THE HIDDEN *DAUBERT* FACTOR: HOW JUDGES USE ERROR RATES IN ASSESSING SCIENTIFIC EVIDENCE

JOHN B. MEIXNER*
SHARI SEIDMAN DIAMOND**

In *Daubert v. Merrell Dow Pharmaceuticals*, the United States Supreme Court provided a framework under which trial judges must assess the evidentiary reliability of expert scientific evidence. One factor of the *Daubert* test, the "known or potential rate of error" of the expert's method, has received considerably less scholarly attention than the other factors, and past empirical study indicates that judges have a difficult time understanding the factor and use it less frequently in their analyses as compared to other factors. In this Article, we examine one possible interpretation of the "known or potential rate of error" standard that would treat the factor more broadly: considering direct assessments of a method's validity as assessments of the method's potential rate of error, even when numerical error rates are not mentioned. To assess the extent to which judges use the error rate factor in this "implicit" sense, we examined 208 federal district court cases, coding for the number of words judges spent analyzing the *Daubert* factors and other evidentiary considerations. We found that judges faced with a *Daubert* challenge often undertake a detailed analysis of the quality of the methodology used by the expert rather than simply relying on proxies for the quality of the method such as peer review and general acceptance. Analysis of a method's potential rate of error was significantly more common and lengthy than analysis using any of the other *Daubert* factors. This implicit error rate analysis also predicted the final admissibility ruling of the evidence and varied across expert disciplines. Our data support the notion that judges put considerable effort into directly assessing the validity of the scientific evidence before them when responding to a *Daubert* challenge. That is, they engage substantially in central processing in making methodological evaluations rather than merely relying on the peripheral cues of peer review and general acceptance. This finding lays the groundwork for future assessments of the obstacles judges face in these demanding evaluations.

---

* Law clerk, United States District Court for the Eastern District of Michigan; J.D., Northwestern University School of Law; Ph.D., Northwestern University Department of Psychology; B.S., University of Michigan.

** Howard J. Trienens Professor of Law and Professor of Psychology, Northwestern University School of Law; Research Professor, American Bar Foundation. We thank Jay Koehler, Ken Paller, and Peter Rosenfeld for helpful comments. This Article has benefitted from feedback from Ted Eisenberg, anonymous reviewers, and audience participants at the 2013 Conference on Empirical Legal Studies.

INTRODUCTION

Imagine you are a judge in a product liability suit brought against an automotive manufacturer.[1] The plaintiff in the suit, a truck driver, was driving his truck with his wife sleeping in the vehicle on I-75 near Toledo, Ohio, when the truck's steering mechanism suddenly gave out. The truck crashed into the median, causing injuries to the driver's wife. The truck's steering mechanism had been repaired about six months prior, following a separate accident. The defendant in the current suit had manufactured the truck's steering mechanism, and the plaintiff suspected that the new steering mechanism was responsible for the failure leading to the accident.

The plaintiff offers a truck mechanic as an expert witness in the case. After the accident, the mechanic inspected the steering gear on the plaintiff's truck and determined that the valve housing bolts on the steering gear were extremely loose. Using an identical steering gear as a test subject, the expert plans to testify that he loosened the valve housing bolts on the test steering gear to the same degree that they were loosened on the plaintiff's steering gear. At this degree of looseness, the steering column easily gave out. Based on this test, the expert plans to testify that the loose bolts were the cause of the accident. However, there is one critical problem with the testimony: the expert's opinions are based on the assumption that the bolts were at the exact same degree of looseness at the time of the inspection as they were at the time of the accident, even though the inspection occurred after the accident. Further, a photograph introduced by the defendant shows that the bolts were manipulated *after* the accident and *before* the expert's examination.

You are tasked with assessing the admissibility of the expert under Federal Rule of Evidence (FRE) 702 and *Daubert v. Merrell Dow Pharmaceuticals*.[2] The expert is well qualified based on his experience as a mechanic,[3] and his testimony, if valid, is clearly relevant to the question of whether the steering column was defective. The evidence is

1. This hypothetical is based on one of the cases used in the empirical analysis described in this Article: *Rose v. Truck Ctrs., Inc.*, 611 F. Supp. 2d 745 (N.D. Ohio 2009).

2. 509 U.S. 579 (1993).

3. We diverge from the actual case here slightly, in which the expert was ruled unqualified to testify both on the basis of his experience and due to the problems with the reliability of his methods. *Rose*, 611 F. Supp. 2d at 749–52. In order to demonstrate the issue of how to assess the reliability issue in the case, we isolate reliability by ignoring the qualifications problem that the case presented.

clearly flawed based on the expert's faulty assumption, but does the evidence fail to satisfy any of the four primary *Daubert* factors?[4] The expert clearly conducted testing on the steering column, and the method of comparing two steering columns is generally accepted in the mechanic community. Though the expert has not submitted his methods to peer review and publication, such publication is generally not standard among mechanics. The expert has not provided any error rate, but it is likely difficult to identify an explicit error rate for methods like these.

Perhaps such an assessment of the completeness of the expert's methodology is beyond the four factors provided by the *Daubert* Court.[5] However, an assessment of the core competency of the method itself used by the expert seems to be right at the heart of what the *Daubert* Court intended in modifying the test for admissibility of expert testimony: a shift from the *Frye* standard (which does not assess the method itself but instead trusts other members of the relevant scientific community) to the *Daubert* standard (which tasks judges with the responsibility to assess "whether the reasoning or methodology underlying the testimony is scientifically valid").[6] Indeed, the judge in the actual case upon which this hypothetical was based ruled the evidence inadmissible in part because the faulty assumption made by the expert rendered the testimony unreliable:

> Thus, [the expert's] opinions are based on the assumption the bolts were at the exact same degree of looseness at the November 2006 inspection as they were at the time of the accident in May 2006. However, it is undisputed that the July 2006 photograph of the steering gear shows the bolts were manipulated *after* the accident and *before* Smith's examination. The July 2006 photo shows at least one of the bolts was completely separated from the steering gear, and not fastened to some degree as it was in November 2006.
>
> [The expert's] opinion failed to account for this manipulation; thus, his opinion is unreliable because any

4. We discuss the factors in detail later in the Article, but the four typically discussed factors are: (1) whether a technique can be or has been tested, (2) whether a technique has been subjected to peer review and publication, (3) the known or potential rate of error of the technique, and (4) whether the technique is generally accepted in the relevant scientific community. For further discussion, see *infra* note 12 and accompanying text.

5. The Court of course made clear that many factors beyond the four provided could bear on the admissibility inquiry. *Daubert*, 509 U.S. at 593 ("Many factors will bear on the inquiry, and we do not presume to set out a definitive checklist or test.").

6. *Id.* at 592–93. Though the mechanic in this example is not a scientist, under *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999), the *Daubert* standard applies to expert testimony from nonscientists as well as scientists. *Id.* at 141.

> testing of bolts was unrelated to the condition of the steering gear at the time of the accident. This error goes beyond a "mere weakness" in the factual basis of Smith's opinion, it destroys the factual foundation upon which he rendered his opinion.[7]

In this case, as in most cases, the expert is unable to provide a numerical error rate regarding his methods. Thus, in determining how accurate the expert's method is likely to be, the judge must examine the methodology for flaws that are *likely to produce* errors. In this case, the major flaw of measuring the bolt tightness after the bolts were manipulated was severe enough to lead to an unreasonably high potential for error and thus render the evidence inadmissible. Though the judge does not explicitly mention error rates, the analysis is an implicit consideration of the likelihood of error: what we will call an *implicit error rate analysis*.

The remainder of this Article examines the role that the error rate factor of *Daubert* plays in trial courts' analysis of the reliability of expert testimony.[8] As a preview, we find that error rate analysis is far more common and more extensive than prior research would suggest, and that the extensiveness of error rate analysis is strongly predictive of admissibility decisions. It is also more predictive than treatment of more peripheral *Daubert* cues of methodological rigor like general acceptance and peer review. Part I examines the *Daubert* decision itself, describes two possible interpretations of the "known or potential rate of error" factor, and examines how that factor may be used *implicitly* to assess the quality of an expert's methodology. Part II reviews the empirical literature examining courts' use of all of the *Daubert* factors, demonstrating that studies to date have found that courts show little interest in using the known or potential rate of error factor. Part III introduces our empirical study of federal district courts' treatment of the

---

7. *Rose*, 611 F. Supp. 2d at 751 (citation omitted).

