where the true mean of the $t$-statistic is 0, depicted by “D”, for a value is roughly 0.19. The worst-case scenario for the plaintiff is that jurors place all defendant-favoring prior probability on this point. In other words, no matter what the jurors’ prior beliefs are over defendant-favoring hypotheses, the Average Likelihood of $H_D$ can’t be more than 0.19.

Thus we have seen that in the case of the ASCOT-LLA data:

- A juror who views the evidence in the light most favorable to the plaintiff would have a Bayes factor numerator equal to 0.4.
- No juror could have a Bayes factor denominator greater than 0.19.

It follows immediately that a juror who views the evidence in the light most favorable to the plaintiff must have a Bayes factor at least equal to the ratio of 0.4 to 0.19, or 2.1. Using equation (2) with overall unconditional odds set to 1, the conditional odds in favor of the plaintiff’s litigation position are 2.1 for a reasonable juror who views the evidence in the light most favorable

---

78 Given the standard error estimated here, a $t$-statistic true mean of 1.6 corresponds to an increase of about 20% in the actual probability of type 2 diabetes onset.

79 This concept was introduced in Gelbach & Kobayashi, supra note 25, at 16.

80 Even when the $t$-statistic’s distribution is not exactly normal, the fundamental default rule derived below often will follow. It is easy to show that it holds whenever the $t$-statistic has a distribution that has a single local maximum and is increasing to the left of that maximum. A non-normal distribution with this property is the Student’s $t$ distribution with fixed degrees of freedom. This distribution has the shape just described but has “fatter tails” than the normal distribution. In sum, what’s important about the normal density for purposes of the fundamental default rule is just its overall shape, rather than its particular percentiles.
to the plaintiff, given the ASCOT-LLA data. In other words, the plaintiff’s most favorable juror believes the plaintiff’s litigation position is more than twice as probable as the defendant’s. This hypothetical juror’s posterior probability in favor of the plaintiff’s litigation position is at least 68%, well in excess of 50%-plus-a-feather’s-weight. To reject this result, one must:

(i) believe that only jurors whose prior beliefs substantially favor the defendant are reasonable; 81
(ii) believe that it is possible to view the evidence in the light most favorable to the plaintiff even while adopting prior beliefs that are not the most favorable possible ones to the plaintiff; 82 or
(iii) believe that the $t$-statistic’s actual distributional properties are very different from those of the normal distribution. 83

81 With Overall Unconditional Odds sufficiently less than $1/2.1$, or roughly 0.48, the product of the hypothetical juror’s prior odds and the Bayes factor would be less than 1.
82 In other words, rejecting Juror A described above in favor of one who places prior probability on values of Lipitor’s type 2 diabetes incidence that are less favorable for the plaintiff.
83 Simple non-normality would not be enough to ruin the result. See supra note 80.
The argument above has important implications beyond the Lipitor litigation and the ASCOT-LLA study. Nothing about this argument is tied to the particular values observed in that study. Consider Figure 2. This figure repeats the likelihood function for the ASCOT-LLA data, plotting it with a thin line, whose maximum occurs where the $t$-statistic equals 1.2. It also includes another likelihood function for the case in which the observed $t$-statistic is positive but lower. This likelihood function is plotted with thicker ink. Its maximum point occurs closer to the vertical intercept than does the maximum for the ASCOT-LLA data’s likelihood function.

As with the ASCOT-LLA data, the value of this hypothetical likelihood function is maximized for a positive observed $t$-statistic value. Once again this means the best story for the plaintiff has a greater likelihood function.

---

84 This line is blue for those reading in color.

85 This line is red for those reading in color.
value (at point “Max”) than the best case for the defendant (at point “D”). The picture is qualitatively the same for any positive value of the observed t-statistic. Thus it follows that whenever the observed t-statistic is positive, viewing the statistical estimation evidence in the light most favorable to the plaintiff will yield a likelihood ratio that must be greater than one. Assuming that a reasonable juror could be neutral toward the parties, prior odds of 1 are reasonable. Accordingly, viewing the statistical estimation evidence in the light most favorable to the plaintiff yields posterior odds greater than 1—and thus a posterior probability in favor of the plaintiff above one-half—whenever the observed t-statistic is positive. This shows that when the plaintiff must prove a variable exceeds zero, the plaintiff’s statistical estimation evidence meets the preponderance standard whenever the observed t-statistic is positive. It is easy to generalize this result to all cases in which (i) the plaintiff must prove that a variable of interest exceeds some specified value, or (ii) the plaintiff must prove that a variable of interest falls short of some specified value. This yields the fundamental default rule:

When estimation evidence points in the plaintiff’s direction, this evidence is legally sufficient to meet the preponderance standard.

Thus, the ASCOT-LLA evidence as to a 10 mg dose of Lipitor is legally sufficient. The data indicated that in the ASCOT-LLA trial, the observed type 2 diabetes incidence was greater in the treatment group, whose members received Lipitor, than in the placebo group. That is enough for legal sufficiency, thanks to the procedural rule that courts must view the evidence in the light most favorable to the plaintiff on a motion for judgment.

86 Let \( \theta \) be the true value of the variable of interest, e.g., the true effect of Lipitor on type 2 diabetes incidence, and let \( \hat{\theta} \) be an available estimator of this effect. Suppose the plaintiff must show that the true effect exceeds some minimum value \( M \), which might not be zero. Define \( \lambda = \theta - M \); this is a “re-centered” version of \( \theta \). Proving that \( \theta > M \) is the same as proving \( \lambda > 0 \). Further, defining \( \bar{\lambda} = \bar{\theta} - M \), the event that \( \bar{\theta} > M \) is the same as the event that \( \bar{\lambda} > 0 \). Because \( M \) is a constant, the variance of \( \bar{\lambda} \) is the same as the variance of \( \bar{\theta} \), so any consistent estimator \( \bar{\sigma} \) for the square-root of \( V(\bar{\theta}) \) is also consistent for \( V(\bar{\lambda}) \). It follows that all the analysis above applies to the t-statistic based on re-centering, i.e., \( T \equiv \bar{\lambda}/\bar{\sigma} \).

87 Using the same definitions as in supra note 86, the plaintiff must now prove that \( \theta < M \). Define \( \lambda = -(\theta - M) \); this is a “re-centered” and “re-signed” version of \( \theta \). Proving that \( \theta < M \) is the same as proving \( \lambda > 0 \), and the rest of the argument from supra note 86 goes through.
E. Conventional Hypothesis Testing Leads to an Inappropriate Rule for Determining Legal Sufficiency of Statistical Estimation Evidence, Unless the Significance Level of 50% is Used

In this Section, I show that in determining legal sufficiency, using conventional hypothesis testing at scholarly levels of significance leads to a too-demanding standard. I then show that there is a neat identity between the fundamental default rule and conventional hypothesis testing, so long as the 50% significance level is used. Thus, a court that wants to take the legal standard for sufficiency seriously, rather than unquestioningly applying scholarly standards in litigation, may act as if conventional hypothesis testing is appropriate, provided that it applies the appropriate significance level of 50 percent—much less demanding than scholarly levels often used now.

I now analyze the standard that plaintiffs must meet if a court uses conventional hypothesis testing. Gelbach and Kobayashi (2018) derive a minimum value for the posterior probability in favor of the plaintiffs for the plaintiff’s most favorable juror.88 Their calculations show that using conventional hypothesis testing with the most common significance level, 5%, is tantamount to requiring the plaintiff to present evidence powerful enough to convince the plaintiff’s most favorable juror that there is at least a 79% chance the plaintiff’s litigation position is correct. With more demanding choices, such as requiring a t-statistic of at least 1.96, that figure rises to 87%.89 Obviously these figures far exceed the 50%-plus-a-feather’s-weight threshold implied by the preponderance standard.

Figure 3 graphs the minimum probability standard the plaintiff must meet implied by various significance levels. The top row of labels on the horizontal axis shows the significance level, and the bottom level shows the corresponding t-statistic (e.g., a significance level of 5% corresponds to a t-

88 Gelbach & Kobayashi, supra note 25. They do so by using the fact that the plaintiff’s most favorable juror’s likelihood ratio is always at least as great as the likelihood ratio based on comparing the maximum likelihood value to the value of the likelihood function under the assumption that the t-statistic’s true mean is 0. Id. at 5. They then use the assumption that the plaintiff’s most favorable juror could reasonably have prior overall odds at least equal to 1, i.e., be non-neutral in ways that disfavor the plaintiff. Id. at 7. Then the plaintiff’s most favorable juror’s posterior odds are at least as great as the likelihood ratio comparing the hypothesis that the t-statistic’s true mean equals the observed t-statistic value to the hypothesis that the t-statistic’s true mean is zero. Id. Gelbach & Kobayashi show that when the value of the observed t-statistic, \( T \), is positive, this likelihood ratio is \( L' = \exp \left[ \frac{1}{2} T^2 \right] \); when the observed t-statistic is negative, the likelihood ratio instead is \( L^* = \exp \left[ -\frac{1}{2} T^2 \right] \). Given overall prior odds equal to 1, the posterior probability can be shown to equal \( L'/(1 + L^*) \), so this is a lower bound on the posterior probability in favor of the plaintiff held by the plaintiff’s most favorable juror.

89 Requiring a t-statistic of at least 1.96 implies that the significance level is 2.5%; this figure is commonly used by experts who incorrectly use two-sided testing. See supra note 39.
statistic of 1.65). This figure reveals an important fact about the relationship between conventional hypothesis testing and the fundamental default rule derived in Section I.D. An expert who set the significance level to 50% would exactly implement the preponderance standard with the evidence viewed in the light most favorable to the plaintiff. Any more demanding significance level (i.e., any level less than 50%) is more demanding than the preponderance standard's 50%, whereas any significance level above 50% yields a less demanding standard. The t-statistic of 1.2 corresponds to a significance level of 0.115 (the p-value for the ASCOT-LLA data), and a posterior probability in favor of the plaintiff of nearly 70%, as discussed above.

Figure 3: Conventional Significance Level and Plaintiff’s Most Favorable Juror’s Minimum Posterior Probability in Favor of the Plaintiff

Figure 3 shows starkly that for an expert using conventional hypothesis testing to follow the preponderance standard test for legal sufficiency, while viewing estimation evidence in the light most favorable to the plaintiff,
requires that the expert use a significance level of 50%. Thus, we can view conventional hypothesis testing as an appropriate method at the legal sufficiency stage as long as the significance level of 50% is used.

It follows that courts may reasonably rely on the testimony of an expert who self-consciously declares that he is testing for statistical significance at the 50% level. I am aware of two cases in which experts have done just this in the context of challenges to the admissibility of expert testimony. These cases did not involve legal sufficiency as such; the issue posed was admissibility under Rule 702. As I shall discuss below, though, the proper role of judicial gatekeeping of the strength of statistical estimation evidence for admissibility purposes is the same as it is for legal sufficiency purposes. In each of the two cases in which an expert announced his reliance on the 50% significance level, the court properly allowed the testimony.

It is useful to offer an additional framing of the fundamental default rule in terms of the $p$-value, because hypothesis testing results often are presented in that form (see discussion at the end of Section I.A). We have seen that the fundamental default rule is equivalent to conventional hypothesis testing with a significance level of 50%. It follows that the fundamental default rule may equivalently be understood as stating that estimation evidence is legally sufficient whenever the $p$-value is less than $0.5$. 

Throughout this discussion, I have assumed that the plaintiff’s most favorable juror could reasonably have overall prior odds equal to one. What if that were unreasonable in a case? For example, what if the court were sure

---


92 See In re High-Tech Emp., 2014 WL 1351040, at *12 n.25 (“Dr. Leamer may testify to . . . the fact that his alternative conduct regression model’s conduct coefficients pass the 50% level ‘suggests that it is more likely than not that the compensation of employees were decreased during the period of the agreements.’”); see also In re Photochromic Lens, 2014 WL 1338605, at *26-27 (approving the use of a significance level of .50 because the testifying expert economist justified the approach on the basis of Type II error probabilities, and also because the opposing expert failed to justify rejecting that approach except because it was “simply out of bounds of what economists do”).

93 I emphasize that this conclusion is not an instance of the oft-seen fallacy of equating the $p$-value with a posterior probability. For example, Kaye & Freedman rightly point out that it is generally mistaken to interpret the $p$-value as “the probability that defendants are innocent,” because the $p$-value “merely represents the probability of” rejecting the null hypothesis when it is correct, so that it is true, in general, that “a $p$-value less than 50% does not demonstrate a preponderance of the evidence against the null hypothesis.” Kaye & Freedman, supra note 9, at 271 n.138. I have not made that argument. Instead, I have provided a distinct argument showing that any $p$-value below 0.5 is sufficient for the plaintiff’s most favorable juror to believe the plaintiff’s story is more probable than the defendant’s. To see the distinction, observe that a juror with overall prior odds of $\frac{1}{2}$ would require a much lower $p$-value than 0.5 to reach this conclusion. See infra note 94 (explaining that a $p$-value less than or equal to 0.19 would be needed in that case).
that a plaintiff could not possibly have overall prior odds greater than some number Z? If Z were less than one, then as a matter of law the plaintiff's most favorable juror would have to have a more skeptical view of the plaintiff's case before seeing the statistical estimation evidence. That means that for any observed value of the statistical estimation evidence, the more skeptical plaintiff's most favorable juror would have a lower posterior probability than what the fundamental default rule would indicate. Accordingly, the plaintiff's most favorable juror would find the plaintiff's position more probable than the defendant's only if the statistical estimation evidence were stronger than the fundamental default rule requires. How much stronger depends on how skeptical the plaintiff's most favorable juror would have to be to be reasonable.94 Nothing in my statistical argument precludes a court from adopting such a view. All my approach requires is that the court be able to articulate a value for Z, the upper bound on the overall prior odds for the plaintiff's most favorable juror.95

F. Summary of Legal Sufficiency and the Preponderance Standard

In this Section of the Article, I started by reviewing conventional hypothesis testing, showing that it maps poorly onto the preponderance standard typically used in civil litigation when scholarly significance levels are used. I then explained how the Bayesian hypothesis testing alternative provides a snug fit to the preponderance standard. I argued that the primary obstacle to this Bayesian approach, the subjectivity of priors, can be hurdled when the question at bar is legal sufficiency—as on a motion for summary judgment or judgment as a matter of law.

In that circumstance, what matters is what the plaintiff's most favorable juror reasonably would believe. I take that question in two steps. First, if it

94 For example, if Z=1/2, then the plaintiff's most favorable juror must start out thinking the defendant is twice as likely as the plaintiff to be right. Then the generalized likelihood ratio would have to be at least 2 for the plaintiff's most favorable juror's posterior odds to exceed one—so that it takes considerably stronger evidence to convince the jury to find for the plaintiff than with a plaintiff's most favorable juror who has overall prior odds of 1. Using the formula above, that for a positive t-statistic the generalized likelihood ratio equals \( L = \exp \left( \frac{t^2}{2} \right) \), a generalized likelihood ratio of 2 would require a t-statistic of roughly 1.18. Treating the t-statistic as having approximately a standard normal distribution, the p-value associated with a t-statistic of 1.18 is 0.119. Thus if the plaintiff's most favorable juror starts out thinking the defendant is twice as likely as the plaintiff to be right, she will continue to think so unless the statistical estimation evidence is statistically significant at level 1.19 percent or lower.

95 Similarly, the court could take the view that Z is greater than one, in which case the plaintiff's most favorable juror would find the plaintiff's case more probable even with some values of the generalized likelihood ratio less than 1 (meaning, evidence that nominally points in the defendant's favor).
would be reasonable for jurors to believe the parties are equally likely to be right—i.e., if a juror with equal overall priors would not be excluded for cause—then a juror may reasonably have overall prior odds of 1. That observation allows us to focus attention on the Bayes factor. Although this is a ratio of likelihood function values averaged over priors, our interest in the plaintiff’s most favorable juror allows us to treat the generalized likelihood ratio as the appropriate value of the Bayes factor. This generalized likelihood ratio has a simple form when the statistical estimation evidence involves an approximately (or exactly) normally distributed random variable, as can usually be shown to be the case.