8. There is an important distinction to make between validity (a principle's ability to show what it purports to show) and reliability (an application's ability to produce consistent results). The *Daubert* Court noted this distinction in a footnote and argued that while the terms have differences, they are "distinct from the other by no more than a hen's kick," and thus the Court stated that its focus was on "*evidentiary* reliability—that is, trustworthiness." *Daubert*, 509 U.S. at 590 n.9 (quoting James E. Starrs, Frye v. United States *Restructured and Revitalized: A Proposal to Amend Federal Evidence Rule 702*, 26 JURIMETRICS J. 249, 256 (1986)) (internal quotation mark omitted). While it is not exactly clear how evidentiary reliability compares to the more common definitions of validity and reliability, it appears to involve some amalgamation of the two but remains closer to the former. As the Court stated, "In a case involving scientific evidence, *evidentiary reliability* will be based upon *scientific validity*." *Id.* Thus, when we refer to reliability here in terms of a court making an admissibility decision, we are speaking of evidentiary reliability.

Daubert factors and describes our methodology in examining over 200 trial-court *Daubert* cases. Part IV provides our hypotheses, and Part V describes and discusses our results. Part VI discusses our findings in the context of *Daubert* doctrine and concludes.

I. THE *DAUBERT* DECISION'S TREATMENT OF ERROR RATES

For nearly 70 years, admissibility of scientific expert testimony in federal courts was determined under the standard laid out in *Frye v. United States*.[9] In that case, the D.C. Circuit Court of Appeals outlined a "general acceptance" test under which any scientific evidence that is to be admissible "must be sufficiently established to have gained general acceptance in the particular field in which it belongs."[10] In 1993, in the landmark decision of *Daubert v. Merrell Dow Pharmaceuticals*, the United States Supreme Court held that the FRE, passed in 1975, superseded *Frye*.[11] The Court in *Daubert* outlined a nonexclusive multifactor test in which trial courts are tasked with assessing the reliability of expert evidence. The Court identified five nonexclusive factors for the judge to consider when determining the reliability of scientific evidence: (1) whether a technique can be or has been tested, (2) whether a technique has been subjected to peer review and publication, (3) the known or potential rate of error of the technique, (4) the existence and maintenance of standards controlling the technique's operation, and (5) the general acceptance of the technique.[12]

Daubert immediately spurred a great deal of scholarly discussion. Opinions on the standard have run the gamut from positive to negative with many writers lamenting that the standard is confusing and ambiguous[13] or did not go far enough,[14] though some lauded the standard

9. 293 F. 1013, 1013–14 (D.C. Cir. 1923) (ruling that the "systolic blood pressure deception test," an early lie-detection test, had not gained "general acceptance" among physiological and psychological authorities and therefore should not be admitted as evidence).

10. *Id.* at 1014.

11. *Daubert*, 509 U.S. at 589.

12. *Id.* at 593–94. There is some dispute as to whether the test contains five separate factors (with maintenance of standards as a separate factor) or whether the error rate and maintenance of standards factors combine to form one single factor, yielding a total of four. *Compare Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 158 (1999) (interpreting four factors), *with United States v. Crisp*, 324 F.3d 261, 265 (4th Cir. 2003) (interpreting five factors). The difference between the two views appears to be purely semantic, though most scholars and judges appear to favor the "four factor" characterization. It should, of course, be noted that the factors provided by the Court are nonexclusive, and "[m]any factors will bear on the inquiry." *Daubert*, 509 U.S. at 593.

13. *See, e.g.*, Daniel J. Capra, *The* Daubert *Puzzle*, 32 GA. L. REV. 699, 704 (1998) ("*Daubert* created many problems for the lower courts, in large part because the opinion gives a mixed message."); David L. Faigman, David H. Kaye, Michael J. Saks &

as enabling the trial judge to keep "junk science" out of court.[15] However, many commentators expected that the decision would do little to change admissibility outcomes, as judges would simply use the multifactor test to arrive at the same outcome they would have reached under *Frye*,[16] and while results are mixed, at least some data indicate that the aggregate effects of *Daubert* on admissibility rates have been small.[17]

Joseph Sanders, *How Good is Good Enough?: Expert Evidence Under* Daubert *and* Kumho, 50 CASE W. RES. L. REV. 645 (2000) (expressing disagreement with other scholars regarding the criteria set by *Daubert*); G. Michael Fenner, *The* Daubert *Handbook: The Case, Its Essential Dilemma, and Its Progeny*, 29 CREIGHTON L. REV. 939, 955 (1996) (describing disagreement among judges as to what *Daubert* means for the expert evidence standard); Lisa Heinzerling, *Doubting* Daubert, 14 J.L. & POL'Y 65, 65–66 (2006) (discussing a number of issues leading to confusion regarding the law post-*Daubert*); Jon Y. Ikegami, *Objection: Hearsay—Why Hearsay-Like Thinking is a Flawed Proxy for Scientific Validity in the* Daubert *"Gatekeeper" Standard*, 73 S. CAL. L. REV. 705, 711 (2000) (quoting a district judge as stating that applying the *Daubert* standard is like being "hit . . . between your eyes with a four-by-four" (alteration in original)); Randolph N. Jonakit, *The Meaning of* Daubert *and What That Means for Forensic Science*, 15 CARDOZO L. REV. 2103, 2106 (1994) (stating that "crucial questions were not addressed" by the *Daubert* opinion); Janine M. Kern & Scott R. Swier, Daubert v. Merrell Dow Pharmaceuticals, Inc.*: "Gatekeeping" or Industry "Safekeeping"?*, 43 S.D. L. REV. 566, 575 (1998); Cassandra H. Welch, Note, *Flexible Standards, Deferential Review:* Daubert*'s Legacy of Confusion*, 29 HARV. J.L. & PUB. POL'Y 1085, 1091 (2006) ("The language of the decision lack clarity.").

14. Ronald J. Allen & Esfand Nafisi, Daubert *and Its Discontents*, 76 BROOK. L. REV. 131, 134 (2010) ("Unfortunately, once past the admission threshold, nothing forbids the presentation of the evidence to the jury in the tired, old, radically-subversive-to-the-goals-of-the-legal-system, deferential fashion. The true problem with *Daubert*, in other words, is that it did not go far enough."); Richard D. Friedman, *Squeezing* Daubert *Out of the Picture*, 33 SETON HALL L. REV. 1047, 1048 (2003) ("[R]arely if at all should the court exclude [scientific evidence] on the mere ground that the jury is likely to over-value it. Thus, I am suggesting that *Daubert* be squeezed out of the picture by other approaches to the problem."); Joseph B. Spero, *Much Ado About Nothing—The Supreme Court Still Fails to Solve the General Acceptance Problem Regarding Expert Testimony and Scientific Evidence*, 8 J.L. & HEALTH 245, 268 (1994) (stating that *Daubert* "did not address anything at all").

15. *See, e.g.*, David E. Bernstein, *The Admissibility of Scientific Evidence After* Daubert v. Merrell Dow Pharmaceuticals, Inc., 15 CARDOZO L. REV. 2139, 2181 (1994) ("In the end, our legal system will be the better [because of the *Daubert* decision]. By providing a flexible test focusing on the criteria used by scientists to determine the trustworthiness and validity of scientific conclusions, *Daubert* will ensure that scientific evidence that is admitted in court is trustworthy and reflects scientific knowledge.").

16. *See, e.g.*, Bert Black, Francisco J. Ayala & Carol Saffran-Brinks, *Science and the Law in the Wake of* Daubert*: A New Search for Scientific Knowledge*, 72 TEX. L. REV. 715, 743 (1994) ("The great *Frye* debate notwithstanding, the real difference in scientific evidence cases is not general acceptance versus relevance/reliability, but whether or not the court is willing to undertake a thorough and active review. Courts that want to dig into the details of an expert's reasoning and the validity of his or her testimony can do so with or without *Frye*.").