It is a short additional step to the fundamental default rule, according to which statistical estimation evidence that points in the direction of the proffering party is legally sufficient. Where it applies, the fundamental default rule turns out to be equivalent to using conventional hypothesis testing at the 50 percent significance level—or, what is the same, finding for the plaintiff whenever the p-value is less than 0.5. Thus, even though conventional hypothesis testing is not generally an appropriate way to implement the preponderance standard, it is appropriate if the correct significance level is used.96

II. How to Analyze Admissibility of Statistical Estimation Evidence Under the Federal Rules and Associated Case Law

Estimation evidence typically enters a trial record through expert testimony.97 This Part discusses how the Federal Rules of Evidence and associated case law properly apply to expert testimony about estimation evidence.

The first consideration is relevance under Rule 401. This rule states that evidence is relevant if “it has any tendency to make a fact” that is consequential to the action’s outcome “more or less probable than it would be without the evidence.”98 As with the preponderance standard, such language is catnip to the Bayesian. It refers directly to how “probable” a fact is, with and without the evidence. It is tailor-made for expression in formal terms, because Bayesian reasoning implies that evidence $E$ as to fact $F$ is relevant if and only if $F$’s conditional probability given $E$ differs from its unconditional probability.

In terms of equation (2), this is equivalent to saying that the evidence is associated with a Bayes factor different from 1, so that the Bayesian fact

---

96 If it would be unreasonable for jurors to have neutral priors toward the parties, then the fundamental default rule doesn’t apply, but it may be appropriately adjusted. See supra note 94.

97 In general, Fed. R. Evid. 801 makes writings describing the results of statistical study hearsay, and therefore presumptively inadmissible.

98 Fed. R. Evid. 401.
finder changes her beliefs. As Professor David Kaye has noted, the link between the language of Rule 401 and simple Bayesian ideas was first pointed out by Richard Lempert in 1977,\(^9\) though the idea dates at least to John Maynard Keynes.\(^1\)

As with the preponderance standard, the idea that Bayesian posterior probability provides a useful framework for relevance under Rule 401 has been controversial.\(^1\) But it shouldn’t be. Professor Bruce Hay has pointed out that in a jury trial case, “a judge doesn’t have to . . . decide whether (she thinks) a piece of evidence makes a fact more or less likely; rather, she has to decide whether a reasonable trier of fact might consider it to have that effect.”\(^1\) This is a natural understanding of what it means to say that a piece of evidence has “any tendency” to make a consequential fact more or less probable. Although one might agree with Professor Allen and coauthors that reasonable jurors need not be Bayesians, it is difficult to see how a juror who processes information into conditional and unconditional probabilities in the way that Bayes’s Theorem counsels—when it is possible to do so—thereby would be unreasonable. Thus, it is appropriate to regard differing conditional and unconditional probabilities as a sufficient condition for relevance under Rule 401.

This conclusion complements the results from Part I. The plaintiff’s most favorable juror has different conditional and unconditional odds when estimation evidence points in the plaintiff’s direction. That is enough to make a consequential fact more likely under the preponderance standard. So, it is enough for relevance under Rule 401.\(^1\)

---


11. For a recent demonstration, see the back-and-forth between Ronald J. Allen, Samuel Gross, Bruce Hay, David Kaye, Michael Pardo, Roger Park, and Michael Risinger in Park et al., *supra* note 54, at 10-20.

12. Park et al., *supra* note 54, at 19. To be precise, whether the judge should make this finding by a preponderance or, in line with Hay’s suggestion, determine whether a reasonable jury could find it, depends on whether “the relevance of evidence depends on whether a fact exists.” *Fed. R. Evid.* 104(b); Huddleston v. United States, 485 U.S. 681, 690 (1988) (stating that to decide whether a Rule 104(b) is satisfied, “the trial court neither weights credibility nor makes a finding that the [proponent] has proved the conditional fact by a preponderance of the evidence,” but instead “simply . . . decides whether the jury could reasonably find the conditional fact . . . by a preponderance of the evidence.”)

13. This result is congenial to the position of Michael D. Green and Joseph Sanders, who argue that “most admissibility decisions regarding expert testimony are best thought of as sufficiency judgments about the scientific evidence supporting the expert’s testimony,” although their reasons differ from mine. Michael D. Green & Joseph Sanders, *Admissibility Versus Sufficiency: Controlling the Quality of Expert Witness Testimony*, 50 *Wake Forest L. Rev.* 1037, 1058 (2015).
I turn now to Rule 702, the second consideration related to admissibility. Rule 702, as elaborated by Daubert, Joiner, and Kumho Tire, and as amended in 2000 following the trilogy, provides a set of considerations for determining whether expert testimony is admissible. The text of Rule 702 requires expert witnesses to have expertise arising from “knowledge, skill, experience, training, or education.” I take it as given that the testifying expert has appropriate qualifications in statistical methodology, as the concepts underlying my argument ultimately are quite basic ones in the theory of statistics and probability. As Rule 702 is now written, such an expert may testify if the conditions of four subdivisions are satisfied.

Subdivisions (a), (c), and (d) are best understood through a discussion of the Daubert trilogy, which I undertake momentarily. As for Rule 702(b), it requires that “the testimony [be] based on sufficient facts or data.” The Committee Note to the 2000 Amendment explains that the phrase “sufficient facts or data” adverts to “a quantitative rather than qualitative analysis.” As one treatise puts it, Rule 702(b) raises the question of “whether the expert considered enough information to make the proffered opinion reliable,” or “whether the expert ignored a significant portion of seemingly important data,” as when “an expert ‘cherry picks’ favorable data in this manner but ignores a significant quantity of other important facts . . . .” Thus an expert should consider multiple studies, if more than one exists. I show how to incorporate statistical evidence from multiple studies in the Appendix.

The remainder of this Part engages Rule 702 (Section II.A) and its proper application to estimation evidence (Section II.B).

A. Reliability Under Rule 702 and the Daubert Trilogy

Reliability entered the Rule 702 lexicon with Daubert, which stated that “under the Rules the trial judge must ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable.”

---

104 Some scholars argue that Rule 702 is unnecessary, or at least ill-justified. See generally Frederick Schauer & Barbara A. Spellman, Is Expert Evidence Really Different?, 89 NOTRE DAME L. REV. 1 (2013). I take the law as it exists, while also acknowledging the strength of their arguments.
106 FED. R. EVID. 702.
107 FED. R. EVID. 702(b).
108 FED. R. EVID. 702 committee notes to 2000 amendment.
110 Id. at 310.
112 Id. at 589.
Blackmun’s opinion explained that in cases “involving scientific evidence, evidentiary reliability will be based on scientific validity,” which the Court equated with “trustworthiness.”113 Justice Blackmun explained that the requirement that expert testimony “assist the trier of fact to understand the evidence or to determine a fact in issue . . . goes primarily to relevance.”114 In addition, Justice Blackmun offered an alternative way to understand the helpfulness requirement:

The study of the phases of the moon . . . may provide valid scientific “knowledge” about whether a certain night was dark, and if darkness is a fact in issue, the knowledge will assist the trier of fact. However (absent creditable grounds supporting such a link), evidence that the moon was full on a certain night will not assist the trier of fact in determining whether an individual was unusually likely to have behaved irrationally on that night.115

Justice Blackmun summed up the idea here by stating that “Rule 702’s ‘helpfulness’ standard requires a valid scientific connection to the pertinent inquiry as a precondition to admissibility.”116 The idea of a “valid scientific connection to the pertinent inquiry” connects to what Justice Blackmun characterized as “fit”117: scientific evidence that doesn’t seem to speak to the questions at issue won’t be helpful.

The Daubert Court declined to provide a bright-line rule for reliability. Instead, it listed several criteria that might inform the reliability inquiry as to the method underpinning an expert’s testimony: (1) “whether it can be (and has been) tested”;118 (2) “whether the theory or technique has been subjected to peer review and publication”;119 (3) where possible, “the known or potential rate of error”;120 (4) “the existence and maintenance of standards controlling the technique’s operation”;121 and (5) whether it is possible to “identify[y] a relevant scientific community and an express determination of a particular degree of acceptance within that community.”122 These are only possible criteria; as both Daubert and amended Rule 702 emphasize, they need

---

113 Id. at 590 n.9 (emphasis omitted).
114 Id. at 591.
115 Id.
116 Id. at 591-92.
117 Id.
118 Id. at 593.
119 Id.
120 Id. at 594.
121 Id.
122 Id. (quoting United States v. Downing, 753 F.2d 1224, 1238 (3d Cir. 1985)).
not be apposite in every case. Responding to defendant Merrell Dow’s “apprehension” about “a ‘free-for-all’ in which befuddled juries are confounded by absurd and irrational pseudoscientific assertions,” the Court declared, “In this regard respondent seems to us to be overly pessimistic about the capabilities of the jury and of the adversary system generally. Vigorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence.”

One key feature of Daubert was its separation of methodology and conclusions. Scientifically reliable methods, it seemed, were the ticket through Daubert’s gateway. But General Electric Co. v. Joiner seems to have kicked over that applecart. Though the Court granted cert only about the standard of review owed a district court’s admissibility decisions on appeal, Joiner appeared to swell the scope of district court discretion to exclude expert testimony. Declaring that “conclusions and methodology are not entirely distinct from one another,” Chief Justice Rehnquist’s opinion in Joiner introduced the requirement that there must not be too wide an “analytical gap” between the facts and data that form the inputs of an expert’s opinion and the conclusions that are its output. Coupled with the abuse of discretion standard that Joiner held would apply to appellate review, this determination displayed the Supreme Court’s willingness to allow district courts considerable discretion in deciding to exclude expert testimony.

In Kumho Tire Co. v. Carmichael, the Supreme Court confronted the question of whether the same gatekeeping principles apply when an expert’s testimony will not involve scientific evidence. Citing the text of Rule 702, which at the time (as now) referred not only to “scientific,” but also to “technical, or other specialized knowledge,” Justice Breyer’s opinion held that

---

123 See FED. R. EVID. 702 committee notes to 2000 amendment (“Daubert set forth a non-exclusive checklist for trial courts to use in assessing the reliability of scientific expert testimony.”) (emphasis added).

124 Daubert, 509 U.S. at 595-596.

125 Id. at 594-595 (explaining that the “overarching subject” of the admissibility inquiry “is the scientific validity and thus the evidentiary relevance and reliability—of the principles that underlie a proposed submission,” and stating that the “focus, of course, must be solely on principles and methodology, not on the conclusions that they generate”).


127 Petition for Writ of Certiorari at (i), Joiner, 522 U.S. 136 (No. 96-188), 1996 WL 33414071.

128 Joiner, 522 U.S. at 146. Arguably, a wide Joiner “analytical gap” exists whenever there is a lack of the “fit,” or “valid scientific connection to the pertinent inquiry,” required under Daubert—and vice-versa. For all the fanfare about Joiner on this point, then, perhaps the gap between it and Daubert is small.

129 Id. at 139.


131 Id. at 141-42.
the gatekeeping requirement *Daubert* established for determining the admissibility of scientific evidence applied as well to other forms of expert evidence.\(^{132}\) Further, Justice Breyer explained that “[t]he objective of [Daubert’s gatekeeping] requirement is to . . . make certain that an expert . . . employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field.”\(^{133}\)

Putting all this together, Rule 702 and the *Daubert* trilogy erect a variety of hurdles to admissibility of expert testimony, over and above relevance as Rule 401 defines it. The testimony must help the fact finder understand other evidence or determine a fact. When it involves scientific evidence, the testimony must be based on the use of methods that fit the issues the testimony addresses. The conclusions embraced in the testimony must not stray too far from those that a district court, operating within its usual discretion, could think are fairly warranted by the facts and data used to draw the conclusion. And the testimony must be founded on the same level of intellectual rigor that is generally used outside the courtroom by those in the testifying expert’s field of expertise.

**B. Understanding the Concepts of Strength and Credibility of Estimation Evidence**

In this Section, I analyze statistical estimation evidence in terms of the two core aspects of evidence on which the law of evidence concentrates. First is probativeness—how convincing the evidence in question will be, if it is admitted. Second is reliability—how trustworthy the evidence is, such that the evidence-law value of accuracy is not threatened by allowing jurors to draw such inferences as a person might naturally draw from the evidence in question.

It is important to understand that statisticians sometimes use the same words as evidence law commentators in different ways. Statistical experts do not necessarily draw a clear line between what evidence law commentators describe as probativeness and reliability. In particular, whereas in evidence law “credible” links to reliability rather than probativeness, a statistical expert who states that “credible evidence indicates $X$” might or might not mean the same thing as one who states that “strong evidence indicates $X$.” In this Section, I present a view that links “credibility” and “strength” of statistical evidence respectively to reliability and probativeness in the evidence law sense. This discussion is useful because although statistical experts usually are

\(^{132}\) *Id.* at 147, 158.

\(^{133}\) *Id.* at 152. As Justice Scalia’s brief concurrence illustrates, the *Daubert* trilogy cases were decided against a backdrop of concern about the extent of made-up expertise in not-very-scientific fields that have limited or no footprint outside of litigation. *Id.* at 159 (Scalia, J., concurring) (“[D]iscretion to choose among reasonable means of excluding expertise that is false and science that is junky”).
able to understand each other via pragmatic context, to outsiders, it might be unclear which notion is involved. Readers should keep in mind the particular link I draw.

Most estimation evidence of interest can be discussed in terms of $t$-statistics. As discussed in Part I, a greater $t$-statistic indicates that the plaintiff’s most favorable juror will regard estimation evidence as more strongly favoring the plaintiff. Experts using conventional hypothesis testing also view higher $t$-statistics as stronger evidence against the null hypothesis. Equivalently, because lower $p$-values correspond to $t$-statistics of lesser magnitude, lower $p$-values indicate stronger estimation evidence. All of this implies that the plaintiff’s most favorable juror and experts using conventional hypothesis testing will agree on the ranking of two pieces of statistical estimation evidence: an estimate with a lower $p$-value provides stronger evidence in favor of the plaintiff.

Nothing in the previous paragraph’s discussion related to whether the estimation evidence came from a well-designed study. The role of $p$-values and $t$-statistics is entirely distinct from that question. A poorly designed study that has little to do with the question of interest can have an enormous $t$-statistic and thus provide strong evidence about some question—just as a perfectly designed and conducted study might yield only quite weak evidence, in the form of a $t$-statistic close to 0.

A thought experiment will help distinguish the qualities of strength and credibility. Suppose all sides to a controversy (legal or not) are behind a veil of ignorance as to the numerical results of a statistical study. After receiving a description of the study’s methods—who is assigned which treatments or placebos, how the data will be collected, how the effect at issue will be measured—all sides to the controversy agree that the study is reasonably designed. It seems beyond question that such a study would be credible for purposes of the controversy involved. Because the criteria for credibility could be stated without regard to the results, credibility in the reliability sense of evidence law is distinct from the character of the results themselves. This allows us to separate our analysis of the strength of estimation evidence and the credibility of the underlying study.

---

134 To be sure, this is not generally the case when multiple parameters’ values are simultaneously tested (in which case a combination of statistics usually forms the test statistic, e.g., via a $\chi^2$ statistic). Nor is it the case in certain situations in which there is good reason to doubt normality. See, e.g., Jonah B. Gelbach, Eric Helland & Jonathan Klick, Valid Inference in Single-Firm, Single-Event Studies, 15 AM. L. & ECON. REV. 495, 498 (2013) (using “analytical arguments” in a study “to illustrate the importance of normality of the distribution of excess returns for achieving valid inference, even asymptotically”); Fisch, Gelbach & Klick, supra note 29, at 575 n. 126 (citing Alon Braw & J.B. Heaton, Event Studies in Securities Litigation: Low Power, Confounding Effects, and Bias, 93 WASH. U. L. REV. 583, 591 n. 17 (2015)) (noting that standard practice is to rely on the assumption of normality).
Within the category of credibility, it is useful to further distinguish two
qualities: technical implementation and credibility for the proffered purpose.