17. *See, e.g.*, Edward K. Chen & Albert H. Yoon, *Does* Frye *or* Daubert *Matter? A Study of Scientific Admissibility Standards*, 91 VA. L. REV. 471, 503 (2005)

While there has been extensive commentary regarding the *Daubert* decision, the scholarly community has paid little attention specifically to the "known or potential error rate" factor.[18]

Perhaps this lack of scholarly attention is in part because the *Daubert* Court itself seemed to deemphasize the error rate factor as compared to the others. Spending fewer words on this factor than on any of the other three, the Supreme Court simply stated that "the court ordinarily should consider the known or potential rate of error."[19] Regarding quantitative error rates, the Court did not specify how broadly the error rate of a method should be defined: Is it measured at a general level, looking to the rate of error of an overall technique broadly (e.g., the national overall rate of error of DNA testing)? Or is it measured at a specific level, looking to the error rate of the individual expert testifying (e.g., the error rate of the individual examiner who conducted the DNA analysis in the case at hand)?[20] Obviously, one or both of these error rates may be unknown in many cases, but the Court did not provide guidance as to which error rate trial courts should focus on. Additionally, a quantitative error rate is made up of more than a single value; a trial court considering an error rate could look to either the false positive error rate (the chance of returning a condition-positive outcome on the test when the true status is negative) or the false negative rate (the chance of returning a condition-negative outcome on the test when the true status is

(finding that "a state's choice of scientific admissibility standard does not have a statistically significant effect on removal rates" which "may support the broader theory that a state's adoption of *Frye* or *Daubert* makes no difference in practice"); Eric Helland & Jonathan Klick, *Does Anyone Get Stopped at the Gate? An Empirical Assessment of the* Daubert *Trilogy in the States*, 20 SUP. CT. ECON. REV. 1, 32–33 (2012) (finding no differences in the types of experts retained in state cases both before and after *Daubert* and concluding that "there is virtually no systematic evidence supporting the view that adoption of *Daubert* makes any difference at all").

18. This has been noted in one recent article focusing specifically on the known or potential rate of error factor. Mark Haug & Emily Baird, *Finding the Error in* Daubert, 62 HASTINGS L.J. 737, 740 (2011). In that article, the authors speculated as to three reasons why the factor may have received less attention: "(1) it is difficult to define, but 'we know it when we see it'; (2) it is merely a detail of 'evidentiary reliability' and therefore, does not warrant such attention; or, (3) it is too difficult to implement." *Id.* To our knowledge, Haug and Baird's article is the most extensive treatment to date on the error rate factor, though one student note has focused specifically on the error rate factor in criminal contexts. Munia Jabbar, Note, *Overcoming* Daubert*'s Shortcomings in Criminal Trials: Making the Error Rate the Primary Factor in* Daubert*'s Validity Inquiry*, 85 N.Y.U. L. REV. 2034 (2010). We discuss Haug and Baird's article in more detail *infra* Parts II and III.

19. *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579, 594 (1993).

20. This unanswered question was noted by Munia Jabbar in her note. Jabbar, *supra* note 18, at 2044–45. Jabbar argues, and we agree, that trial judges should consider error rate at the specific level when making admissibility decisions because general error rates may not take into account expert-specific shortcomings. *See id.* at 2037.

positive). The *Daubert* Court did not point out this nuance or mention whether trial courts should consider differences between the two types of error as relevant. From a normative standpoint, certain types of errors may be especially undesirable in particular contexts. For example, we may worry particularly about false positive errors made by forensic experts in criminal trials because of the increased burden of proof in that context.[21] *Daubert*, however, remains silent as to whether courts should make any context-based distinction when evaluating error rates.

Perhaps most importantly, the *Daubert* Court did not specify whether the error rate factor is intended to apply only to quantitative error rates that can be identified by the expert (or the field more generally) or whether it can apply more broadly to the chance that the expert may have made a mistake in his methods that could lead to erroneous testimony being given to the trier of fact. While the former (which we term the "restricted" view of the error rate factor) would be applicable only in those limited circumstances in which an error rate could be identified based on testing, the latter characterization (which we term the "broad" view of the error rate factor) would be applicable in a wide array of circumstances, such as in the product liability example provided at the start of this Article. When the judge does not have a known error rate to assess, she can assess the potential for error in evaluating flaws in the expert's methodology. A plain-language interpretation of the "known or potential" language written by Justice Blackmun in *Daubert* could be thought of to encompass these two types of error rate analysis: (1) the more explicit "known" error, which can be evaluated simply by assessing a numerical value, and (2) the more implicit "potential" error, which can be assessed by examining the methodology and evaluating its potential for producing erroneous results.

Which of these two interpretations more accurately reflects the Court's intent in *Daubert*? We do not know for certain, but, for several reasons, we suspect that the court intended the error rate factor to be closer to the restricted view. First, the case that Justice Blackmun cited as a past example in which a lower court considered the known or potential rate of error of a method, *United States v. Smith*,[22] contained clearly defined quantitative error rates.[23] That case involved a bank and wire fraud criminal charge in which spectrographic voice analysis was used to identify the defendant as the same person who had made phone calls to a

21. This observation has also been made by others in the literature. *See, e.g.*, David L. Faigman et al., *Admissibility of Scientific Evidence*, *in* 1 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY § 1:21 (2014); Jabbar, *supra* note 18, at 2045.

22. 869 F.2d 348 (7th Cir. 1989).

23. *Id.* at 353–54.

bank as part of the fraudulent scheme.[24] Justice Blackmun cited pages 353 to 354 of the *Smith* opinion, in which the Seventh Circuit described a field study assessing the error rate of spectrographic voice identification:

> [The expert] also testified as to studies performed in the field. He first discussed a study performed by Professor Oscar Tosi of Michigan State University in conjunction with the Michigan State Police from 1968 to 1970. Of the 35,000 comparisons made in this study, the error rate for false identifications was 2.4% and the error rate for false eliminations was about 6%. This study previously has been cited as authoritative by other federal courts of appeal. A follow up to that study conducted by Dr. Tosi involving only actual cases examined by trained voice examiners found no errors whatsoever.
>
> Nakasone also discussed a more recent report published by the FBI in June, 1987 in the Journal of the Acoustical Society of America. The cases in that report which were submitted to actual determinations yielded a .31% rate of false identifications and a .53% rate of false eliminations. Finally, Nakasone testified that variations, such as use of tapes not recorded under laboratory conditions and attempts by the speaker to disguise her voice, will increase the error rate of false eliminations. That is, instead of resulting in more false identifications, these variations will result in more false eliminations.[25]

In addition to this citation, there is another reason to believe that the Court may have been thinking narrowly when it listed the error rate factor. If, instead of a narrow conception of error rate, the Court had intended a broad version with great sweep and potential importance, we might have expected a more detailed treatment of the error rate factor in the *Daubert* opinion. It appears that the Federal Rules Committee also assumed a narrow conception of the error rate factor in *Daubert*. The 2000 amendments to FRE 702 provided "other factors" in the advisory notes that would likely overlap with our broad definition of the error rate factor, such as whether an expert "is being as careful as he would be in his regular professional work."[26] Some lower courts have adopted these

---

24. *Id.* at 349–50.

25. *Id.* at 353–54 (footnote omitted) (citation omitted). Interestingly, the case does make the distinction between false positives and false negatives, *see id.*, though the *Daubert* opinion itself does not.

26. FED. R. EVID. 702 committee's notes on rules—2000 amendment (quoting *Sheehan v. Daily Racing Form Inc.*, 104 F.3d 940, 942 (7th Cir. 1997)).

factors explicitly in their analysis.[27] If the Supreme Court in *Daubert* intended to outline a broad error rate factor, the Federal Rules Committee has not reflected that broad interpretation through FRE 702.

However, there are some arguments for a broad interpretation of the error rate factor. The *Daubert* decision can be seen as a shift from the *Frye* general acceptance test to a test that focuses on scientific validity, and the broad interpretation of the error rate factor invites the trial judge to assess scientific validity more directly than any of the other factors by looking to whether the expert's methods are likely to lead to the conclusion that the expert claims they will.[28] Thus, the broad interpretation of the error rate factor could be said to better fit the spirit of *Daubert* than the restricted view by calling on the judge to directly assess the quality of the expert's methods, as was done in the example case at the beginning of this Article. And even if the Court intended a restricted meaning when it listed the error rate factor in *Daubert*, trial courts taking a serious view of their *Daubert*-assigned role as gatekeeper could still be encouraged to use the broad version in their analysis because whether an expert is likely to make an error in his assessment goes to the heart of the validity question. In the empirical project we describe in Parts III, IV, and V, we examine the extent to which federal district courts conduct both restricted (or "explicit") and broad (or "implicit") error rate analysis in *Daubert* decisions.

Beyond this fundamental interpretation issue, other major questions of how to deal with error rates were entirely ignored in *Daubert*. Although the Court said that lower courts should "consider" the known or potential rate of error,[29] it did not specify whether lower courts should (1) examine the rate of error to determine whether it stays underneath some unknown threshold, above which the factor cuts against admissibility (we term this the "threshold" standard), or (2) simply ascertain whether an accurate rate of error has been produced, leaving the

27. *E.g.*, *In re Unisys Sav. Plan Litig.*, 173 F.3d 145, 166 & n.10 (3d Cir. 1999); *Sheehan v. Daily Racing Form, Inc.*, 104 F.3d 940, 942 (7th Cir. 1997).

28. One possible objection to the broad view of the error rate factor is that it impermissibly looks to the expert's conclusions rather than his methods. *Daubert* specifically forbids such post-hoc analysis: "The focus, of course, must be solely on principles and methodology, not on the conclusions that they generate." *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579, 594–95. However, while the broad view of the error rate factor considers the likelihood that the expert's conclusions will be erroneous, it bases that likelihood on the quality of the expert's methods themselves, in the same way an explicit error rate gives the likelihood that the expert's conclusions will be erroneous based on past empirical testing. Thus, the only reason any analysis of the expert's methods is relevant to admissibility is in light of the conclusions that the expert is likely to generate; reliable methods are not an end in themselves, but rather a means to the end of achieving valid conclusions to present to the trier of fact.

29. *Id.* at 594.

trier of fact to assess the probative value of the evidence in light of that error rate (we term this the "simple provision" standard). If the Court meant that a certain threshold of error is intolerable, the Court left it entirely to lower court judges to determine what the threshold is. Is there a lower bound to the requirement, such as chance accuracy? Does the required accuracy level change in different contexts, such as civil versus criminal trials?[30]

To make matters even more confusing, the Supreme Court seemingly changed the error standard in *Kumho Tire Co. v. Carmichael*,[31] the case that held that the *Daubert* standard applies to all experts.[32] While the *Daubert* Court characterized the error rate factor in a way that does not distinguish between the threshold standard and the simple provision standard, stating that "the court ordinarily should consider the known or potential rate of error,"[33] *Kumho Tire* clearly characterized the factor as a threshold, asking the question "[w]hether, in respect to a particular technique, there is a high 'known or potential rate of error.'"[34] Because the *Kumho Tire* Court did not state that its intent was to change the error rate standard, one could assume that the *Daubert* court intended a threshold standard all along. However, this is in tension with the general "liberal thrust" of the FRE, which favor admissibility where evidence is relevant and not misleading to the jury.[35] Thus, based on the unexplained inconsistency between *Daubert* and *Kumho Tire*, we would expect some confusion from the lower courts on the matter, especially prior to the *Kumho Tire* decision in 1999.[36]

---

30. *See supra* note 21 and accompanying text.

31. 526 U.S. 137 (1999).

32. *Id.* at 148–49.

33. *Daubert*, 509 U.S. at 594.

34. *Kumho Tire*, 526 U.S. at 149. We note that our wording of this standard as a "threshold" is a bit different than the testability, peer review, and general acceptance standards, which are all worded as qualities that the evidence must achieve in order to foster admissibility. The error rate factor as characterized in *Kumho Tire* and in our "threshold" wording, in contrast, is a quality that evidence must avoid to be admissible.

35. *Daubert*, 509 U.S. at 588.

36. Unsurprisingly, the academic literature is also inconsistent in deciding between the simple provision standard and the threshold standard. *Compare* Robert J. Goodwin, *The Hidden Significance of* Kumho Tire Co. v. Carmichael*: A Compass for Problems of Definition and Procedure Created by* Daubert v. Merrell Dow Pharmaceuticals, Inc., 52 BAYLOR L. REV. 603, 611 (2000) (phrasing the *Daubert* standard as whether a method "has a known or potential error rate"), *and* Pamela J. Jensen, Note, Frye *Versus* Daubert*: Practically the Same?*, 87 MINN. L. REV. 1579, 1583 (2003) (phrasing the standard in this way: "does the technique have a 'known or potential rate of error'"), *with* Mark Lewis & Mark Kitrick, Kumho Tire Co. v. Carmichael*: Blowout from the Overinflation of* Daubert v. Merrell Dow Pharmaceuticals, 31 U. TOL. L. REV. 79, 83 (1999) (phrasing the standard as "whether the methodology is accurate").

The lack of attention paid to the error rate factor of *Daubert* is disconcerting, as it is the only factor that speaks directly to the probative value of the evidence itself. Both the peer review and general acceptance factors are only proxies for scientific quality—though the *Daubert* Court specifically stated that peer review is useful because "it increases the likelihood that substantive flaws in methodology will be detected"[37] and thus is not simply a test of scientific agreement. That is, it reassures the judge that competent scientists have vetted the method in question for flaws that the judge herself may not be able to identify. Likewise, the general acceptance factor has nothing to do with the technique itself—as many others have noted, a theory or technique may be widely accepted despite a lack of evidentiary reliability.[38] While the testability factor is a "substantive" one in that it looks to the particular technique itself, it can be viewed as a threshold question: the Court noted that "whether [a theory] can (and has been) tested" helps determine not only whether the technique is reliable enough to be considered as evidence, but also whether the technique is scientific knowledge at all and whether an error rate for the technique could even be generated.[39] With these three factors of the *Daubert* standard seemingly not providing an avenue to broadly assess the validity of a scientific technique—that is, its ability to accurately measure what it purports to measure—it would appear that the known or potential rate of error should potentially be the most important in assessing the science itself that is at issue in a *Daubert* hearing.

Of course, because factor tests like the *Daubert* standard are well known for providing great leeway to the trial judge to weigh and apply the factors as she deems appropriate, trial judges are likely to form and apply their own interpretations of how to use the error rate factor. Surprisingly, despite the now 21 years of data accumulated since *Daubert*, little empirical research has examined how trial judges have interpreted the *Daubert* standard. In the small body of empirical research on *Daubert*, only a fraction relates to the error rate factor, which appears to be the most difficult factor to understand due to its extremely vague language and seemingly technical nature. In the next Part, we summarize this literature and describe the gap we fill with our study.

II. PAST EMPIRICAL RESEARCH ON THE *DAUBERT* DECISION

Initial studies of the effects of *Daubert* on judges' analyses of scientific evidence appeared in the early 2000s. In a report from the

37. *Daubert*, 509 U.S. at 593.
38. *E.g.*, Faigman et al., *supra* note 21, § 1:5.
39. *Daubert*, 509 U.S. at 593.

RAND Institute for Civil Justice, Lloyd Dixon and Brian Gill[40] analyzed 399 federal district court opinions[41] from 1980 to 1999 and 601 elements of expert evidence in those 399 cases.[42] Coding for a variety of factors to assess changes in rates of offer, admission, and usage of scientific evidence, Dixon and Gill found that while *Daubert* did not bring about major changes in the overall rate at which scientific evidence was admitted or the extent to which plaintiffs were successful in their claims, there were small changes in the way judges discussed the evidence,[43] as would be expected following a shift in the legal standard. Dixon and Gill found that post-*Daubert*, judges increasingly reviewed all types of expert evidence as opposed to just natural science evidence, adapted their analyses to fit the new factors, and began to mention relevance and qualifications more frequently than before.[44] Coding for the four *Daubert* factors as well as several other indicators of reliability,[45] Dixon and Gill found that all of the factors were addressed more frequently, and reliability in general was discussed more extensively following *Daubert*.[46] However, the study simply coded whether or not each factor was addressed at some point in each opinion.[47] No attempt was made to assess the extent to which judges analyzed and weighed the various factors.

---

40. LLOYD DIXON & BRIAN GILL, RAND INST. FOR CIVIL JUSTICE, CHANGES IN THE STANDARDS FOR ADMITTING EXPERT EVIDENCE IN FEDERAL CIVIL CASES SINCE THE *DAUBERT* DECISION (2001), *available at* http://www.rand.org/content/dam/rand/pubs/monograph_reports/2005/MR1439.pdf. This timeline falls within all three of the *Daubert* trilogy of cases that make up the current standard for scientific evidence, along with FRE 702. *Daubert* itself was decided in 1993. *General Electric Co. v. Joiner*, 522 U.S. 136 (1997), was decided in 1997 and held that appellate courts should use an abuse of discretion standard when reviewing a trial judge's admissibility decision. *Id.* And *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999), was decided in 1999 and held that the *Daubert* standard applies to all expert testimony, not just scientific testimony. *Id.* Of course, the Dixon and Gill sample only contains a few years post-*Kumho*, which limits the extent to which the data can assess the effect of that decision.

41. DIXON & GILL, *supra* note 40, at xiii. The 399 cases were selected using a search string much like the one we use in our empirical study. *See id.* at 16–17 n.3; *infra* note 76.

42. DIXON & GILL, *supra* note 40, at xiii–xvi. Dixon and Gill term any separate discussion of expert evidence as an element. *See id.* at 18–19 ("For example, a judge might address a challenge to the valuation of lost profits or wages in one part of the opinion and a challenge to toxicological evidence in another. We instructed the coders to be driven by the structure of the opinion in deciding whether judges addressed multiple elements of evidence.").