1. Technical Implementation

Technical implementation concerns whether the statistical evidence in
question was calculated as it should have been, given the representations the
expert makes about the results.

An expert might accidentally forget to count some observed values in
calculating an average, might divide by the wrong number of observations,
and so on. Or, the data used to calculate the statistical evidence might have
been the result of data entry errors, with otherwise correct mathematical
formulas then applied to the contaminated data set. Statistical evidence
generated in such ways would be mechanically inaccurate; such inaccuracies
might or might not be important, depending on the context.\textsuperscript{135}

A third problem arises when an expert estimates a model in a way that is
highly likely, if not guaranteed, to result in biased estimates.\textsuperscript{136} A fourth
example arises when an analyst engages in deliberate manipulation of data in
searching for a model specification that yields results in line with the analyst’s
pre-conceived idea of what the results should be.\textsuperscript{137}

In all of these cases, many professional statisticians would be skeptical of
an expert’s analysis, even without knowing the specific results of the
estimation evidence.

2. Credibility for the Proffered Purpose

This subsection discusses credibility for the proffered purpose, which is
about the overall fit between the statistical evidence flowing from data at hand
and the purpose for which a party intends to use it. That distinguishes
credibility for the proffered purpose from strength of evidence and technical
implementation. Those features can be thought of as characteristics that are

\textsuperscript{135} For an example and brief discussion of evidence law considerations, see D.H. Kaye, \textit{The
Dynamics of Daubert: Methodology, Conclusions, and Fit in Statistical and Econometric Studies}, 87 VA. L.

\textsuperscript{136} For example, it is well known that using ordinary least squares to estimate regression
coefficients yields biased estimates when the dependent variable is the change in some variable,
\(y_t\), over time, when \(y_t\) is serially correlated, and when the set of independent variables included
in the estimation includes \(y_{t-1}\). For more on bias, see Kaye & Freedman, supra note 9, at 249 and
references therein.

\textsuperscript{137} This practice is variously known as specification searching, data snooping, and data mining
(though the last term has recently acquired a neutral or even positive connotation in many fields). Halbert
White, \textit{A Reality Check for Data Snooping}, 68 ECONOMETRICA 1097, 1097-98 (2000); see also
E.E.O.C. v. Datapoint Corp., 570 F.2d 1264, 1270 (5th Cir. 1978) (determining that an EEOC
statistician had engaged in such conduct).
internal to a particular study, because they directly involve calculations using the data at hand. Credibility for the proffered purpose cannot.

To illustrate credibility for the proffered purpose, take an infamous example from outside the law: the prediction of the Literary Digest that Republican Alf Landon would defeat President Franklin D. Roosevelt in the 1936 election by a popular vote margin of 3 to 2.138 The Literary Digest prediction was made from a large sample of mail-poll respondents.139 From an internal point of view, the direction, magnitude, and strength of this statistical evidence were all very impressive. Yet Roosevelt went on to win with 62% of the popular vote, carrying 46 of the then-48 states.

What went wrong is that the survey’s respondents were highly self-selected: As statistics professor Maurice Bryson would point out 40 years later, “the minority of anti-Roosevelt voters felt more strongly about the election than did the pro-Roosevelt majority.”140

Because the survey had a huge sample size—2.3 million141—it yielded an estimate so precise that no one could seriously view the discrepancy between the poll and the election’s outcome as driven by random variation in the particular sample chosen. The strength of evidence was enormous, but the evidence powerfully answered the wrong question.142

The problem, Bryson’s explanation indicates, was that the population the respondents represented did not represent the electorate. No statistician who was aware of this non-representativeness before the election would have considered the poll credible for the purpose of predicting who would win the election, or by what margin.

In Daubert terms, the Literary Digest poll did not “fit” the question at hand because it was ill-suited to uncover information about the population of interest. In Joiner terms, there’s an “analytical gap” between the population about which we could reasonably expect to learn from this poll, and the population whose characteristics were actually at issue. And in Kumho Tire terms, use of the Literary Digest poll is outside the ken of what an informed statistical expert would rely on for her out-of-court work. Accordingly, in a hypothetical case about the 1936 election, a district court should keep the gates shut to testimony based on that poll’s results.

139 Id. at 185.
140 Id. This account of the doomed poll also includes an interesting rebuttal of the usual explanation for the Literary Digest poll’s face-plant—that the problem had to do with systematic over-representation of Republicans among those with telephones. See id. at 184-185.
141 Id. at 185.
142 Nor, to my knowledge, is there any reason to think the survey suffered from any technical implementation problem.
Thus the Literary Digest poll well illustrates the issue of credibility for the proffered purpose, and its link to evidence-law reliability. Statistical evidence is credible for the proffered purpose if the object it accurately measures is the object the evidence is supposed to illuminate. When there are good reasons to doubt that the object the study actually estimated was the object of interest, an estimate can point in the direction of a party’s position, have a substantial magnitude, and be highly statistically significant (or have a high likelihood ratio)—and still not be credible for the proffered purpose.

By its very nature, credibility for the proffered purpose cannot easily be assessed using the data at hand. An epidemiological study relating, say, smoking and lung cancer is credible for the purpose of establishing that smoking causes lung cancer only if one is willing to rule out—say—the possibility that those who ultimately will be diagnosed with lung cancer are just more likely to have taken up smoking. That would be an example of omitted variables bias, also sometimes known as confounding. Another alternative explanation might be that lung cancer causes people to smoke, or—more plausibly—that the same third variable that causes people to smoke also causes lung cancer. How does one rule out such alternative explanations? In the case of smoking and lung cancer, the answer was a combination of publicly released epidemiological and lab research, as well as smoking-gun research results the tobacco companies hid from the public for years, which materialized only as a result of litigation.

Perhaps the leading methodological framework for assessing causation—and thus credibility for the proffered purpose of proving causation—in toxic torts litigation is now the Bradford Hills criteria. Except for the magnitude of the effect—according to which smaller effects are more likely to be the result of chance variations than true causation—these criteria tend to involve the assessment of results from multiple statistical analyses. Thus, they cannot answer the question of whether testimony as to any single piece of statistical

---


144 Fisher made this weird argument, too. See Ronald Fisher, Cigarettes, Cancer, and Statistics, 2 CENTENNIAL REV. ARTS & SCI. 151, 162 (1958) (“Is it possible, then, that lung cancer—that is to say, the pre-cancerous condition which must exist and is known to exist for years in those who are going to show overt lung cancer—is one of the causes of smoking cigarettes? I don’t think it can be excluded.”); see also Stolley, supra note 143, at 419 (noting the possibility of this causal relationship).


evidence should be admissible into evidence or legally sufficient, which is my primary focus in this Article.

Sometimes, results from randomized controlled trials [RCTs] will be available. In these settings, credibility for the proffered purpose might seem to follow directly from the fact of randomization, provided the sample is large and the study is well executed. Indeed, it is widely accepted that RCTs provide causal evidence. But even this is too facile, because in the strictest sense, the causal evidence RCTs provide is limited to the population represented by the collection of people studied at the time they were studied.\textsuperscript{147} To believe RCT evidence is applicable still requires one to believe that the study’s population, dosage, and other characteristics are not too different from the corresponding facts of the litigation at hand.

Whole careers have been spent discussing these and related questions, and for the sake of manageability I shall not go further here. The take-home point is that in litigation as in life generally, an analytical leap over at least some gap is virtually always required to believe that estimation evidence can credibly answer a specific question of interest.\textsuperscript{148}

3. Analogy to Non-Statistical Evidence

The strength and credibility characteristics just described have analogues in non-statistical evidence. Suppose Paula sues Dave for fraud. At trial, Paula plays an audio recording of Dave, who stipulates to the authenticity of the recording.\textsuperscript{149} On the recording, Dave says, “Let’s make sure we don’t mislead Paula in this transaction.” Paula argues to the jury that Dave’s tone of voice on the recording is dripping with conspiratorial sarcasm. At closing, Dave’s attorney testifies that a reasonable person could listen to the recording and conclude that Dave was speaking sincerely. The jury finds for Dave, and in a

\textsuperscript{147} An asthma remedy proven to help in a climate with few airborne particulates would presumably do little good for an asthmatic caught in a building enveloped by a smoke-spewing inferno. More generally, randomization cannot solve every interesting inferential problem. See, e.g., James J. Heckman & Jeffrey A. Smith, Assessing the Case for Social Experiments, 9 J. ECON. PERSP. 85, 99 (1995) (describing randomization bias, which “occurs when random assignment causes the type of persons participating in a program to differ from the type that would participate in the program as it normally operates”); Nancy Cartwright, Are RCTs the Gold Standard? 2 BIOSOCIETIES 11, 11 (2007) (noting that randomization may not be the best to ensure reliability in studies); Angus Deaton, Instruments, Randomization, and Learning about Development, 48 J. ECON. LITERATURE 424, 426 (2010) (noting that “[r]andomized controlled trials cannot automatically trump other evidence . . . ”).

\textsuperscript{148} People get less carried away about such gaps than do scholars. As a former economics-department colleague of mine opined wryly, few people wonder whether those likely to die within a period of seconds of study are prospectively more likely to be shot in the chest. Similarly, I do not doubt that I should wait to cross a busy street, even though I’ve never seen a compellingly executed RCT of whether it hurts to be run over by cars. These examples show the value of practical reasoning starting from reasonable assumptions.

\textsuperscript{149} Such a recording is not hearsay when offered against a party. FED. R. EVID. 801(d)(2)(A).
post-verdict interview, jurors report that they found the recording equally consistent with sincerity and sarcasm. So although the recording was highly reliable—highly credible—evidence as to its subject matter, the jury did not see it as being highly probative—strong—evidence in favor of Paula's position.

Now twist the hypo two ways. First, suppose any listener would hear the voice on the recording as dripping with sarcasm. Second, instead of stipulating to the recording’s authenticity, Dave claims he never made the statement in question—he testifies that the recording was fabricated. The recording is strong evidence for its subject matter: if it is taken at face value, reasonable jurors could find for Paula (for just this reason its improper admission over Dave's objection would be reversible error). The question for a juror considering whether to rely on the recording is whether it is sufficiently credible to do so.\(^{150}\)

These two examples underscore the distinction between credibility for the proffered purpose, which links to evidentiary reliability, and strength of evidence, which links to probativeness. Just as evidence law can regard non-statistical evidence as highly reliable but not very probative in supporting the party who offered it, so can the law regard statistical evidence as highly credible but weak. The reverse is true as well: just as evidence law can view non-statistical evidence as relatively unreliable but highly probative if believed, so can evidence law treat statistical evidence as having low credibility but quite powerful given that it is taken at face value.

C. The Proper Application of Rule 702 to Estimation Evidence Is Friendly to the Fundamental Default Rule\(^{151}\)

Having drawn the distinction between strength and credibility of evidence, we are in position to discuss the appropriate role of judicial gatekeeping under Rule 702. My central contention is that appropriate gatekeeping is largely unrelated to the strength of estimation evidence. Instead, gatekeeping should be focused primarily on ensuring that the statistical evidence used was generated in a technically competent way and has reasonable credibility for the proffered purpose.

\(^{150}\) The admissibility determination here typically would be based on FED. R. EVID. 901 (“To satisfy the requirement of authenticating or identifying an item of evidence, the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is.”). This involves a Rule 104 question, and as usual, the jury would be within its rights to view admitted evidence as non-credible.

\(^{151}\) This Section includes some text drawn verbatim from an amicus brief I recently authored, Brief for Amicus Curiae Jonah B. Gelbach in Support of Plaintiffs-Appellants 7-11, In re Lipitor (Atorvastatin Calcium) Mktg., Sales Practices & Prods. Liab. Litig. (No. II) MDL 2502, 892 F.3d 624 (4th Cir. 2018), 2017 WL 1628475 (C.A.4).
Rule 702 gatekeeping is supposed to be about reliability, understood as involving the “fit” between expert testimony and the question to which that testimony speaks. Suppose an expert proposes to faithfully describe estimation evidence from a study that was credibly designed and competently implemented, and which only weakly supports the plaintiff’s litigation position. Such testimony is reliable for establishing the proposition that the statistical facts weakly support the plaintiff’s position.

In the rest of this Section, I first argue that Rule 702 and associated case law do not support the exclusion of expert testimony about estimation evidence merely because the estimation evidence fails to satisfy conventional hypothesis testing at conventional significance levels. As my discussion above suggests, the problem with such exclusion is that significance level is best understood as a measure of strength, or probativeness, rather than of reliability in the evidence law sense. I then discuss the ways in which federal evidence law appropriately does limit the admissibility of expert testimony about estimation evidence. Proper limitations are anchored in unreliability rather than weakness; failure of technical implementation and lack of credibility for the proffered purpose both provide bases for exclusion. By contrast, questions related to the strength of evidence—e.g., the weight accorded to it—are left to the jury in our system.

1. Rule 702 and the Daubert Trilogy Do Not Allow Gatekeeping Based on Conventional Hypothesis Testing at Conventional Significance Levels

Rule 702(a) requires that expert testimony about estimation evidence “help the trier of fact to understand the evidence or to determine a fact in issue.”\footnote{Fed. R. Evid. 702(a).} Consider the Lipitor case again. Dr. Singh wrote in his expert report that the ASCOT-LLA estimation evidence was not statistically significant at conventional significance levels.\footnote{Dr. Singh Supplemental Report at 3.} He also wrote that “the direction of effect was consistent with an increased risk of diabetes,”\footnote{Id. (emphasis in original).} and he stated that based on analysis of the ASCOT-LLA data as well as other information he used in applying the Bradford Hills criteria, he found “beyond a reasonable degree of certainty that [Lipitor] 10 mg increases the risk of diabetes.”\footnote{Id. at 32. I do not take a position on whether the additional information Dr. Singh considered was consistent with Rule 702; my position is that the ASCOT-LLA evidence alone should have been enough. I note that Dr. Singh did write that the ASCOT-LLA trial is “unreliable in establishing” the relationship between Lipitor at 10 mg and type 2 diabetes, id. at 3-4, but it is clear from context that he is arguing that this supposed unreliability will “bias . . . findings toward the null” hypothesis of no effect. Id. at 3.} The district
court's exclusion of Dr. Singh's testimony was partly based on its view that he had misapplied the Bradford Hills criteria, and for purposes of this Article, I take no position on that issue.

But imagine if Dr. Singh had testified in his deposition simply that the ASCOT-LLA evidence was more consistent with causation of type 2 diabetes than with its absence. A reasonable person who understands Bayesian hypothesis testing could believe this based on the ASCOT-LLA estimation evidence. That testimony surely could have been helpful to the jury.

Rule 702(b) requires that an expert's testimony be "based on sufficient facts or data." This has been understood to mean that the expert testimony must not cherry pick, or disregard inconvenient facts or data without good reason. In addition to contravening Rule 702(b), ignoring studies that reach conclusions unhelpful to an expert's proponent might also diminish the credibility for the proffered purpose of the estimation evidence, raising direct questions about the reliability of the expert's testimony or the estimation evidence underlying it. In Part I, I implicitly assumed that only one study is available concerning the question at issue. Roughly speaking, that was the situation for the case of Lipitor at the 10mg dose. In other cases, there will be multiple sources of statistical estimation evidence—for example, multiple RCTs or other studies evaluating the effect of interest. My argument from Part I can be adapted to account for those cases; see the Appendix to this Article for details.

I turn now to the parts of Rule 702 that directly require reliability. Rule 702(c) requires that expert testimony be the "product of reliable principles and methods," and Rule 702(d) further requires that "the expert has reliably applied the principles and methods to the facts of the case." Bayesian hypothesis testing involves fundamental principles of mathematical statistics that certainly qualify as reliable principles and methods. Thus expert testimony is reliable if it involves the correct use of Bayesian hypothesis testing to draw inferences from estimation evidence that emerges from a credibly designed and competently executed study.

For the reasons just given, expert testimony as to statistical results satisfies the text of Rule 702 provided that (i) the expert correctly applies Bayesian hypothesis testing methods to draw conclusions from underlying estimation

156 See Section I.D.
157 See WRIGHT & MILLER, supra note 109.
158 ASCOT-LLA was the only study that directly measured type 2 diabetes incidence based on a 10mg dose. Dr. Singh Supp. Rep. at 3-4. Although Dr. Singh argued that there were other reasons to think this dose would cause type 2 diabetes onset, these reasons involved reasoning based on "indirect and mechanistic evidence" rather than direct estimates of the effect of a 10mg dose. Id. at 32.
159 FED. R. EVID. 702(c) & (d).
evidence that was generated from a study that is both (ii) credible for the purpose proffered and (iii) is implemented in a technically appropriate way.

Rule 702’s current text is the result of a major amendment in 2000, following the Daubert trilogy. Although the Rule itself is what controls, courts regularly look to the trilogy cases for elaboration, so it is important to engage these cases. I do so now.

a. Daubert

Daubert instructs that reliability means scientific validity, understood to mean that the expert evidence at issue “support[s] what it purports to show.” Applying Bayes’s Theorem to statistical evidence meets this requirement because Bayes’s Theorem is true as a matter of mathematical logic. Accordingly, use of the theorem’s implications supports any proposition within the scope of the theorem’s subject matter.

Using Bayes’s Theorem also satisfies the several factors identified by the Daubert Court as indicia of reliability. Daubert explains that a “pertinent consideration is whether the theory . . . has been subjected to peer review and publication.” Bayes’s Theorem is among the most venerable results in statistics. A special case of the theorem was published, posthumously under Bayes’s name, in 1763. The role of the likelihood ratio in updating from prior odds to posterior odds is both mathematically apparent and the subject of countless published discussions.

Daubert also states that “the court ordinarily should consider the known or potential rate of error” of a scientific approach that is the subject of expert testimony or evidence. As explained above, the likelihood ratio’s role in updating prior beliefs is a matter of mathematical logic, not estimation or

---

160 See Fed. R. Evid. 702 committee notes to 2000 amendment (“Rule 702 has been amended in response to Daubert . . . and to the many cases applying Daubert . . .”).

161 This language comes from Brief for Amicus Curiae Jonah B. Gelbach in Support of Plaintiffs-Appellants, supra note 151, at *19 (citing Daubert, 509 U.S. at 590 n.9).

162 Id.

163 Id.

164 Id.

165 Id. (citing Daubert, 509 U.S. at 593).

166 Id.

167 Id. (citing Thomas Bayes, An Essay Towards Solving a Problem in the Doctrine of Chances, 53 Phil. Trans. 370 (1763)). Interestingly, there is some doubt as to whether Bayes was the first to state the result. See Stephen M. Stigler, Who Discovered Bayes’s Theorem?, 37 Am. Statistician 290, 290 (1983) (discussing theories of the origin of Bayes’s theorem).

168 Id.; see, e.g., James O. Berger, Statistical Decision Theory and Bayesian Analysis (1985).

169 Brief for Amicus Curiae Jonah B. Gelbach in Support of Plaintiffs-Appellants, supra note 151, at *20 (citing Daubert, 509 U.S. at 594).
measurement, so its use cannot itself cause error. An empirical test of a mathematically provable proposition can yield a result falsifying the proposition only if the empirical test itself is faulty. So Bayes-based expert testimony that evidence is more consistent with causation than with its absence satisfies Daubert with respect to “whether [the scientific approach at issue] can be (and has been) tested.”

Finally, there can be no question that Bayes’s Theorem and the likelihood ratio are “generally accepted” tools for a scientific community under Daubert. Bayesian statistics is a theoretical field of its own, as well as the foundation for much reasoning and estimation in many cognate applied fields, such as psychology, economics, medicine, and so on. At a minimum, the substantial community of self-identified Bayesians working in applied fields—including psychologists, biostatisticians, econometricians, and decision theorists—subscribe to all the principles described above. No member of the wider statistical community—including frequentists—denies

---

170 Id.
171 Id.
172 Daubert, 509 U.S. at 593. None of this is inconsistent with the fact that empirical evidence indicates there’s lots of non-Bayesian behavior in the world. See, e.g., Nancy Pennington & Reid Hastie, Explaining the Evidence: Tests of the Story Model for Juror Decision Making, 62 J. PERSONALITY & SOC. PSYCHOL. 189 (1992) (discussing the prevalence of causal reasoning in decision making processes and suggesting that jurors in a criminal trial follow the “Story Model” for their decision-making processes). That shows only that particular people do not update according to Bayes’s Theorem—not that the theorem is an unreliable way to update beliefs. An important class of violations of the theorem involve what is often referred to as the “base-rate fallacy,” and methods to avoid it are often referred to by psychologists as “de-biasing.” See, e.g., Baruch Fischhoff, Judgment and Decision Making, 1 WIREs COGNITIVE SCI. 724, 727 (2010) (explaining how the “base rate fallacy” involves allowing even weak information about specific cases to outweigh knowledge of what generally happens (the base rate”)); Baruch Fischhoff, Debiasing, in JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES 431-44 (Daniel Kahneman, Paul Slovic & Amos Tversky eds., 1982) (describing efforts to ameliorate hindsight and overconfidence biases in studies).

173 This language comes from Brief for Amicus Curiae Jonah B. Gelbach in Support of Plaintiffs-Appellants, supra note 151, at *20 (citing Daubert, 509 U.S. at 594).


175 This language comes from Brief for Amicus Curiae Jonah B. Gelbach in Support of Plaintiffs-Appellants, supra note 151, at *20.
the mathematical truth of Bayes’s Theorem and the likelihood ratio’s role in updating probabilities, given their initial values.176

In sum, because the probability of a fact relative to its negation is made mathematically more probable when the likelihood ratio in favor of that fact exceeds 1, to say that “the trial judge must ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable,” requires only that the statistical evidence be competently generated and credibly related to the fact it is offered to establish.177

One way to make sense of all this is to observe that evidence law principles apply to statistical significance as much as they apply to witness testimony in general.178 Just as lay testimony may be both reliable and weak—as in the first hypo involving the recording of Dave’s voice179—statistical evidence may be more consistent with the plaintiff’s litigation position than the defendant’s and also have a $p$-value higher than what scholars would consider sufficient to announce a scientifically notable finding.

Properly understood, then, statistical significance goes not to admissibility but to weight.180 The Daubert Court explicitly considered that distinction when it suggested that the respondent’s argument for excluding expert testimony in that case seems to us to be overly pessimistic about the capabilities of the jury and of the adversary system generally. Vigorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence . . . Additionally, in the event the trial court concludes that the scintilla of evidence presented supporting a position is insufficient to allow a reasonable juror to conclude that the position more likely than not is true, the court remains free to direct a judgment . . . and likewise to grant summary judgment. . . . These conventional devices, rather than wholesale exclusion

176 Id.; see, e.g., ANDREW GELMAN ET AL., supra note 174, at 3 (promoting the use of Bayes’s theorem in statistical analysis, noting that “[t]he essential characteristics of Bayesian methods is their explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis.”); see also BERGER, supra note 168 (addressing both Bayesian and non-Bayesian decision theory). There are major methodological disagreements as to the appropriateness of Bayesian estimation among those who use statistical methods in professional work. But non-Bayesians do not question the mathematical correctness of any of the claims made supra. Rather, they question the appropriateness of Bayesian estimation techniques because these require specification of particular prior beliefs. The arguments used supra make use of objective legal standards to solve this subjectivity problem (except where useful in demonstrative examples).

177 This language comes from Brief for Amicus Curiae Jonah B. Gelbach in Support of Plaintiffs-Appellants, supra note 151, at *20 (citing Daubert, 509 U.S. at 589).

178 Id. at *22.

179 Id.; see subsection II.B.3, supra.

180 That is, statistical significance is like any other type of evidence, such as witness credibility, whose importance should be left up to the fact finder to decide.
... are the appropriate safeguards where the basis of scientific testimony meets the standards of Rule 702.181

Daubert’s reference to “shaky but admissible evidence” obliterates the basis for requiring statistical evidence to meet significance levels conventionally required by scholars in their scholarly capacities. By design, significance levels used in scholarship are not “shaky.”182 They are meant to be powerful and convincing.183 Otherwise, the standards in use would not have been adopted to establish that statistical evidence is sufficient to constitute knowledge.184 If such significance levels could be required by trial courts for statistical evidence to be admissible, “shaky” statistical evidence could not be admitted at all, contravening the above passage in Daubert.185 Even more tellingly, it would render pointless the adversarial devices and procedural mechanisms, such as summary judgment, that Daubert emphasized.186

b. Joiner

Joiner is most connected to my analysis because of its insistence that conclusions and methodology are not entirely distinct, such that too great an “analytical gap” between the two will doom an expert’s testimony.187

181 This language comes from Brief for Amicus Curiae Jonah B. Gelbach in Support of Plaintiffs-Appellants, supra note 151, at *20 (citing Daubert, 509 U.S. at 596 (citations omitted)).
182 Id. at *23.
183 Id.
184 That is not to say that everyone regards, say, the conventional 5% significance level as powerful and convincing. For example, Berger, supra note 168, at 151-152, provides calculations showing that support for the null hypothesis in conventional hypothesis testing is often much stronger than a naïve guess based on the p-value would indicate. And a recently released paper co-authored by seventy-two prominent applied and theoretical statisticians argues that the threshold p-value of 0.05 is too liberal for use in scientific contexts, because it leads to too many “discoveries” of non-existent relationships. Daniel J. Benjamin et al., Redefine Statistical Significance, 2 NATURE HUM. BEHAV. 6, 6-10 (2018). The authors “emphasize that this proposal is about standards of [scientific] evidence, not standards for policy action,” which could as well be applied to litigation. Id. at 8. See also Sander Greenland et al., Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations, 31 EUR. J. EPIDEMIOLOGY 337, 340 (2016) (describing the prominent misconceptions of p values); Ronald L. Wasserstein & Nicole A. Lazar, The ASA’s Statement on p-Values: Context, Process, and Purpose, 70 AM. STATISTICIAN 129, 129 (2016) (discussing the challenges surrounding the American Statistical Association’s decision to “develop[] a policy statement on p-values and statistical significance ..... ”); see, e.g., STEPHEN T. ZILLI & DEIRDRE N. McCLOSKEY, THE CULT OF STATISTICAL SIGNIFICANCE: HOW THE STANDARD ERROR COSTS US JOBS, JUSTICE, AND LIVES 2 (2008) (criticizing conventional hypothesis testing).
185 This language comes from Brief for Amicus Curiae Jonah B. Gelbach in Support of Plaintiffs-Appellants, supra note 151, at *23.
186 Id.
187 Interestingly, the sole question presented in the petition for certiorari in Joiner was: “What is the standard of appellate review for trial court decisions excluding expert testimony under Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993)?” Petition for Writ of Certiorari at (i),
This renunciation of Daubert’s assertion that the “focus, of course, must be solely on principles and methodology, not on the conclusions that they generate” might seem to reopen the question of whether strength of evidence is the most that the degree of statistical significance can usefully measure.\(^{188}\)

In the most widely cited part of his opinion for the Joiner Court, Chief Justice Rehnquist wrote:

> Conclusions and methodology are not entirely distinct from one another. Trained experts commonly extrapolate from existing data. But nothing in either Daubert or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the ipse dixit of the expert. A court may conclude that there is simply too great an analytical gap between the data and the opinion proffered.\(^{189}\)

There is much truth to this statement. Statistical evidence is often founded either on events that have already occurred outside the immediate context of the case at bar, or on some sample of the events that form that context. The same is usually true outside the litigation context, actually. It would be mock-worthy to question the probative value of the results of a flawlessly designed and executed double-blind RCT of a drug’s efficacy. At least, it would if the skepticism were directed at the experimental population actually included in the study at the time it was conducted. But if the question in litigation involves substantially different people, substantially different circumstances, or both, then one needn’t be a gadfly to question the applicability of the study’s results.\(^{190}\)

And all of that is true of the very highest-quality studies—those that are highly credible for the purpose for which they were designed. Studies that do not use randomized assignment—or involve animals rather than humans, or humans exposed to different substances than the ones at issue in litigation, and so on—require a further leap of analytical faith to be treated as credible.

---

\(^{188}\) Daubert, 509 U.S. at 595.

\(^{189}\) Joiner, 522 U.S. at 146.

\(^{190}\) See Merck & Co., Inc., v. Garza, 347 S.W.3d 256, 262-66 (Tex. 2011) (finding that well-executed studies suggesting an increase of injury must be considered alongside other factors excluding other possible causes of the injury for there to be sufficient evidence to find for the plaintiff); Merrell Dow Pharm., Inc. v. Havner, 953 S.W.2d 706, 724-30 (Tex. 1997) (holding that three different studies on the cardiovascular effects of the drug in question or a similar drug did not show statistically significant doubling of risk for someone like the plaintiff); Jonah B. Gelbach, The Triangle of Law and the Role of Evidence in Class Action Litigation, 165 U. PA. L. REV. 1807, 1819 (2017) (arguing that in Tyson Foods, Inc. v. Bouaphakeo, 136 S. Ct. 1036 (2016), “[t]he operative question for the plaintiffs’ evidence [wa]s . . . whether [measured data] could be assumed to be sufficiently similar so that it would be reasonable to use an overall measure—such as an average—in place of . . . unknowable actual’ data).
This is as true for an expert assembling an opinion from such studies as it is for a layperson. Where expertise runs out, and runs into generalized inference, it is appropriate for a judge to exercise reasoned discretion in applying her own personal intuitions. When the question is whether to allow inferences from a particular chunk of statistical evidence gathered from a different population or under different circumstances, it is no answer for a party to merely point to the label “expert.” Ex hypothesi, the expert’s expertise concerns something other than the reasonableness of actually using the evidence in question. So when an expert’s conclusions are linked to the statistical evidence at issue only by the expert’s “ipse dixit,” the trial court owes the expert no deference as to whether to draw or reject that link.

Two of the studies relied upon by plaintiffs’ experts in Joiner—referred to by Chief Justice Rehnquist as the third and fourth studies at issue—fit this picture well. The third study involved an examination of the effects of mineral oil, rather than the PCB chemicals that were the alleged cause of Mr. Joiner’s lung cancer; Chief Justice Rehnquist pointed out that this study “made no mention of PCB” chemicals. The fourth study involved people who had been exposed to PCB chemicals, but also to “numerous [other] potential carcinogens.” The trial court could reasonably determine that these studies weren’t relevant to the Joiners’ case, because too great a leap of faith would be required to believe them relevant—in Joiner’s terms, because the analytical gap between the statistical evidence and the case at bar was too wide.