43. *Id.* at xv–xvi.

44. *Id.* at 61–63.

45. *Id.* at 37, 38 tbl.5.1.

46. *Id.* at 39 tbl.5.2, 40.

47. *See id.* at 37.

More recent studies have further examined judicial understanding of the *Daubert* factors. First, survey projects have found that judges appear to better understand the more external factors—that is, peer review and general acceptance—and in some cases consider these factors more important than error rate and testability. A survey conducted by Sophia Gatowski and colleagues in 2001 gathered responses from 400 state court judges on questions designed to test whether they understood the scientific meaning of the *Daubert* guidelines and were able to apply them in practical situations.[48] The survey raised some concerns regarding judges' ability to evaluate the *Daubert* factors—notably, less than 5% of all judges demonstrated a "clear understanding" of the testability and error rate factors,[49] while over 70% demonstrated a clear understanding of the peer review and general acceptance factors.[50] Despite this difference in ability to assess the various factors, the judges were split roughly evenly as to which factor they thought should be given the most weight in the analysis outside of general acceptance, which judges tended to favor.[51] In another survey of 325 state trial judges, Veronica Dahir and colleagues asked judges who were experienced in assessing the admissibility of syndrome evidence which admissibility considerations were most important in their assessments of the evidence.[52] While judges frequently mentioned qualifications and general acceptance first among their considerations in admissibility, only one of 216 judges who gave a codable response mentioned testability first,[53]

48. Sophia I. Gatowski et al., *Asking the Gatekeepers: A National Survey of Judges on Judging Expert Evidence in a Post*-Daubert *World*, 25 LAW & HUM. BEHAV. 433, 435 (2001).

49. *Id.* at 444, 447; *see also id.* at 445 fig.1. Coding of the judges' responses in the Gatowski study was done by separating responses into three possible categories: "judge understands concept," "judge's understanding of concept is questionable," and "judge clearly does not understand concept." *Id.* at 441. One caveat of this study is the difficultly in coding such diverse responses into three discrete categories. While the interrater reliability of the coders was relatively high at .84, *id.*, the indeterminate "questionable" category encompassed the majority of responses for the testability and rate of error factors, where understanding may be difficult to evaluate. *See id.* at 441, 445 fig.1. Thus, the very low rates of "clear understanding" may be somewhat misleading.

50. *Id.* at 447–48; *see also id.* at 445 fig.1.

51. *Id.* at 448.

52. Veronica B. Dahir et al., *Judicial Application of* Daubert *to Psychological Syndrome and Profile Evidence: A Research Note*, 11 PSYCHOL. PUB. POL'Y & L. 62, 71 (2005).

53. In this Article, we term the first *Daubert* factor as "testability," though Dahir and some others in the scholarly literature have termed it the "falsifiability" factor. *See e.g.*, *id.* at 64. Although the two words have overlapping, but not completely identical, connotations, both we and other authors in this area are referring to the same *Daubert* factor when using this terminology: "determining whether a theory or technique . . . can be (and has been) tested." *Daubert v. Merrell Dow Pharm.*, 509 U.S. 579, 593 (1993).

while no judges mentioned error rate first.[54] Similar results were found for profile evidence, and the pattern of results remained the same when judges were asked to identify what aspects of syndrome and profile evidence they found most problematic.[55] These results, along with the results of the Gatowski survey, raise concerns that judges may particularly struggle with the two more technical factors that actually examine the substantive strength of the evidence—testability and error rate.

Studies examining judicial opinions and litigation strategy have also tended to find limited use of the error rate and testability factors. In a case-coding study similar to Dixon and Gill's 2001 report, Jennifer Groscup and colleagues coded for the number of words judges spent discussing each of the *Daubert* factors as well as other evidentiary rules that are relevant when expert evidence is proffered.[56] Unlike the Dixon and Gill study, which coded only whether or not a particular factor was mentioned in the case, this study provides some insight into how judges weigh the various factors by examining the number of words spent discussing each.[57] Groscup found that judges spent the fewest words discussing the error rate and testability factors, devoting an average of fewer than 10 words to each factor per opinion.[58] Discussion of the peer review and general acceptance standards was greater (approximately 15 and 55 words, respectively), though judges spent more time discussing FRE 702 than they did discussing the *Daubert* factors at all.[59] Groscup also estimated the influence that each factor had on the outcome of the admissibility decision,[60] and both error rate and testability were judged to

54. Dahir et al., *supra* note 52, at 71.

55. *Id.* at 72–73.

56. Jennifer L. Groscup et al., *The Effects of* Daubert *on the Admissibility of Expert Testimony in State and Federal Criminal Cases*, 8 PSYCHOL. PUB. POL'Y & L. 339, 342–43 (2002). Some examples of other evidentiary rules that may come into play are FRE 104, which requires that all evidence must be relevant, and FRE 403, which allows the judge to exclude relevant evidence if its probative value is substantially outweighed by a danger of prejudice. FED. R. EVID. 104, 403.

57. We used word counts as our principal dependent measure. We acknowledge that although widely used, word count may be an imperfect proxy for importance or weight. We explore the benefits and drawbacks of using word counts *infra* notes 165–69 and accompanying text.

58. Groscup et al., *supra* note 56, at 350.

59. *Id.*

60. To determine this, coders gave a subjective 0–9 rating for each criterion. *Id.* at 353–54 ("For each of these criteria, a rating of its influence on the decision was made. This rating was on a 10-point scale, where 0 indicated that the criterion was not mentioned, 1 indicated that it was mentioned but was not at all influential, and 9 indicated that it was mentioned and was very influential."). Each variable was coded by three independent raters; interrater reliability scores are not given for each variable in the

be less important criteria than peer review and general acceptance.[61] "Reliability" more generally under FRE 702 was judged to be more important, but criteria unrelated to the science itself were judged as the most important, such as qualifications of the expert.[62] Thus, like the Gatowski study, the Groscup et al. study implies that judges are less concerned with error rates than they are with other factors. Notably, however, all of the cases sampled in the study were criminal appellate cases,[63] leaving unanswered the question of whether the results accurately characterize court behavior in unappealed decisions and civil cases.

In a more recent related study from 2011, Mark Haug and Emily Baird examined both federal district and circuit cases, specifically looking at judges' use of the error rate factor.[64] Though they sampled only 107 cases (which contained 200 total experts), they similarly found that rate of error was underused compared to the other *Daubert* factors, with only 33 of 200 assessments of expert admissibility containing error rate discussion and none focusing on the error rate factor alone.[65] The parties themselves may also consider error rates less important than other factors in litigating the admissibility of expert testimony: in a study examining the grounds for *in limine* challenges to expert evidence in federal district cases in South Carolina, David Flores and his colleagues found that such motions included a challenge based on error rates in only one of 25 cases, though testability was a common challenge, appearing in 14 of 25 cases.[66]

The research described in this Article builds on the work summarized above, but importantly includes a more in-depth examination of the error rate analysis.[67] The general thrust of prior

study, though 87% of the variables showed significant correlations between coders. *See id.* at 343 n.1.

61. *Id.* at 355 tbl.5.

62. *Id.*

63. *Id.* at 342–43.

64. Haug & Baird, *supra* note 18, at 744. Haug and Baird do not specify whether they sampled only civil cases, only criminal cases, or both. Because the search term provided by the authors would likely return both civil and criminal cases, we assume that both were sampled, though we do not know what proportion of cases fall into each category.

65. *Id.* at 744–45, 746 tbl.2.

66. David M. Flores, James T. Richardson & Mara L. Merlino, *Examining the Effects of the* Daubert *Trilogy on Expert Evidence Practices in Federal Civil Court: An Empirical Analysis*, 34 S. ILL. U. L.J. 533, 556–57, 557 tbl.7 (2010).