The first two studies at issue in Joiner appear to have been quite different from the third and fourth. As Chief Justice Rehnquist describes the first two, each involved workers who had been exposed to PCB chemicals. Both of these studies found that deaths from lung cancer in the studied populations exceeded what would normally be expected. Joiner held that the district court was within its discretion to rule that these studies provided an insufficient basis for the expert’s opinion. As to the first study, that was because the study’s authors “were unwilling to say that PCB exposure had caused cancer among the workers they examined.” As to the second, it was because the extent of the elevated incidence of lung cancer deaths “was not statistically

191 Joiner, 522 U.S. at 145. Mr. Joiner alleged that his exposure to PCB chemicals occurred via contamination of the mineral oil-based coolant he used in his work as an electrician; thus, it was the carcinogenic effect of contaminating chemicals, and not anything about mineral oil per se, that was at issue. Id. at 139.
192 Id. at 146.
193 Id. at 145. The district court explicitly mentioned statistical significance in quoting from this study. See Joiner v. Gen. Elec. Co., 864 F. Supp. 1310, 1324 (1994) (“The numbers were small, the value of the risk estimate was not statistically significant, and such risk had never been suggested before.”).
significant and the authors of the study did not suggest a link between the increase in lung cancer deaths and the exposure to PCB's.\textsuperscript{194}

To see the problem with the \textit{Joiner} opinion as to these studies, just ask: Would these two studies have provided an adequate basis for an expert to find causation if the results \textit{had} been statistically significant? If so, then there couldn’t have been anything unreliable about the studies in \textit{Daubert}'s sense, so there would be no meaningful analytical gap in relying on them in forming an opinion as to causation. As to the two PCB studies in \textit{Joiner}, neither the Supreme Court nor the trial court raised an issue that can fairly be regarded as an analytical gap.

But here’s one: arguably it is unreasonable for an expert to simply assume that it is possible to understand the cause of cancer for the plaintiff in \textit{Joiner} from the first two studies at issue. Mr. Joiner was an electrician who would “occasionally” have PCB chemicals “splash . . . into his eyes and mouth.”\textsuperscript{195}

The two studies in question involved (i) data on workers in an electrical capacitor plant, who for years worked with \textit{substantial quantities} of PCB chemicals\textsuperscript{196} (the first study) and (ii) data on workers in a plant that actually \textit{produced} PCB chemicals (the second one).\textsuperscript{197} The differences between the types, frequency, and duration of exposure suggest the possibility that these studies might have been quite unreliable as a basis for concluding that Mr. Joiner’s lung cancer was caused by PCB exposure.\textsuperscript{198}

Statistical significance at more demanding significance levels is helpful, where it is, because it indicates that the evidence is stronger—i.e., more probative. But if there is no “analytical” gap in using evidence \(E\) when it strongly points in the direction of the expert’s conclusion, how can there be an “analytical” gap in using the same \(E\) when it points in exactly the same direction, just less probatively?

A final point about the language in \textit{Joiner} and statistical significance is that it is under-specified. It is vacuous to say only that a result “is statistically insignificant.” Statistical significance is necessarily defined by reference to a significance level. That is crucial because by the very nature of statistical significance, every estimate is statistically significant at \textit{some} significance level. Suppose the authors of the first study in \textit{Joiner}, the one involving the workers

\textsuperscript{194} \textit{Joiner}, 522 U.S. at 145.
\textsuperscript{195} \textit{Id.} at 139.
\textsuperscript{196} Pier Alberto Bertazzi et al., \textit{Cancer Mortality of Capacitor Manufacturing Workers}, 11 AM. J. INDUS. MED. 165, 167 (1987) (stating that peak annual PCB consumption in the plant was 250 tons).
\textsuperscript{198} For a wide-ranging discussion of \textit{Joiner} from the perspective of philosophy of science, see Susan Haack, \textit{An Epistemologist in the Bramble-Bush: At the Supreme Court with Mr. Joiner}, 26 J. HEALTH POL’L, POL’Y, & L. 217 (2001).
in the capacitor plant, had declared: “our results are statistically significant at the 49.99% level.” That statement would have been true given the study’s findings, actually, because the number of lung cancer deaths observed among those exposed to PCB chemicals exceeded the number that were expected.

Thus, the inadmissibility of expert testimony founded on this study in Joiner cannot be traced to some mythic, wholesale absence of statistical significance. It must instead be because the results were not statistically significant at a sufficiently demanding significance level. So as a matter of definition, there is no way around the conclusion that Joiner’s language as to the two PCB studies amounts to rejecting them because they yielded insufficiently strong evidence.

But one can hardly conclude that Joiner licenses district courts to require whatever strength of evidence they like while marching under the banner of Rule 702 reliability. That would be a judicial hiding of elephants in mouseholes: surely if the Court were selecting a particular level of statistical significance for use in litigation, it would have done so with more than an offhand mention or two. For another, it’s doubtful that when he wrote Joiner, Chief Justice Rehnquist had a particularly clear grasp of what statistical significance does and doesn’t mean.

One would have to pole vault over an analytical chasm to conclude that Joiner’s inexact language about statistical significance somehow announces a general common law policy in favor of—what? A 5 percent significance level? 10 percent? 1 percent? This indeterminacy is a clue that the Joiner Court had little awareness of the significant can of worms it popped open. And other Supreme Court cases that take a positive view of scholarly levels of statistical significance do so at most tentatively.

---

199 Cf. Whitman v. Am. Trucking Ass’n, 531 U.S. 457, 468 (2001) (Scalia, J.) (“Congress, we have held, does not alter the fundamental details of a regulatory scheme in vague terms or ancillary provisions—it does not, one might say, hide elephants in mouseholes.”).

200 On this front, see Chief Justice Rehnquist’s partial dissent from Daubert. It self-consciously evinces a lack of understanding of the concept of scientific falsifiability of hypotheses, which is closely related to the logic of null hypothesis significance testing. Daubert, 509 U.S. at 600 (Rehnquist, C.J., dissenting in part) (“I defer to no one in my confidence in federal judges; but I am at a loss to know what is meant when it is said that the scientific status of a theory depends on its ‘falsifiability,’ and I suspect some of them will be, too.”).


unanimously rejected the idea that it is always necessary for estimation evidence to meet demanding significance levels to prove causation.\footnote{Id. The Court noted that “[a] lack of statistically significant data does not mean that medical experts have no reliable basis for inferring a causal link between a drug and adverse events.” Id. at 40. In fact, the Court noted, relying on the Amici Brief for Medical Researchers, “medical experts rely on other evidence to establish an inference of causation . . . ‘Medical professionals and researchers do not limit the data they consider to the results of randomized clinical trials or to statistically significant evidence.’” Id. at 40–41. Similarly, “[t]he FDA . . . does not limit the evidence it considers for purposes of assessing causation and taking regulatory action to statistically significant data.” Id. at 41.}

In sum, although Joiner’s embrace of the district court’s reasons for rejecting the two studies that involved PCB chemicals is inconsistent with my argument in this Article and with Daubert, it also sits poorly with any reasonable understanding of Joiner’s otherwise sensible language about analytical gaps and ipse dixits. One can save both the result in Joiner and the overall reasoning without understanding Joiner to say that a district court is within its discretion to reject expert testimony merely because it is founded on results that are statistically insignificant at conventional significance levels.

To put it another way, there are only three alternatives. One is to treat Joiner as announcing a strength-of-evidence standard of proof in the most sottó of voce. I reject this option, and so should you. The second alternative is to treat Joiner’s language about statistical significance as dicta. And the third is to view Joiner as consistent with admitting expert testimony about estimation evidence so long as the expert testifies that estimation evidence is statistically significant at some level. Both the second and third understandings of Joiner on the statistical significance point are consistent with the fundamental default rule. The last approach was recently taken by a district court hearing an antitrust case in which the expert economist testified that he had adopted the 50% significance level (he gave reasons) and that the estimation evidence was statistically significant at that level.\footnote{See In re High-Tech Emp. Antitrust Litig., No. 11-cv-02509-LHK, 2014 WL 1351040, at *12 (N.D. Cal., Apr. 4, 2014); id. at n.25 (“Dr. Leamer may testify . . . that the fact that his alternative conduct regression model’s conduct coefficients pass the 50% level ‘suggests that it is more likely than not that the compensation of employees were decreased during the period of the agreements.’”); see also In re Photochromic Lens Antitrust Litig., No. 8:10-cv-00984-T-27EA, 2014 WL 1338605, at *26 (M.D. Fla. Apr. 3, 2014) (approving the use of a significance level of .50 because the testifying expert economist justified the approach on the basis of Type II error probabilities, and also because the opposing expert failed to justify rejecting that approach except because it was “simply out of bounds of what economists do”).}

c. Kumho Tire

Kumho Tire explains that experts must “employ[] in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in
the relevant field.”205 It might seem that this criterion disqualifies an expert from using a significance level of just below 50 percent in expert testimony, if that expert follows the common scholarly norm of using a more demanding level such as 5 percent in her scholarly activities.206 But such a determination would be inapt, because “employs . . . the same level of intellectual rigor” cannot reasonably mean “requires the same level of statistical significance.”207

Intellectual rigor demands consistent reasoning and attentiveness to contextual variation.208 It is intellectually rigorous for an expert viewing a fixed set of facts to provide different answers to different questions. The legal sufficiency question related to expert testimony in litigation is whether statistical results make the fact at issue more probable than not. That is the correct standard in expert testimony, even when the expert giving that testimony is a scholar who uses more demanding levels of statistical significance in her work intended for publication in scholarly journals.

Scholars operating in a scholarly capacity have the luxury of following social norms that require powerful statistical evidence before a null hypothesis can be rejected.209 That’s okay in scholarland, because scholars rarely if ever have to use their statistical results to make consequential decisions on short time.210 In fact, a common way to describe statistically insignificant results is that they “fail to reject the null hypothesis,” not that they prove the alternative hypothesis; the non-rejecting researcher finds herself in an intermediate state of doubt as to which hypothesis is true. That kind of epistemic luxury is alien to litigation. Courts owe no deference to such social conventions merely because some—or even many—scholars follow them.211 The level of statistical significance to be required is an instance in which, as Daubert put it, academic considerations simply do not line up with the judicial “project of reaching a quick, final, and binding legal judgment—often of great consequence—about a particular set of events in the past.”212

Consider a scholar who has written, in a scholarly journal, that a study meant to speak to injury causation by a drug was poorly conceived and executed, such that its quantitative results cannot be taken at face value. If this scholar comes to court and relies on that same study’s quantitative results,
then he is applying a different standard of intellectual rigor inside the courtroom than in his scholarly work. He has a *Kumho Tire* problem.

But what if all of that were reversed? Instead, the scholar has written that the study in question was flawlessly designed and perfectly executed. And he has written that although the injury rate is higher among those who took the drug than among those who didn't, the difference isn't statistically significant at the 5% level. If he then testifies in court that the statistical evidence is more consistent with injury causation than not, we should not understand *Kumho Tire* to bar his testimony for lack of intellectual rigor.

We can understand the issue here as involving the proper level of generality to ensure that experts use the same level of intellectual rigor for their courtroom and outside work. To insist that experts always use conventional hypothesis testing at conventional significance levels is to insist not that they rigorously use statistical methodology both inside and outside the courtroom, but rather that they always answer causation questions by reference to the same evidentiary standard. Yet the standard at issue in a civil lawsuit is different from the one scholars use when considering whether evidence is strong enough to constitute scientific knowledge. That can't make sense. Rather, when it is appropriate to do so, an expert who analyzes the same facts using the same methodology in an intellectually rigorous way sometimes will give different answers to different questions.

2. The Proper Focus in Gatekeeping is on Technical Implementation and Credibility for the Proper Purpose

Evidence law does properly cabin the use of statistical evidence. It establishes rigorous standards for the expert’s use of underlying statistical data, as well as for the collection and generation of that data. Applying these standards might properly lead a district court to exclude expert testimony in many settings.

Even if she followed Part I’s playbook to measure the strength of evidence, an expert’s testimony as to statistical evidence could still be unreliable in numerous ways. The expert might have committed mathematical errors in calculating the likelihood ratio or statistics based on it. He might have used a statistical model unrelated to the fact in question. He might have testified as to statistical results reported by others that suffer from either of those deficiencies. And, as my discussion of the PCB studies in *Joiner* suggested, the underlying study that generated the data on which the expert built his opinion might lack credibility for the proffered purpose.

To further illustrate the last point, I now consider two quite different cases and then offer some further general comments.
a. Credibility for the Proffered Purpose: Mayor of Philadelphia v. Educational Equality League

*Mayor of Philadelphia v. Educational Equality League*\(^{213}\) was a racial discrimination case brought in the early 1970s, long before *Daubert*, but it illustrates well two kinds of concerns at which gatekeeping should be directed. The plaintiff organization alleged the Mayor had violated the Equal Protection Clause by discriminating against blacks in making appointments to an advisory panel that nominated city School Board members.\(^{214}\) The panel consisted of thirteen members in total, of whom the Mayor had discretion to appoint four.\(^{215}\) Neither party introduced an expert to discuss statistics at the trial court,\(^{216}\) but the Court of Appeals concluded *sua sponte* that the small number of black members on the panel—either one or two out of thirteen over three appointment cycles between 1967 and 1971—was evidence of a statistical disparity considering that blacks made up 33.5% of the city’s population and 60% of the public school system.\(^{217}\)

The Supreme Court reversed.\(^{218}\) But it did not reject the use of statistical evidence; indeed, the Court took pains to state that “[s]tatistical analyses have served and will continue to serve an important role as one indirect indicator of racial discrimination in access to service on governmental bodies . . . .”\(^{219}\) Nor did the reversal hinge on the *sua sponte* nature of the lower court’s use of statistical reasoning. Rather, the basis for reversal as to the statistical evidence was that “the simplistic percentage comparisons undertaken by the Court of Appeals lack[ed] real meaning in the context of this case.”\(^{220}\) One reason was that the city charter required that

nine of the 13 seats [be] restricted to the highest ranking officers of designated categories of citywide organizations and institutions. . . . [T]his is not a case in which it can be assumed that all citizens are fungible for purposes of determining whether members of a particular class have been unlawfully excluded. At least with regard to nine seats on the Panel and assuming, *arguendo*, that percentage comparisons are meaningful in a case involving discretionary appointments, the relevant universe for comparison purposes consists of the highest ranking officers of the categories of organizations and

\(^{214}\) Id. at 609-10.
\(^{215}\) Id. at 607.
\(^{216}\) Id. at 620.
\(^{218}\) Mayor of Phila., 415 U.S. at 621, 629.
\(^{219}\) Id. at 620.
\(^{220}\) Id.
institutions specified in the city charter, not the population at large. The Court of Appeals overlooked this distinction.\footnote{Id. at 620-21 (emphasis added).}

So the Supreme Court believed that when the city charter requires nine panelists to be drawn from a particular population, it is that population, and not the city-wide population, that is at issue when discrimination is alleged. The statistical evidence under consideration—even if it had been the subject of good faith trial testimony from a competent expert—compared the composition of the panel to the overall city population, which was the wrong comparison group because the charter required panelists be drawn from a particular subpopulation of the city.\footnote{Id. at 614.}

In addition, the Court also held that the district court was right to worry about “the smallness of the sample presented by the 13-member Panel,” holding that the “Court of Appeals erred in failing to recognize the importance of this flaw in straight percentage comparisons.”\footnote{Id. at 620-621.} The Court’s concern about sample size did not spring from a belief that some statistic was statistically insignificant due to small sample size. Rather, the Court seems to have generically worried about the appropriateness of basing any conclusions on evidence from such a small number of panelists (of whom only four could be chosen on an entirely discretionary basis).\footnote{Here it is worth noting that in Title VII cases, the EEOC Four-Fifths Rule is an alternative metric for courts faced with statistical evidence. The Four-Fifths Rule will sometimes be satisfied when the formal statistical significance tests fail to reach conventional levels of significance such as 5%, in which case plaintiffs will argue for the Four-Fifths Rule and defendants for a rule of statistical significance. For a discussion of this pattern, see Jennifer L. Peresie, Toward a Coherent Test for Disparate Impact Discrimination, 84 IND. L.J. 773, 781-84 (2009).}

Mayor of Philadelphia thus provides an early illustration of the Supreme Court assessing statistical evidence for its fit with the question at bar, and finding it wanting. Whatever the other evidence in the case, there was little reason to regard the statistical evidence as offering a reliable way to explain the underrepresentation of blacks on the city panel.

b. Credibility for the Proffered Purpose: Merck v. Garza

Now consider the Texas Supreme Court case of Merck v. Garza.\footnote{347 S.W.3d 356 (Tex. 2011).} That was a product liability case alleging that Leonel Garza’s fatal heart attack was caused by his use of Vioxx. The plaintiffs, relatives of Mr. Garza, sought to introduce estimation evidence from a clinical trial showing that Vioxx caused a large increase in the relative risk of heart attack. The issue on appeal was
sufficiency rather than admissibility, but the court couched its discussion in terms of reliability just the same. Observing that Mr. Garza had taken “a much smaller dosage of Vioxx for much less time” than subjects in the clinical trial,226 the court stated that the clinical trial “study suggests nothing at all about significantly lesser exposure.”227 Thus while the “usage involved in a study need not match the claimant’s usage exactly,” the court found that the instant case failed the requirement that “the conditions of the study should be substantially similar to the claimant’s circumstances,”228 and the court held that the study was of no help to the plaintiffs.

Garza illustrates the fact that a study can be very credible for some purposes—the study in question was a randomized controlled trial—and still have too little credibility with respect to the question for which it is offered.229

c. Cherry Picking

Another basis for exclusion would involve cherry picking results when there are multiple statistical studies available. Rule 702(b) requires that an expert base her opinion on “sufficient facts or data.”230 The Committee Note to the 2000 Amendment to Rule 702 briefly observes that “sufficient” in this context means “quantitative” rather than “qualitative.”231 According to Wright & Miller, the point of this language is to ensure that experts take into account all the pertinent facts or data and not engage in cherry picking.232

Kumho Tire is also on point. An expert operating in her scholarly capacity would not ignore the results of studies she thought were well designed, well executed, and otherwise were credible for a particular purpose. Kumho Tire’s requirement that experts use the same “intellectual rigor” in the courtroom as in their scholarly activities thus requires that in forming opinions about statistical evidence, testifying experts adequately take into account all the reasonably applicable statistical evidence—not just whatever supports the position of the party proffering their testimony.

Another way to express the proscription on cherry picking is to observe that ignoring some of the credible statistical evidence when drawing

---

226 The clinical trial involved a dose of 50 mg, and a median duration of nine months, by comparison to a dose half as great for Mr. Garza, for only twenty-five days. Id. at 266.
227 Id.
228 Id.
229 The legal standard in Garza was substantially based on a Bendectin case from the same court a decade and a half earlier, Merrell Dow Pharm., Inc. v. Havner, 953 S.W.2d 706, 720 (Tex. 1997), in which the court had stressed that a plaintiff seeking to use epidemiological studies would have to show similarity of the same type the Garza court found wanting with respect to a clinical trial study. Garza, 347 S.W.3d at 262.
230 FED. R. EVID. 702(b).
231 FED. R. EVID. 702, committee notes to 2000 amendment.
232 WRIGHT & MILLER, supra note 109.
conclusions opens up Joiner’s analytical gap. An expert who discounts some of
the evidence can close the gap by giving good reasons to do so, but one who
cannot do so has a Joiner problem.

Even in the absence of Rule 702(b), Kumho Tire, or Joiner, Rule 403 would
provide additional reason to require experts to take into account all the
evidence. Testifying on the basis of a cherry-picked set of results is an
excellent way to mislead a jury.233

d. Summary

In sum, this Section has shown that the Federal Rules of Evidence and
the Daubert trilogy have plenty of work to do with respect to statistical
evidence. That work is properly directed at issues related to technical
implementation, to the credibility of statistical evidence for the proffered
purpose, and to cherry picking. Gatekeeping of testimony about statistical
evidence on these bases is not only acceptable, it is required under the Rules.
Practical reasoning helps courts here, as do the incentives litigants have as a
result of our adversarial process. Moreover, there is extensive and excellent
qualitative guidance about credibility in the Reference Manual on Scientific
Evidence published by the Federal Judicial Center.234

III. POLICY CONSIDERATIONS RELATED TO THE FUNDAMENTAL
DEFault RULE

This Part discusses normative issues implicated by the fundamental default
rule introduced in Part I. Section A discusses the administrability advantages
of the fundamental default rule when the preponderance standard applies.
Section B considers the impact on both the volume of litigation, and primary
behavior, of adopting my position. Finally, Section C discusses how Bayesian
hypothesis testing can be used to conceptualize the basis for (i) the Supreme
Court’s choice of the preponderance standard and (ii) alternative standards.

233 See FED. R. EVID. 403 ("The court may exclude relevant evidence if its probative value is
substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing
the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative
evidence." (emphasis added)).

234 FEDERAL JUDICIAL CENTER, RMSE, supra note 9. See, e.g., chapters on statistics (by
David H. Kaye and David A. Freedman), multiple regression (by Daniel L. Rubinfeld), and
epidemiology (by Michael D. Green, D. Michal Freedman, & Leon Gordis).
A. The Fundamental Default Rule Would Improve Administrability and Litigation Practice

There's not all that much to say about administrability considerations with my proposal, but what there is to say is good. I claim that the Rules imply a simple standard for the admissibility and sufficiency of estimation evidence for the questions at which they are directed: the evidence is generally acceptable when it supports the party offering it. It is difficult to conceive of a more easily administered rule than that (except a rule instructing judges to either never or always allow statistical evidence). So, administrability is a mark in favor of the default rule.

But the benefit of the rule doesn't stop there. Experience suggests that using conventional hypothesis testing at conventional significance levels has had its own share of administration challenges. Some cases have devolved into classic battlefields of the experts, with the proper significance level being the rough equivalent of Alsace-Lorraine. Methodological skirmishes likely would be less prevalent if significance levels were de-emphasized in favor of simply looking at which side the evidence supports.

In addition, the U.S. reports have plenty of opinions in which judges have butchered the meaning of statistical significance, wrongly equated it with posterior probabilities, or both. And even courts that do get this meaning right can't avoid the embarrassing fact that statistical significance at conventional significance levels does not answer the question that's actually at issue in the litigation.

235 See, e.g., Erica P. John Fund, Inc. v. Halliburton Co., 309 F.R.D. 251, 262 (N.D. Tex. 2015) (describing district court decision resolving disagreements between experts over, e.g., technical statistical issues related to multiple testing); In re Photochromic Lens Antitrust Litig., No. 8:10-cv-00984-T-27EA, 2014 WL 1338605, at *16-27 (M.D. Fla. Apr. 3, 2014) (describing plaintiff’s expert’s reasons for using the 50% significance level, noting that defendant’s expert “did not controvert those explanations” except to state that using the 50% significance level is “simply out of bounds of what economists do,” and ultimately finding for the plaintiff’s side on the ground that the “use of a 50% measure of statistical significance, by itself, is” not “sufficient justification for denying class certification”) (internal quotations omitted); In re Int'l Grp., Inc. Sec. Litig., 265 F.R.D. 157, 186-97 (S.D.N.Y. 2010), vacated and remanded, 689 F.3d 229 (2d Cir. 2012) (stating that the defendant’s expert “contended that 5% is the minimum level of statistical significance that conventional statistical methodology would accept,” noting that the plaintiff’s expert testified that the standard is at least 10% in the financial economics field, and ultimately finding for the defendant on the issue in question because “there is a distinction between reporting and drawing conclusions based on a 10% level of statistical significance”); Segar v. Civiletti, 598 F. Supp. 690, 701 (D.D.C. 1981), aff'd in part, vacated in part sub nom. Segar v. Smith, 738 F.2d 1249 (D.C. Cir. 1984) (noting that “[p]laintiffs’ experts consider a .10 level . . . to be statistically significant,” that “Defendants [sic] experts stated that anything above the .05 level was statistically insignificant,” and that because the “probative value of a study is affected by that study’s statistical significance . . . this Court accords minimal value to Plaintiffs’ promotions analysis . . .”).

236 See, e.g., Kaye & Freedman, supra note 9, at 252 n.103 (collecting cases).
Thus, Daubert hearings might be considerably different, and perhaps even less frequent, if the fundamental default rule were applied. Because scholarly levels of statistical significance would not be required to allow an expert’s testimony about statistical estimation evidence, it is likely that there would be many fewer arcane arguments over statistical details, which not all judges really understand. What arguments did occur at Daubert hearings would presumably turn on the issues I identified in Section II.C as the proper ones under Rule 702—namely, issues related to technical implementation and credibility for the proffered purpose.

Understanding these issues, and coming to a reasonable view about them, requires less specialized knowledge than does understanding the details of hypothesis testing. Moreover, credibility for the proffered purpose is an area where laypeople’s capacities might not be all that much less than experts’. For example, one doesn’t need a degree in applied statistics to understand the source and importance of the non-representativeness of the Literary Digest survey discussed in subsection II.B.2. Thus, there’s reason to believe that adopting the fundamental default rule would de-mystify Daubert hearings.

An additional question is how adopting the fundamental default rule would affect the set of studies on which experts rely in litigation. Because studies would no longer be considered uninformative merely because they don’t meet scholarly standards of statistical significance, one might think the set of studies available for use would expand. But under current practice, a study that doesn’t meet scholarly significance levels is helpful to the side wishing to argue there isn’t “an effect” (i.e., the defendant in my analysis above, e.g., with respect to the Lipitor litigation). Switching to the fundamental default rule just changes whom the study favors, not whether it should be considered (assuming it is otherwise appropriate).

At trial, an expert could be allowed to testify on the condition that the expert address the strength of statistical evidence during direct examination. And on cross-examination, a good trial attorney will be able to skewer any expert who overstates this strength. Attorneys could also ask testifying experts to describe the degree of confidence they hold concerning the conclusions they describe as reasonable. And cross-examiners could ask what assumptions, including about prior beliefs, the expert made in coming to their conclusions. And of course, attorneys could attack the technical implementation or credibility for the proffered purpose of the studies underlying experts’ testimony, and a jury is within its rights to find those requirements unmet. So even under the fundamental default rule, there is room for lawyers at trial to flesh out the details of statistical estimation evidence. In fact, it would be salutary to divert lawyer effort away from arguments about significance levels and the (ill-)fitting nature of conventional
hypothesis testing, and toward the question of how credible and how strong the statistical estimation evidence really is.

On the disadvantage side, adopting the default rule would increase the set of cases in which Rule 403 appropriately comes into play. That Rule allows judges to exclude relevant evidence if its “probative value is substantially outweighed” by various dangers, including misleading the jury.237 Using conventional hypothesis testing at demanding levels ensures that admitted statistical evidence is far stronger than necessary to meet the preponderance standard. By its nature, such evidence has quite high probative value, so Rule 403 dangers would have to be quite severe to “substantially outweigh[]” the probative value of this evidence. Switching to my default rule would mean that evidence with more marginal probative value would be admissible in the absence of Rule 403. So, Rule 403’s balancing test would be satisfied more frequently under my default rule. To the extent that a Rule 403 assessment doesn’t merely duplicate the Rule 702 analysis judges already do, that would temper the administrability gain from adopting the fundamental default rule, because we would have more Rule 403 fights as opposing parties argued that juries would be too easily persuaded by the supposed magic of statistics. But perhaps judges should not often exclude even weak statistical estimation evidence under Rule 403. Judge Posner writes: “fears that jurors are dazzled by evidence which involves explicit probability estimates, and so give such evidence more weight than a good Bayesian would do, appear to be unfounded; jurors appear to give statistical evidence less weight than [sic] they should.”238

My sense is that judges should rarely find that Rule 403 warrants excluding weak statistical evidence. As Daubert declares, the need to attack “shaky but admissible evidence” is not unknown to trial practice, and it is traditionally entrusted to opposing counsel’s “[v]igorous cross-examination” and “presentation of contrary evidence,” as well as the judge’s “careful instruction on the burden of proof.”239 Jurors need not understand all the ins and outs of statistical significance testing, or the generalized likelihood ratio, in order for these means to work.

Putting all this together, it seems likely that adopting the default rule would simplify litigation and make judges’ gatekeeping job easier. It would free them from the task of understanding and describing—or, in some unfortunate cases, from misunderstanding and mischaracterizing—the

237 FED. R. EVID. 403.
238 Posner, supra note 8, at 1511.
concept of statistical significance. For the same reasons, it would simplify the job of attorneys in cases that involve experts.240

B. The Quantity of Litigation Activity Might or Might Not Increase with the Fundamental Default Rule, but the Rule Likely Would Shift Bargaining Power to Plaintiffs

What about the possibility that the fundamental default rule would spark a dramatic increase in the volume of litigation? To ask whether one or another rule would lead to “too much” litigation is to invite the question of what the “right” amount of litigation is.

Writing in what functioned as dicta in a case involving due process and state-law civil commitment proceedings, the Supreme Court long ago gave one answer:

The standard [of proof] serves to allocate the risk of error between the litigants and to indicate the relative importance attached to the ultimate decision.

Generally speaking, the evolution of this area of the law has produced across a continuum three standards or levels of proof for different types of cases. At one end of the spectrum is the typical civil case involving a monetary dispute between private parties. Since society has a minimal concern with the outcome of such private suits, plaintiff’s burden of proof is a mere preponderance of the evidence. The litigants thus share the risk of error in roughly equal fashion.241

How should one conceptualize the relationship between the “risk of error” involved in litigation and the standard of proof? A common approach among evidence law scholars is to use decision theory. This approach assigns a cost, say, $C_D$, to mistakenly finding against the defendant, and another cost, say, $C_P$, to mistakenly finding against the plaintiff. In hypothesis testing terms, mistaken determinations against the defendant entail rejecting the null

240 An additional question is how adopting the fundamental default rule would affect the set of experts used in litigation. The current system presumably favors experts who can give simple-sounding explanations of the complex concepts underlying conventional hypothesis testing. If courts adopting the fundamental default rule required experts to explain Bayesian concepts, that would additionally favor experts who can explain that approach; these might be different experts from the first set, but the same type of skills would presumably be favored. Perhaps the biggest impact on the set of experts would be that experts comfortable with the Bayesian approach but not with conventional hypothesis testing would be ushered into court, and litigation more generally. That kind of diversity of opinion would better reflect the world of practicing applied statisticians, which could be a good thing.

hypothesis, so such mistakes are Type I errors (false positives). Mistaken determinations in favor of the plaintiff entail Type II errors (false negatives). Decision theory seeks to find a decision rule that minimizes total expected error costs. To do this, the following principles are helpful:

- The expected error cost associated with mistaken allowance of estimation evidence is the product of cost \( C_D \) and the probability of a Type I error (again, a Type I error in this context means finding for the plaintiff when the defendant’s position is the correct one).
- The expected error cost associated with mistaken rejection of estimation evidence is the product of cost \( C_P \) and the probability of a Type II error (again, a Type II error in this context means finding for the defendant when the plaintiff’s position is the correct one).

The Supreme Court’s declaration that litigants should equally share the risk of errors is often interpreted to mean that the two costs, \( C_D \) and \( C_P \), should be treated as equal. It can be shown that if a decision rule minimizes error costs, the ratio of cost \( C_D \) to cost \( C_P \) must equal the ratio of Type II error probability to Type I error probability (i.e., the ratio of false negatives to false positives). If the costs are equal, then the first ratio is one, so the decision rule must cause the Type I and Type II error probabilities to be equalized.

---

242 As Professor Frederick Schauer has pointed out to me in correspondence, “back in the old days, Type I and Type II errors were not labeled according to false negatives and false positives, but instead to ‘the error that we are most eager [or less eager] to avoid.’” Email from Frederick Schauer, Professor of Law, Univ. of Va. Sch. of Law, to author (July 2, 2019, 10:23 EST) (on file with author) (citing ERNEST KURNOW, GERALD J. GLASSER & FREDERICK R. OTTMAN, STATISTICS FOR BUSINESS DECISIONS 235 (1959)).