67. We also note that while we are most interested here in studying the way judges use the *Daubert* factors and other evidentiary considerations in assessing expert evidence, our data also speak to the issue of the relative success of the various parties involved in the cases (i.e., civil plaintiffs and defendants as well as criminal prosecutors and defendants). On this point, we would be remiss not to mention a comprehensive

research is that judges give little consideration to error rates and rely more heavily on proxies for good science, such as the easier-to-apply general acceptance and peer review factors.[68] The Gatowski study implies that judges have serious difficulty understanding the concept of error rates and are thus relatively unequipped to use the factor in their decisions.[69] Additionally, the rating results from the Groscup study suggest that judges put less emphasis on error rates than they do on the other *Daubert* factors.[70] Likewise, the recent Haug and Baird study finds that judges are dismissive of the error rate factor and rarely use the standard, even when they mention it in their discussion of the law.[71]

Yet by restricting the error rate factor to the narrow discussion of numerical error rates provided by experts, these prior studies may have underestimated the consideration of error rates in judicial analyses and thus misunderstood how judges think about them. Because applicable error rates are not available in many disciplines and thus can rarely be provided by experts or assessed by judges, judges may adopt a less formal approach to analyze what an expert's error rate is *likely to be*. We thus would predict that judges in such situations will spend significant time talking about the dependability of the methods used by an expert, whether the conclusions made by the expert align with those methods, whether those methods seem sound, and so forth. In doing this, the judge,

---

study by Michael Risinger comparing the outcomes of challenges to expert testimony in both civil and criminal cases. *See generally* D. Michael Risinger, *Navigating Expert Reliability: Are Criminal Standards of Certainty Being Left on the Dock?*, 64 ALB. L. REV. 99 (2000). Risinger's study found that decisions in both federal and state courts tended to favor civil defendants and strongly disfavored criminal defendants. *See generally id.* We find similar results here and further discuss this *infra* Parts V and VI.

68. *See, e.g.*, DIXON & GILL, *supra* note 40, at 40; Gatowski, *supra* note 48, at 445 fig.1, 446 tbl.1; Groscup et al., *supra* note 56, at 350 tbl.3; Haug & Baird, *supra* note 18, at 744–46 tbl.2. *But see* Flores, Richardson & Merlino, *supra* note 66, at 556–57 tbl. 7; Mara L. Merlino, Colleen I. Murray & James T. Richardson, *Judicial Gatekeeping and the Social Construction of the Admissibility of Expert Testimony*, 26 BEHAV. SCI. & L. 187, 196 (2008) (surveying judicial opinions assessing toxicology, damages, and psychological expert testimony and finding frequent use of the testability and error rate factors). *See generally* Edward K. Cheng & Albert H. Yoon, *Does* Frye *or* Daubert *Matter? A Study of Scientific Admissibility Standards*, 91 VA. L. REV. 471 (2005) (demonstrating that removal rates from state to federal courts in *Frye* states did not change after the introduction of the *Daubert* standard, indicating that the analysis of the evidence likely did not change either); Dahir et al., *supra* note 52, at 71; Leah H. Vickers, Daubert, *Critique and Interpretation: What Empirical Studies Tell Us About the Application of* Daubert, 40 U.S.F. L. REV. 109, 137 (2005) (reviewing the literature empirically studying results of the *Daubert* decision and concluding that while "*Daubert* has indeed raised the bar to admissibility . . . judges are not frequently utilizing the reliability factors suggested in the decision").

69. *See* Gatowski et al., *supra* note 48, at 445.

70. *See* Groscup et al., *supra* note 56, at 348–53.

71. *See* Haug & Baird, *supra* note 18, at 745.

though not explicitly discussing a specific error rate, is trying to determine the likelihood that the expert's opinion is distorted by weaknesses in methodology, revealing an implicit error rate problem. Thus, unlike previous studies, our examination of judicial responses to expert testimony considers both explicit and implicit error rate analyses. In addition to this unique (and in our opinion more realistic) coding method, we also attempt to capture the *importance* of error rates analyses to the judges through word counts, similar to Groscup et al.

III. METHOD

A. Identification and Selection of Cases

Our primary aim in this study was to investigate the types of analyses (and the extent of those analyses) undertaken by trial court judges when faced with a challenge to the reliability of an expert witness. Thus, unlike some previous scholars, we elected to study trial court cases rather than appellate cases.[72] Examination of trial-level cases allows us to determine how *Daubert* and its progeny are used in the full range of everyday cases, as opposed to the smaller subset of cases that are appealed. Additionally, we selected only federal cases.[73] We based that choice on the fact that not all states have adopted the *Daubert* standard;[74] some states continue to use the *Frye* standard while others use hybrid standards that combine elements of *Daubert* and *Frye*.[75]

---

72. *See, e.g.*, Groscup et al., *supra* note 56. The Groscup study examined state-level appellate opinions "because of their potential to demonstrate trends in judicial decision making about expert testimony." *Id.* at 342. Haug and Baird examined federal appellate opinions in addition to district court cases. Haug & Baird, *supra* note 18. While appellate opinions may be a better source of determining the likely direction of the law in the future, we are more interested here in assessing how courts of first instance are applying the law already in place—the influence of *Daubert* rather than its evolution. Of course, this means that our results cannot be directly compared to those of the Groscup et al. study.

73. Some past studies have used state cases instead. *E.g.*, Groscup et al., *supra* note 56.

74. Thirty-two states currently follow the *Daubert* standard. MARTIN S. KAUFMAN, ATLANTIC LEGAL FOUND., THE STATUS OF *DAUBERT* IN STATE COURTS (2006), *available at* http://www.atlanticlegal.org/daubertreport.pdf (describing 30 states that had adopted the *Daubert* standard as of March 31, 2006); Robert Ambrogi, *Two More States Adopt* Daubert, *Bringing Total to 32*, BULLSEYE (Oct. 7, 2011), http://www.ims-expertservices.com/bullseye-blog/october-2011/two-more-states-adopt-daubert,-bringing-total-to-32/ (describing the adoption of *Daubert* by Alabama and Wisconsin).

75. KAUFMAN, *supra* note 74. For example, Minnesota follows the "*Frye-Mack*" standard, which adopts the *Frye* general acceptance standard but also includes a requirement that expert evidence have a scientifically reliable foundation. *Id.* at 34.

We searched for cases using the Westlaw database of all federal district court cases. We developed a search string designed to capture cases that contained challenges to the admissibility of expert testimony.[76] As a general matter, the search captured cases that met two primary requirements: (1) a citation to *Daubert* and (2) a word involving an admissibility decision (e.g., "admit" or "admitted") within five words of a phrase describing an expert (e.g., "expert witness" or "expert testimony").[77] Though analysis of expert testimony in federal courts is governed by FRE 702,[78] we did not require that the court cite to the rule, as occasional cases that undertake a full *Daubert* analysis do not do so (though most cases in our sample did). This decision reflects our general approach: we were intentionally over inclusive in the initial search and then removed cases from the sample if they did not address admissibility based on the content of an expert's testimony (e.g., if a report was offered after discovery closed). We sampled 18 years of cases from 1994 (one year after the *Daubert* decision) through 2011.

The search initially captured a total of 6,834 cases. Of the entire population of cases, Westlaw designated 1,107 as criminal cases, or 18% of the total sample. From the population, we randomly selected 208 cases for coding.[79] Though our search string captured more cases in later years,[80] in order to examine potential changes in patterns over time, we used random stratified sampling to sample an equal, random subset of cases from each year. Thus, with our 208 cases across 18 years of

76. The search string was adapted from past similar studies of *Daubert* decisions, including DIXON & GILL, *supra* note 40, at 15–17, and Groscup et al., *supra* note 56, at 342. The exact search string was "509 U.S. 579" & ((admiss! or inadmiss! or admit! or exclud! or preclud! or strike or stricken or unqualif! or qualif! or bar or barred or barring) /5 expert & (witness or testi!)). Note that the requirement of a citation to *Daubert* is captured by the case's United States Reports number rather than a use of the word "Daubert." We used this strategy because cases that contain an analysis of scientific evidence tend to provide the complete citation to *Daubert*, including the reporter number. In contrast, cases that do not involve an analysis of scientific evidence sometimes refer to the *Daubert* decision, without providing a full citation, and we sought to avoid capturing those cases in the search. *See, e.g.*, *Federico v. Lincoln Military Hous., LLC*, No. 2:12cv80, 2013 WL 5409910, at *11 (E.D. Va. Sept. 25, 2013) (discussing a prior case as "raising a *Daubert* challenge" but discussing other aspects of that case); *LendingTree, LLC v. Zillow, Inc.*, No. 3:10-cv-00439-FDW-DCK, 2013 WL 4522512, at *1 (W.D.N.C. Aug. 27, 2013) (case involving timeliness of a party's discovery responses but also mentioning that "the Court has already extended the deadlines for . . . the filing of dispositive and *Daubert* motions"); *Daubert Chem. Co. v. Cigna Prop. & Cas. Co.*, No. 90 C 6587, 1991 WL 113201 (N.D. Ill. June 19, 1991) (containing the name "Daubert" in one of the parties but not dealing with a scientific evidence issue).

77. *See supra* note 76.

78. *See* FED. R. EVID. 702.

79. We used a random number generator in Microsoft Excel to select cases identified by the Westlaw search.

80. *See infra* Table 2.

sampling, we selected approximately 12 cases per year for coding. Because we were also especially interested in how forensic expert testimony would be analyzed by judges, particularly with regard to error rates, we over-sampled criminal cases; while 18% of the population of cases identified by our Westlaw search were criminal cases, our final sample contained about one-third criminal cases.[81] Additionally, many cases in our sample contained admissibility analyses regarding multiple proffered experts. We treated each expert as a separate unit and coded all variables for each expert. Prior to coding, we checked each case to ensure that it contained expert evidence being assessed for its admissibility under FRE 702. Because our initial search was intentionally broad, a number of cases were captured by the search that did not actually contain an FRE 702 analysis.[82]

B. Dependent Variables and Coding Methods

After developing the search string, we constructed a coding rubric using 15 pilot cases that were drawn from the search results but not included in the final sample. Based on the 15 pilot cases and on prior research,[83] we identified six broad categories of coding variables that we applied to the full sample of 208 cases (see Table 1 for summary of all variables). *Identifier variables* provide basic information about the jurisdiction and type of case (e.g., circuit and district in which the case was heard, date of the opinion). *Expert information* describes characteristics of the experts in the case (e.g., the party offering the expert, the type of expert evidence). *Legal standards* code for whether the court mentioned each of the five[84] *Daubert* factors and the way those factors were framed. *Error rate analysis* measures the extent to which the judge analyzed the known or potential rate of error of the testimony—critically, this category includes separate codes for explicit discussion of error rates and for implicit discussion of error rates, as

81. We chose 30% criminal cases to produce a large enough sample of cases involving forensic testimony that would enable us to examine how forensic experts are assessed differently under the *Daubert* factors.

82. Situations in which cases met our Westlaw search criteria but were excluded from coding included: (1) ineffective assistance of counsel cases in which *Daubert* was cited because a lawyer failed to make proper *Daubert* arguments; (2) purely procedural cases, such as Rule 26 cases in which the only question was whether evidence was proffered in a timely manner; (3) *Daubert* cases in which the opinion simply stated that a *Daubert* hearing would be required; (4) slip opinions that stated the admissibility outcome from the *Daubert* hearing but not the reasoning; and (5) other cases that cited *Daubert* and contained expert evidence, but did not contain any *Daubert* analysis, such as summary judgment rulings in which the reliability of an expert was not contested.

83. *See supra* notes 38–52 and accompanying text.

84. We include the maintenance of standards factor in our analyses here.

detailed below. *Other*-Daubert-*factors* analysis measures the extent of any analysis involving the other three nonerror *Daubert* factors, applying them to the facts of the case. Finally, *other admissibility analysis* measures the extent of any analysis of the expert evidence involving legal standards not provided in *Daubert* but still applicable to the case (e.g., relevance, expert qualifications). We explain each of these six categories in further detail below.

C. General Coding Principles

Most of our primary dependent variables involved counts of the number of words a judge spent on a particular type of analysis.[85] We coded all citations as part of the analysis. We attempted to code single paragraphs in single categories where possible, only coding multiple variables in a single paragraph where the distinction was clear.[86] Although most of the passages we coded fell in a single category, on occasion the language straddled two categories, and the passage was counted in both categories.

On some occasions, judges conducted analysis of the reliability of expert testimony but then stated that the problems with the testimony were not great enough to rule the testimony inadmissible. When this was the case, we still coded the discussion; that is to say, our coding was outcome independent. Regardless of whether the factor cut in favor of admission or in favor of rejection, we still considered it analysis because the judge was assessing the evidence under the factor being discussed.[87]

---

85. We discuss the pros and cons of using word counts as a measure *infra* Part VI.

86. This decision is particularly important because a court's analysis of an expert's testimony based on the *Daubert* factors is frequently bound up with a court's descriptive statements of the parties' arguments and the expert's claims. In general, if the paragraph containing the assessment also included those descriptions of claims and objections, we coded the entire paragraph in the relevant category. In the example below, the first sentence is merely a statement of what the defendant argued rather than peer-review discussion; but because the sentence naturally leads into the peer-review discussion and cannot easily be separated from it, the entire paragraph was coded as peer-review discussion: "Defendant also claims that Laughery 'never produced literature substantiating his opinion the warning systems were inadequate.' Laughery cites to five peer-reviewed publications that he authored or edited which are relevant to his testimony in this case . . . ." *Cochran v. Brinkman Corp.*, No. 1:08-cv-1790-WSD, 2009 WL 4823858, at *13 (N.D. Ga. Dec. 9, 2009) (internal citation omitted) (the list of five publications Laughery cited is included in the case at *13). Thus, our general rule was that where a paragraph consisted mainly of a single type of codable analysis but other non-codable discussion was bound up in that paragraph, we coded the full paragraph as the single codable type of analysis.

87. For example:

We did not code passages in which the judge simply mentioned that a party had raised an objection to an expert based on a particular factor and that evaluation of that factor would go to weight, rather than admissibility. In these cases, where the judge did not conduct any analysis of the expert's methodology but instead simply said that the jury would be able to make the proper determination, we did not code the discussion as analysis, as the judge was not evaluating the expert's methodology.[88]

D. Dependent Variables

1. IDENTIFIER VARIABLES

In addition to basic identifying information, such as case name, citation, and year, we coded for the case's jurisdiction (i.e., district and circuit of the case). District courts from all 11 federal circuits were represented in the sample. We did not stratify based on jurisdiction, as we had no hypotheses of differences in analysis by jurisdiction. We also coded cases according to the primary substantive dispute category, adopting (with some modifications) the categories used in Groscup et al.

> Could [the expert] have offered more scientific support for his conclusions? Absolutely. He could have included studies showing the risk of battery failure, or the risk of electronic interference, or the ease with which hearing aids can become dislodged during physical confrontations, but it is not necessary that he include these things in order for the Court to find his testimony reliable. Contrary to the suggestions offered by Allmond's arguments, Rule 702's reliability requirement does not mean that the expert's testimony is beyond refute. Indeed, many of the bases put forth by Kramer may not withstand challenge under cross-examination. Nevertheless, the ultimate conclusion that he puts forth is supported by reasons, drawn from his training and education and experience in the field, and sufficient to satisfy the reliability requirement of Rule 702.

Allmond v. Akal Sec., Inc., No. 4:05-cv-96(HL), 2007 WL 988757, at *5 (M.D. Ga. Mar. 29, 2007). Though the judge was clear that the expert did not account for all possible alternatives in his methodology, the judge's view was that enough scientific support was shown such that the evidence was sufficiently reliable. *Id.* We consider this an analysis—the judge examined the methodology undertaken by the expert and evaluated whether it was sufficient under *Daubert*.

88. For example: "Defendant notes that Dr. Paul's findings have not been verified by others in his field. That, however, is not in itself a sufficient reason to exclude his testimony." *Colombo v. CMI Corp.*, 26 F. Supp. 2d 574, 576 (W.D.N.Y. 1998); s*ee also FDIC v. Suna Assocs., Inc.*, 80 F.3d 681, 687 (2d Cir. 1996) (citing *Daubert* for the proposition that "publication . . . does not necessarily correlate with reliability"). "Indeed, '[d]isputes as to the strength of [an expert's] credentials, faults in his use of . . . a methodology, or lack of textual authority for his opinion, go to the weight, not the admissibility of his testimony.'" *Colombo*, 26 F. Supp. 2d at 576 (quoting *McCullock v. H.B. Fuller Co.*, 61 F.3d 1038, 1044 (2d Cir. 1995)).

(2002).[89] When a case involved more than one category of law, we coded the case for the single legal category that was most predominant.

2. EXPERT INFORMATION

As mentioned above, we coded each expert separately if the judge conducted separate assessments on the admissibility of the experts in the case. For each expert, we coded:

- the party offering the evidence (plaintiff, civil defendant, prosecution, or criminal defendant);
- whether the expert was admitted to testify (fully admitted, fully rejected, or partially admitted); and
- the type of expert offered, adapting the categories used in Groscup et al. (2002).

Like Groscup et al., we separated experts into four primary categories: medical/mental health experts, technical/engineering experts, scientific experts, and business experts. We also included forensic experts under the technical/engineering category, but conducted some separate analyses of the forensic expert group, based on our hypotheses.[90] We also divided scientific experts into natural science experts and social science experts.[91]

3. LEGAL STANDARDS

We began by coding whether the court, in its description of the law governing expert evidence, mentioned each of the five *Daubert* factors: testability, peer review, known or potential rate of error, general acceptance, and maintenance of standards. Note that in this first

89. Civil cases were divided into the following categories: tort, contract, property, intellectual property, habeas, civil rights, bankruptcy, tax, antitrust, deportation, employment, and other. Criminal cases were divided into the following categories: drug, violent, sex crime, fraud, theft, conspiracy, and other. Intercoder reliability in identifying case types was 1.0 across the 20 test cases.