243 For treatments of this point in the legal domain, see, e.g., Michelle M. Burtis, Jonah B. Gelbach & Bruce H. Kobayashi, Error Costs, Legal Standards of Proof, and Statistical Significance, 35 SUP. CT. ECON. REV. 1, 11 (2018) (discussing an error-cost minimization criterion for optimality, showing that the preponderance standard meets it when the cost of Type I and Type II errors are equal, and noting that the Supreme Court’s dicta in In re Winship, 397 U.S. 358 (1970) “suggests such a weighting is appropriate in certain civil cases”); Cheng, supra note 50, at 1261 (stating that the legal system does not demonstrate a preference between finding erroneously for the plaintiff or defendant, with the implication that \( C_D \) and \( C_P \) should be equal); Michael S. Pardo, The Nature and Purpose of Evidence Theory, 66 VAND. L. REV. 547, 561 (2013) (citing In re Winship, 397 U.S. 358, 361 (1970), for the proposition that error costs in civil cases should be treated as equal); Id. (citing Grogan v. Garner, 498 U.S. 279, 286 (1991), for the proposition that “the preponderance standard ‘results in a roughly equal allocation of the risk of error’”). A possible alternative interpretation is that the two expected error costs are equal. On this understanding, “risk of error” means the cost of error weighted by the probability it occurs. I am unaware of any authors who have adopted this approach and shall not discuss it further.

244 See, e.g., Burtis, Gelbach & Kobayashi, supra note 243, at 11 (finding that the preponderance standard and optimal standard mathematically derived coincide “when the cost of Type I and Type II errors are equal”); Cheng & Pardo, supra note 51. Cheng, supra note 50, at 1260-61 (describing how, when error costs between plaintiff and defendant are equal, the likelihood of a Type I or Type II error are equal).
The rest of the analysis is somewhat complicated because of the role that prior beliefs play in determining the Type I and Type II error probabilities. But if we view the evidence through the eyes of the plaintiff’s most favorable juror, then under some conditions it can be shown that following the fundamental default rule leads to Type I and Type II error probabilities that are both 50 percent. Thus, the fundamental default rule can be viewed as the optimal decision rule given that error costs are treated as equal.

All of that said, many observers would question Addington’s statement that “society has a minimal concern with the outcome of . . . private suits.” A lower standard of evidence will induce more lawsuits, or credible threats of suit. That can be expected to induce changes in primary behavior. Whether these changes in primary behavior are socially beneficial depends on context and empirical facts outside the scope of any individual lawsuit.

This discussion of how the standard of evidence affects primary behavior naturally raises the question of what such effects mean for the optimal standard of evidence. There is a substantial theoretical law and economics literature that attempts to answer this question, taking into account the fact that the volume and quality of litigation is endogenous to the chosen standard. This literature tends to abstract from important doctrinal and institutional facts about litigation, most likely because these features complicate economic modelling. Perhaps as a result, this literature has had no apparent impact on procedural law (whether it has affected practice untransparently is a different question). A serious engagement with this

245 These conditions require a quite technical explanation and so are omitted from this Article.
246 There are discussions in the literature of what should happen when costs vary across case types. See, e.g., Daniel L. Rubinfeld, Econometrics in the Courtroom, 85 COLUM. L. REV. 1048, 1062–63 (1985).
248 See works cited in Eric L. Talley, Law, Economics, and the Burden(s) of Proof, in RESEARCH HANDBOOK ON THE ECONOMICS OF TORTS 325 (J. Arlen, ed. 2013) (stating that law and economics “scholars have begun to deliver insights into how the equilibrium properties of litigation are likely to have feedback effects on primary behavior of plaintiffs and defendants—the subject matter that is most closely associated with substantive law (rather than procedural rules”).
250 For example, Kaplow, Burden of Proof, supra note 249, ignores settlement and does not discuss the role of doctrines such as viewing the evidence in the light most favorable to the plaintiff.
251 Kaplow acknowledges that “in order to undertake the important constructive task of making sensible recommendations for system design, one would need context-specific empirical evidence that is not readily available.” Kaplow, Burden of Proof, supra note 249, at 751–52. An approach that requires unavailable data seems unlikely to have an impact on practice. A Westlaw search of sources
literature is beyond the scope of the present Article. But it is possible to imagine that a future engagement could yield useful heuristics for how the standard of evidence should be adjusted when estimation evidence is involved.

Beyond that observation, I shall take it as given that some observers oppose pretty much anything that increases litigation, while others take roughly the opposite position. Under either view, the normatively interesting question as to the volume of litigation is the simple positive one of whether adopting the fundamental default rule would increase or decrease the amount of litigation.

There are arguments to be made in each direction. Adopting the default rule would liberalize evidentiary standards compared to what many judges already do. That raises the expected value of potential plaintiffs’ claims in those disputes in which the liberalization of evidentiary standards would make a difference were the case litigated to judgment. Holding constant primary behavior, an increase in the expected value of potential plaintiffs’ claims can be expected to yield more actual plaintiffs—or anyway, more persons with a credible threat to litigate.

One possible result is that more lawsuits would be filed. Another is that defendants would settle more cases before filing. Thus, it is possible that adopting the default rule would reduce the amount of actual litigation, even as it increased the amount of settlement activity even more. In that event, the net result would be a bigger footprint for the civil justice system, broadly considered, even as the actual burden on the courts fell due to a reduction in the number of filed suits.252

Leaving aside the impact of adopting the fundamental default rule on the number of filed suits, we might also ask how it would affect litigation in those cases that would be filed regardless of the standard applied. Defendants could expect to win less often on summary judgment, holding constant the set of cases that have summary judgment motions. Some of these cases can be expected to settle after the denial of summary judgment. Most likely, though, there would be at least some additional post-summary judgment litigation activity. But it is also possible that there would be a drop in the number of Daubert and summary judgment motions courts must resolve, because the default rule has brighter lines than current practice. And under the fundamental default rule, at least some cases that would have Daubert and

---

252 A third possibility is that prospective defendants would adjust their primary behavior to reduce the chances that they find themselves faced with credible threats to litigate. This is the domain of the optimal-standard law and economics literature discussed just above.
summary judgment motions under current practice would instead settle before those motions are filed, reducing the burden on the courts.

Thus, adopting the default rule would have a complex set of effects on the amount and nature of litigation activity, as could be expected from any change in important procedural rules.\textsuperscript{253} What can be said for sure is that, holding primary activity constant, adopting the default rule would shift bargaining power toward those likely, under current practice, to have trouble getting statistical evidence admitted and accepted as legally sufficient. As a result, plaintiffs’ bargaining power would increase, and there would likely be a net transfer of wealth to them from defendants.

For precisely this reason, potential defendants might be expected to change their primary behavior. Employers would be likely to change HR practices to reduce the circumstances under which statistical evidence would point against them. Manufacturers, including drug and device makers, would take greater precautions to reduce the chances that credible studies would find their products are associated with injuries. And so on.

The prospect of the primary behavior effects just described will be cheered by some and derided by others. Neither is in itself a reason to support adopting the default rule. As discussed above, optimal evidentiary rules involve the balancing of desirable deterrence and the unfortunate chilling of desirable behavior.\textsuperscript{254}

IV. FEDERAL COURTS COULD USE COMMON LAW POWERS TO ADJUST THE EVIDENTIARY STANDARD IN CASES BASED ON FEDERAL LAW

We have seen that gatekeeping with conventional hypothesis testing at conventional significance levels erects a more demanding standard of proof than that required by the Federal Rules, as understood through associated case law. This implies that current gatekeeping practice generally is inconsistent with the Rules to the extent that courts require evidence stronger than whatever would support the plaintiff’s case.\textsuperscript{255}

But judicial practice is not necessarily unlawful merely because it is without the Federal Rules. Here it is helpful to recall that law comes in procedural and substantive flavors. The line between the two is famously

\textsuperscript{253} See Jonah B. Gelbach, \textit{Locking the Doors to Discovery? Assessing the Effects of Twombly and Iqbal on Access to Discovery}, 121 YALE L.J. 2270, 2280-21 (2012) (discussing predictable changes in litigation behavior following the implementation of \textit{Twombly} and \textit{Iqbal} and providing empirical evidence that at least some litigation behavior did change).

\textsuperscript{254} See supra note 249 and related citations.

\textsuperscript{255} Case-by-case review for reliability, under Rule 702, and to avoid jury confusion, under Rule 403, are still appropriate, as I discussed in Part VII.
fuzzy; just as it’s tough to eat one flavor at a time in a cone of chocolate-and-vanilla swirl, some legal propositions have one foot in procedure and the other in substance. Care is needed to avoid too facile a set of conclusions about the judicial development of evidentiary standards.

A first proposition is that the Federal Rules of Evidence were originally enacted as a statute. Thus, the policies originally embodied in them are backed with the full force of Congress’s constitutional power to legislate. A second proposition is that the Federal Rules of Civil Procedure were promulgated and have been amended pursuant to the Rules Enabling Act of 1934.\(^{256}\) The Enabling Act also gives the Supreme Court the power to amend the Federal Rules of Evidence,\(^{257}\) which it has done from time to time. For example, the original text of Rule 702 mentioned nothing about reliability;\(^{258}\) its starring role didn’t begin until the Supreme Court promulgated the 2000 Amendment through the Rules Enabling Act process.\(^{259}\) Thus, any form of gatekeeping over and above the substantive content of the original Rule 702 is a child of either the Rules Enabling Act process or the common law powers of federal judges (or both).\(^{260}\)

The Enabling Act famously mandates that “[S]uch rules shall not abridge, enlarge or modify any substantive right.”\(^{261}\) That might appear to mean that neither the Federal Rules of Civil Procedure nor amendments to the Federal Rules of Evidence may be understood to affect the strength-of-evidence standard required for proof.\(^{262}\)


\(^{257}\) Id. (granting the Supreme Court similar powers with respect to the rules of evidence).

\(^{258}\) The original text of Rule 702 was “If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise.” Pub. L. No. 93–595, 88 Stat 1926, 1937 (1975).

\(^{259}\) See George Fisher, Evidence 803 (2013) (explaining that subdivisions (c) and (d) of Rule 702, which is where the word “reliable” appears, were added after Daubert as part of the 2000 amendment).

\(^{260}\) That is another basis for my claim that neither Daubert nor its subsequent textual incorporation in Rule 702 lawfully can be the source of a transsubstantive standard of proof that requires statistical significance at conventional significance levels. Either that requirement existed previously in Rule 702—an unlikely proposition—or it is illegitimate.


In assessing that understanding, it will help to distinguish two types of litigation claims. The first type includes those claims for which the law defining the claim engenders no special consideration of the standard of proof that a plaintiff’s evidence must meet. A garden variety contract claim would fit in this group, as might a claim brought under, say, the Telephone Consumer Protection Act. Federal courts adjudicating such claims apply the preponderance standard thanks to the Supreme Court’s default rule. Of course these courts must also apply the various sets of Federal Rules, but as I have argued, those Rules do not alter the substantive standard of evidence. Accordingly, lower federal courts have no warrant to hold statistical evidence to more demanding proof requirements than the preponderance standard where it applies. Requiring conventional hypothesis testing at conventional significance levels in such cases is inconsistent with the law of evidence.

The second type of claims includes cases for which substantive law—often, the law creating or otherwise shaping the claim—embodies a particular set of evidentiary standards of proof. Consider again the Texas Supreme Court’s Havner case. In that Bendectin-related tort case, the Court held that tort plaintiffs seeking to use statistical evidence must present at least two studies that support their contention. This is a statement about the conditions under which the elements of a tort claim are proved. Erie R.R. Co. v. Tompkins and related cases tell us that if the effect of the Texas Supreme Court’s decision is to reject the preponderance standard, then it must be state tort law, and not the Supreme Court’s common law policy in favor of the preponderance standard, that governs. Another example, this time from a claim rooted in federal law, is the clear and convincing evidence standard that applies when a public figure brings a libel action.

To understand the appropriate role of federal common law powers in this sphere, it is useful to separately consider claims that arise solely from state law and those that arise from federal law. I treat these two types of claims in the next two Sections.

---

264 See discussion in text at supra note 74 as well as cases cited in that note.
265 See Merrell Dow Pharm., Inc. v. Havner, 953 S.W.2d 706, 727 (Tex. 1997) (“[i]f scientific methodology is followed, a single study would not be viewed as indicating that it is ‘more probable than not’ that an association exists”).
266 304 U.S. 64 (1938).
267 New York Times Co. v. Sullivan, 376 U.S. 254, 285-86 (1964) (invoking a standard of “convincing clarity,” which has been treated subsequently as equivalent to the “clear and convincing” standard).
A. In Cases Based on State Law Claims, the State’s Law as to Statistical Estimation Evidence Should Control in Federal Court

Federal courts have been bound since Erie to interpret the Rules of Decision Act to respect the substance of state common law claims heard in federal court, except when valid federal law preempts that substance.

Consider a case with all the basic facts of Havner, but assume it is heard in federal rather than state court, after Havner; this essentially happened in Cano v. Everest Minerals Corp, a case heard in the United States District Court for the Western District of Texas. Havner had announced a rule that causation in a product liability suit cannot be proved with only a single study providing statistical evidence. Subsequently, the Texas Supreme Court described Havner as requiring that its “standards of reliability are met in at least two properly designed studies.” These standards require that the studies (1) report an estimated relative risk above 2.0 and (2) are able to reject the null hypothesis of a true relative risk of 1.0 at the 95% confidence level. Accordingly, the federal courts must apply this rule unless valid federal law preempts it.

And federal law does not preempt that rule. Texas law defines what showing is needed to establish the fact at which the statistical evidence is directed, namely injury causation. Havner’s required showing establishes a greater-than-preponderance standard of proof under Texas law. If for some reason the U.S. Constitution mandated the preponderance standard in federal civil litigation, then by dint of the Supremacy Clause, federal courts constitutionally would have to apply the preponderance standard to all Texas tort claims. If instead a federal statute mandated that standard, then the question would be whether Congress could lawfully enact such a pre-empting statute.

269 362 F. Supp. 2d 814, 821 (W.D. Tex. 2005) (finding that because Havner dealt with the legal sufficiency of evidence, it is a substantive law issue, and thus should be applied.).
270 Havner, 953 S.W.2d at 717.
272 Havner, 953 S.W.2d at 718-19, 723-24. Relative risk is the ratio of the rate of occurrence of an outcome in one group to another. For example, in the ASCOT-LLA study, the relative risk is 3.0% divided by 2.6%, or about 1.15. Sever, supra note 26, at 1153.
273 Under the most expansive reading of Erie, Congress lacks this power. Erie, 304 U.S. at 78 (“Congress has no power to declare substantive rules of common law applicable in a State whether they be local in their nature or ‘general;’ be they commercial law or a part of the law of torts.”). Of course, if the statute were limited to tort suits involving interstate commerce, then subject to clear statement rules and the like, it would be constitutional. (Query whether the duty of care owed a trespasser along railroad tracks, at issue in Erie, doesn’t also implicate interstate commerce for at least some track lines.)
regulatory law) actually does such a thing. And certainly neither set of Federal Rules at issue here purports to define a substantive standard of proof.

Thus, if Havner’s standard were to be preempted in federal court, it would have to be because judge-made law requires federal courts to apply the preponderance standard. But such a practice would spark forum shopping in search of pro-plaintiff outcomes, implicating the core concern of Guaranty Trust Co. v. York’s reworking of Erie.274 Accordingly, the general federal court practice of applying the preponderance standard for determining legal sufficiency must give way when a case within Havner’s domain is heard in federal court.275

What if the Federal Rules did say the preponderance standard must govern as to elements of substantive law? If such a provision had been legislated by Congress, the only pertinent question would be the provision’s constitutionality. Of course a constitutional statute would preempt thanks to the Supremacy Clause.276 But no Federal Rule—not Rule 702 nor any other—actually contains a standard of proof as to the elements of the substantive law in a case.

So what should happen as to the admissibility of expert testimony in a federal court case where legal sufficiency is controlled by Havner? The District Court in Cano v. Everest Minerals Corp.277 got it exactly right:

whether expert testimony will assist the trier of fact is governed in part by whether the testimony is relevant to the plaintiff’s burden of proof under the

---

274 Guaranty Trust Co. v. York, 326 U.S. 99, 112 (1945) (holding that in diversity cases, the outcome should be substantially the same in federal court as it would be in state court).

275 That is the position taken in Cano v. Everest Minerals Corp., 362 F. Supp. 2d 814, 821-22 (W.D. Tex. 2005) (concluding that because it was an issue under Texas substantive law, “that Havner controls the issue of what evidence is required to establish causation in a toxic tort case”). Accord Wells v. SmithKline Beecham Corp., No. A-06-CA-126-LY, 2009 WL 364303, at *8 (W.D. Tex. Feb. 18, 2009) (holding that “[t]his Court concludes that Havner establishes substantive Texas law on a plaintiff’s causation burden of proof.”). A three-judge panel of the Fifth Circuit addressed this issue, but after assuming without deciding that Havner applied, it determined that the Havner standard was met, and in any event its opinion was subsequently vacated. Bartley v. Euclid, Inc., 158 F.3d 261, 272-73 (5th Cir. 1998), vacated, 169 F.3d 215 (5th Cir. 1999).

276 The same would not be true if Daubert or an amendment to Rule 702 promulgated by the Supreme Court through the REA process were the source of a conflict. In the case of Daubert, that would amount to federal common law as to the area reserved to states by the Rules of Decision Act, per Erie. Although Hanna v. Plumer, 380 U.S. 460 (1965), gives the Supreme Court substantial leeway through the REA process, it is clear there is a limit, including with respect to admissibility of evidence. Tyson Foods v. Bouaphakeo, 136 S. Ct. 1036, 1046 (2016) (“[W]here representative evidence is relevant in proving a plaintiff’s individual claim, that evidence cannot be deemed improper merely because the claim is brought on behalf of a class. To so hold would ignore the Rules Enabling Act’s pellucid instruction that use of the class device cannot ‘abridge . . . any substantive right.’”). Increasing the standard of proof via an amendment to Rule 702 pursuant to the REA would amount to an alteration of substantive rights, therefore contravening 28 U.S.C. § 2072(b).

substantive law, and testimony that will not assist the trier of fact by advancing an element of the plaintiff’s case should be excluded.278

Consequently, the Cano court concluded, “Havner controls the issue of what evidence is required to establish causation in a toxic tort case and therefore what evidence is relevant.”279 Even though it holds that Havner vitiates the preponderance standard as to statistical evidence, the Cano court does not say Rule 702 fails to apply. Indeed, this court applied Rule 702(a)’s helpfulness requirement and determined that evidence should be excluded for failing to satisfy it.280 Thus, it is possible, and, I think, best, to read Cano as accommodating both the substantive law set forth by the Texas Supreme Court in Havner and binding federal evidence law set forth in the Rules.

Another way to put it is that the Federal Rules of Evidence do not exist in a vacuum—they must be understood with reference to the governing substantive law.281 Expert testimony based on studies that cannot meet Havner’s requirements does not make it more probable that Havner’s requirements are met. Consequently, such evidence cannot help the trier of fact. In other words, Havner need not directly control the issue of admissibility to affect what counts as helpful to the trier of fact.

My analysis of the Havner example shows that if state law imposes a standard of proof requiring conventional hypothesis testing at conventional significance levels, then federal courts hearing state claims are bound to respect that standard in their own legal sufficiency determinations. Federal courts must still follow the Federal Rules of Evidence for determining admissibility. But that determination will be informed by what the substantive law requires for a party to prevail on the merits. That is as it should be in our small “f” federal system, in which states define important areas of governing law, at least where that law has not been displaced via constitutionally valid, and therefore supreme, federal law.282

278 Id. at 822 (citing Judge Kozinski’s opinion on remand in Daubert v. Merrell Dow Pharm., 43 F.3d 1311, 1320 (9th Cir. 1995) (“In assessing whether the proffered expert testimony ‘will assist the trier of fact’ in resolving this issue, we must look to the governing substantive standard . . . ”)).
279 Id.
280 Id. at 858.
281 For a qualitatively similar position about the connection between evidence law, substantive law, and civil procedure in the context of class certification when statistical evidence is involved, see Jonah B. Gelbach, The Triangle of Law and the Role of Evidence in Class Action Litigation, 165 U. PA. L. REV. 1807 (2017).
282 U.S. CONST. art. VI, § 2 (“[T]he laws of the United States which shall be made in pursuance [of the Constitution] . . . shall be the supreme law of the land”).
B. The Supreme Court Has Common Law Powers to Alter the Standard of Evidence for Federal Law Claims, but It Should Use Them transparently Rather Than Characterizing These Powers in Terms of the Federal Rules of Evidence

Civil claims under federal law are generally governed by the standard Congress sets, if it has set one.283 Congress has done so in many areas,284 but it has also left the standard of proof unstated in many statutes. When a claim arises from a federal statute that specifies no standard of proof, courts must step in. As discussed above, the Supreme Court’s default rule has been the preponderance standard.285 The Court has explained that “[a]ny other standard expresses a preference for one side’s interests.”286 Accordingly, it has self-consciously embraced elevated standards of proof only when “particularly important individual interests or rights are at stake.”287 Most such examples involve Fourteenth Amendment Due Process as applied to state law actions.288

The practice of raising the standard of proof when statistical evidence is involved does not fit this pattern. Liability related to statistical evidence may generate real public policy concerns, but it does not generally create a situation in which “particularly important individual interests or rights are at stake.”289 Justice Scalia’s opinion in Wal-Mart Stores, Inc. v. Dukes290 might be misread to the contrary. In a part of his opinion in which he spoke for a unanimous court, he emphasized that a defendant has a right to mount any

283 See Microsoft Corp. v. I4I Ltd. P’ship, 564 U.S. 91, 91 (2011) (“Where Congress has prescribed the governing standard of proof, its choice generally controls.”).
284 See Grogan v. Garner, 498 U.S. 279, 288 (1991) (collecting examples from the federal statutes in which Congress announced the preponderance standard for various forms of fraud, including, inter alia, those brought under the False Claims Act and those involving Medicare and Medicaid fraud).
287 Grogan, 498 U.S. at 286 (quoting Herman & MacLean, 459 U.S. at 389-390).
288 E.g., Herman & MacLean, 459 U.S. at 389 (citing Santosky v. Kramer, 455 U.S. 745 (1982), and noting that Santosky proceeded to terminate parental rights); Addington, 421 U.S. at 425-27 (involuntary commitment proceeding); Woody v. Immigration & Naturalization Serv., 385 U.S. 276, 285-86 (1966) (deportation). There are exceptions. One is patent validity. See Radio Corp. of Am. v. Radio Eng’g Labs., 293 U.S. 1, 8 (1934) (subsequently held by the Supreme Court to be codified by 35 U.S.C. § 282 (2018)) (noting that there is a presumption of validity of patents, and “one otherwise an infringer who assails the validity of a patent fair upon its face bears a heavy burden of persuasion, and fails unless his evidence has more than a dubious preponderance.”); Microsoft Corp., 564 U.S. at 104-05 (noting the same presumption of validity). Another, somewhat obscure, example is union bargaining with multiple employers. See United Mine Workers of Am. v. Pennington, 381 U.S. 657, 665 (1965) (holding that a union’s normal antitrust exemption is vitiated when it is “clearly shown that [the union] has agreed with one set of employers to impose a certain wage scale on other bargaining units.”).
289 Herman & MacLean, 459 U.S. at 389.
defense it has to a claim. He then described the district court’s trial plan as “Trial by Formula,” in which

A sample set of the class members would be selected, as to whom liability for sex discrimination and the backpay owing as a result would be determined in depositions supervised by a master. The percentage of claims determined to be valid would then be applied to the entire remaining class, and the number of (presumptively) valid claims thus derived would be multiplied by the average backpay award in the sample set to arrive at the entire class recovery—without further individualized proceedings. We disapprove that novel project. Because the Rules Enabling Act forbids interpreting Rule 23 to “abridge, enlarge or modify any substantive right,” a class cannot be certified on the premise that Wal–Mart will not be entitled to litigate its statutory defenses to individual claims.291

Two things are clear from this passage. First, what was at issue was not estimation evidence as it is usually understood. The statistical information here would have involved applying information about the judgments in a small set of cases to the adjudication of claims of other class members. But leaving aside traditional issue preclusion, it is difficult to see how the master’s determinations would even satisfy Rule 401’s relevance definition with respect to the non-sampled cases. Second, the modification of the substantive right at issue has nothing to do with the standard of proof. So Dukes cannot possibly stand for any principle about especially important individual rights affected by ordinary estimation evidence.

A third important fact about Dukes is that the majority’s disposition of the case was in important ways a holding about substantive Title VII law as such. In the part of Dukes that garnered only a bare majority’s support, the Court forbade class litigation to attack Wal–Mart’s corporate policy of allowing store-level managers the discretion to make various employment decisions.292 In so doing, the Court overtly relied on Federal Rule of Civil Procedure 23(a)(2)’s commonality requirement, declaring that store-level discretion “is just the opposite of a uniform employment practice that would provide the commonality needed for a class action.”293 But as Professor Tobias Barrington Wolff has explained, this holding says much more about substantive Title VII law than it does about what’s needed for commonality under Rule 23(a)(2).294

291 Id. at 367 (citations omitted).
292 Id. at 355-60.
293 Id. at 355.
294 This is laid bare by imagining that “the Court had found that a company-wide policy of reposing discretion in store-level managers could support a Title VII injunction because of its capacity to impose a disparate impact upon women, regardless of how that policy plays out in particular stores.” Tobias Barrington Wolff, Managerial Judging and Substantive Law, 90 WASH. U.
Thus *Dukes* is a useful prism for present purposes because it reflects how substantive law can get made in what looks like a procedural case. When federal courts impose an elevated standard of proof as to statistical evidence, they are doing what the *Dukes* Court did—making substantive law. That is so even if they say they are merely following federal evidence law.

Now, as I noted above, jacking up the standard of proof is something federal courts traditionally have done only when especially important individual interests ride on the outcome. Critics of the fundamental default rule will predictably challenge it on the ground that adopting it would bring substantial economic costs. And they may be right. But if the Supreme Court is worried about such baleful effects, then rule-of-law values counsel that it should not hide behind evidence law. Instead, the Court should be transparent that it is using its common law powers—where they exist—to make new law.

Does the Supreme Court have such common law powers in important litigation areas? It acts as if it does. We have already seen that *Dukes* is in part a determination about the substantive law of Title VII, and it certainly reads like common law. As Professor Wolff explains, in *Dukes* (and other Title VII cases),

> the court has taken portions of a regulatory statute that do not specify the methods of evaluating proof or administering remedies and set forth a body of judge-made law designed to carry into effect the express provisions of the statute and the policies underlying them. . . . [It]he rulings . . . also constitute affirmative statements of policy by the federal courts, making substantive decisions within the framework Congress set forth about the balance between reasonable opportunities for plaintiff recovery, on the one hand, and protection of defendants from unwarranted liability or settlement pressure, on the other.\(^{295}\)

If the Supreme Court can do all that, why can’t it determine the appropriate balance between “opportunities for plaintiff recovery. . . and protection of defendants” when the issue is what standard of proof to apply for statistical evidence? I think it can. In fact, it already has, in the line of cases, discussed above, that set up the preponderance standard as a default rule in the first place.

If the Court is to modify the default rule of preponderance, or make exceptions in specific substantive areas, it should state clearly and forthrightly the policy concerns that motivate it. That would improve democratic accountability by allowing the public and Congress a basis to evaluate the

\(^{295}\) *Id.* at 1044.
desirability of whatever rule the Court announces, facilitating legislative action if Congress or the public think the Court has gotten it wrong. In addition, explaining the substantive basis for policies raising the standard of proof for statistical evidence would promote rule-of-law values such as predictability and clarity—something that is in short supply in the federal courts when statistical estimation evidence is at issue.

CONCLUSION

In this Article I construct a theory of how federal courts should treat statistical estimation evidence at key pre-trial moments in civil litigation. Using black-letter law and the theory of statistical estimation together, I show that the Federal Rules require much more liberal treatment of plaintiffs’ statistical evidence than much current practice reflects. When statistical evidence fits well with the litigation question at which it is directed, the fundamental default rule of statistical estimation evidence holds that such evidence generally is admissible and legally sufficient for the proposition to which it speaks whenever it points in the direction of the party offering it. Using the fundamental default rule will also be much easier to administer than the status quo, alleviating a substantial amount of confusing and confused writing about probability theory by judges and attorneys who lack the expertise to do it well.

Despite my conclusion that present applications of Rule 702 are too stingy with respect to statistical evidence, the Supreme Court can use federal common law policy making powers—where it has them—to imposes elevated standards of proof in cases where they are warranted. When it deploys these powers, the Supreme Court should do so overtly in order to facilitate legislative responses and transparency. And of course, Congress and state bodies with legislative powers are available, too.

APPENDIX: MULTIPLE STUDIES

There is a simple way to handle multiple studies. Let $\theta$ be the true effect size whose value is the object of statistical estimation. Suppose we have data from $S$ studies that are thought to be reliable. Let $\hat{\theta}_s$ be the estimator of effect size from study $s$. Assume that each estimator is consistent and asymptotically normal. Different studies’ estimates will have different estimated standard errors due to differences in sample sizes and possibly other conditions such as the set of covariates available for use. So we must find a way to account for variation in $\sigma_s$, the standard error for study $s$.

When the various studies’ estimators are all statistically independent of each other, it can be shown that the weighted average of them that has
minimum asymptotic variance is $\bar{\theta}_S \equiv \sum_{s=1}^{S} \hat{\theta}_s$, where the weight $\hat{w}_S \equiv \frac{\sigma^{-2}_{-s}}{\sum_{s=1}^{S} \sigma^{-2}_{-s}}$. In words, the denominator of the weight is the sum of the inverses of the estimated variances of the $S$ available estimators. The numerator of the weight is the inverse estimated variance of estimator $s$. Assuming each estimator $\hat{\theta}_s$ is consistent, each has probability limit equal to the true value $\theta$. This means the probability limit of $\bar{\theta}_S$ is a weighted average of the constant $\theta$, so $\bar{\theta}_S$ also has probability limit equal to $\theta$.

Because each $\hat{\theta}_s$ is asymptotically normal, a weighted average of them is asymptotically normal as well. The variance of a weighted sum of independent estimators is the same weighted sum of the individual variances, so it is approximately true that $V(\bar{\theta}_S) = \sum_{s=1}^{S} (\text{plim } \hat{w}_S)^2 \sigma^2_s = (\sum_{s=1}^{S} \sigma^{-2}_s)^{-1}$, because the probability limit of the estimated variance $\hat{\sigma}^2_s$ is $\sigma^2_s$ and the probability limit of $\hat{w}_S$ is $\frac{\sigma^{-2}_{-s}}{\sum_{s=1}^{S} \sigma^{-2}_s}$.

Now let $\bar{T}_S = \frac{\bar{\theta}_S}{(\sum_{s=1}^{S} \hat{\sigma}^{-2}_s)^{-1/2}}$. This is the ratio of an asymptotically normal statistic to a consistent estimator of the square-root of its variance, so its limiting distribution is standard normal. Accordingly, we may treat the statistic $\bar{T}_S$ just as we treated the $t$-statistic from a single study. In effect, it is as if there is a single meta-study of studies, whose evidence may be combined in the best (i.e., minimum variance) way and then used as if it were the evidence provided by a single study.

Repeating the analysis from the main text would show that the fundamental default rule continues to hold, as applied to the statistic $\bar{T}_S$ rather than any one study’s $t$-statistic.\footnote{Because of the variance reduction entailed by averaging the $t$-statistics, though, using conventional critical values for the average $t$-statistic would turn out to be more demanding than in the one-study case.}