90. *See infra* Parts IV.F., V.G.

91. The more complete composition of the categories was as follows: medical and mental health experts (consisting of examining physicians, pediatricians, social workers, psychiatrists, clinical psychologists, and other medical experts), forensic and police procedure experts, technical and engineering experts (consisting of accident reconstruction experts, fire & arson experts, and other engineering experts), natural scientific experts (consisting of chemists, biologists, physicists, and other natural science experts), social science experts (consisting of experimental psychologists, economists, and other social scientists), and business experts (consisting of accountants, business practice experts, attorneys, securities experts, and other business experts).

enumeration of judicial activity, we were not coding whether the judge undertook an analysis of the expert evidence based on each factor, but rather whether the judge mentioned the factor in describing the law that was to be applied. For all factors other than error rate, we coded a binary yes/no decision—whether or not the judge mentioned the factor in his description of the law.[92]

Because of our particular interest in the way judges understand the error rate factor of *Daubert*, in addition to coding whether the court mentioned the factor in its outline of the law, we also coded the way that the court framed the factor as one of three possibilities:

- a direct quote or paraphrase of the *Daubert* language (which leaves ambiguous whether there only must be an error rate provided in order to satisfy the factor or whether the error rate must stay below a certain threshold);
- a "simple provision" standard, in which merely providing an error rate satisfies the factor; or
- a "threshold" standard, in which the judge must decide whether the method's known or potential rate of error is low enough to be acceptable.

Because *Kumho Tire*'s language implied that the threshold standard should be used, we also coded for whether the court cited *Kumho Tire* in explaining the error rate factor.

4. ERROR RATE ANALYSIS

We coded for two types of error rate analysis: explicit and implicit. At a very general level, the two can be thought of as two faces of the same coin: explicit error rate analysis occurs when the judge directly discusses whether he can assess the rate of possible error of the method or testimony and, in some cases, whether that rate of error is acceptable. Implicit error rate analysis is also aimed at determining how likely it is

92. Frequently, the discussion of the factors would be in a single paragraph as in this example:

> In [*Daubert*], the Supreme Court identified four non-exclusive factors that may be helpful to the court in assessing the relevance and reliability of expert testimony, including (1) whether a theory or technique has been tested; (2) whether the theory or technique has been subjected to peer review and publication; (3) the known or potential error rate and the existence and maintenance of standards controlling the theory or technique's operation; and (4) the extent to which a known technique or theory has gained general acceptance within a relevant scientific community.

Moore v. Weinstein Co., No. 3:09–CV–00166, 2012 WL 1884758, at *3 (M.D. Tenn. May 23, 2012).

that the expert will give erroneous testimony to the trier of fact. It is done, however, through analysis of the quality of the methodology itself, without explicitly discussing error. We discuss our coding methods for each of these analyses in more detail below, with examples given in the footnotes.

a. Explicit error rate analysis

We considered an "explicit error rate analysis" to be what is traditionally thought of as discussion of *Daubert*'s error rate factor. Any situation in which the court directly discussed the quantitative error rate of a method was coded as explicit error rate analysis.[93]

---

93. We also looked to several cues in coding explicit error rate, though this is not an exhaustive list:

- language referencing the error rate factor itself (e.g., "Turning to the error rate factor" or "Expert X's testimony does not satisfy the rate of error factor for the following reasons");
- frequent use of the word "error," synonyms of error, or related words (e.g., mistake, false, miscalculation, accuracy) in the analysis (e.g., "Expert X's methods have a very high rate of error" or "Based on discussion of past research, the odds of Expert X making a mistake are high");
- any discussion of signal detection terms, such as "false positive," "false negative," "hit," or "miss;"
- discussion of related studies or experiments that make conclusions regarding error rates; and
- discussion of terms indicating the diagnosticity of the test (e.g., "Doctor X is highly accurate in making diagnoses in this field").

An explicit error rate discussion may have any combination of some of these factors, though at least one was nearly always present. For example:

> There is evidence that PMRB can be distinguished from environmental banding within an acceptable rate of error. A group of FBI analysts, led by Stephen Shaw, conducted a study for which they collected 600 hairs and subjected them to a range of environmental conditions. Although these hairs exhibited signs of decomposition, they did not present PMRB. These hairs were then mixed with hairs known to have come from deceased subjects. According to the abstract of the study (whose publication is forthcoming), two hair examiners were able to distinguish post-mortem root-banded hairs from environmentally-banded hairs with 99.5% accuracy. When the two examiners double-checked each other's work, their accuracy increased to 100%. Suffice it to say, this is a tolerable error rate.

Kogut v. Cnty. of Nassau, 894 F. Supp. 2d 230, 243 (E.D.N.Y. 2012). Explicit error rate discussion may also be even simpler, merely stating whether an error rate is present or not: "There is no information on the known or potential rate of error of the technique, nor the existence and maintenance of standards controlling its operation." *Banta Props., Inc. v. Arch Specialty Ins. Co.*, No. 10–61485–CIV, 2011 WL 7118542, at *4 (S.D. Fla. Dec. 23, 2011) (citation omitted). Another example: "Because Kelsey has not conducted any experiments or testing of any kind, there cannot be a known rate of error for his results.

b. Implicit error rate analysis

A major innovative feature of our analysis is the identification and measurement of implicit error rate analysis. It recognizes the range of methodological assessments a judge may engage in even when the judge does not explicitly label her assessment as an evaluation of error rate. We define an implicit error rate analysis as a direct assessment of the validity of the method at issue that speaks to the potential rate of error of the test where that rate of error is unknown. Unlike discussion of the peer review and general acceptance factors, the issues discussed in an implicit error rate analysis are substantive critiques of the methods used by the expert, not external considerations like others' opinions of the method. An implicit error rate analysis might be characterized as an analysis in which the judge is attempting to discern the likely accuracy of the expert even if an error rate has not been explicitly provided.

Because the implicit error rate analysis is an assessment of validity, it can be broken down into the three major categories of threats to scientific validity, which we briefly outline here and explain in full detail in Appendix A:

- construct validity (the extent to which the expert's measurements properly reflect what they purport to measure), which includes unwarranted extrapolations, sampling biases, improperly operationalized variables, or experimental confounds;
- external validity (the method's generalizability outside of the unique setting of the testing itself), which often deals with the question of whether a sampled population is similar enough in relevant ways to the population in question in the case (e.g., drawing conclusions regarding human disease from an animal study of the same disease); and
- internal validity (the extent to which a methodology can accurately determine whether a cause—effect relationship exists), including the inclusion of appropriate controls and the ability to rule out competing hypotheses.

Judges in our sample on occasion referred to the *ipse dixit* problem as outlined in *General Electric Co. v. Joiner*.[94] In that case, the Court opined that *Daubert* does not require a court to admit "evidence which is connected to existing data only by the *ipse dixit* of the expert. A court

Likewise, there is no evidence concerning a potential rate of error." *Pillow v. Gen. Motors Corp.*, 184 F.R.D. 304, 308 (E.D. Mo. 1998) (citation omitted).

94. 522 U.S. 136 (1997).

may conclude that there is simply too great an analytical gap between the data and the opinion proffered."[95] As mentioned above, in some cases this may be an issue of construct validity or external validity, but in other cases the judge may opine that the expert is simply speculating rather than basing his opinion on data.

Thus, there are two situations where an expert may make "too great of an analytical leap" to arrive at a conclusion: (1) the expert may have conducted some type of testing or based an opinion on some scientifically derived data, but those data could not justify the conclusion that was drawn, or (2) the expert did no testing at all, but simply drew a conclusion. The *ipse dixit* problem as described in *Joiner* falls into Category 1. In *Joiner*, the plaintiff's experts testified that the plaintiff's cancer was caused by the defendant's chemicals, citing laboratory animal studies rather than epidemiological studies.[96] The Supreme Court agreed with the lower courts that the studies the experts examined could not be used to draw conclusions about the cause of cancer in humans and criticized the defense for failing to "explain[] how and why the experts could have extrapolated their opinions from these seemingly far-removed animal studies."[97] The problem was not that the methods were not scientific or were untested, but rather that their conclusions were unwarranted based on the data; essentially an external validity question. We consider this to be an implicit error rate analysis. As in cases with other external validity questions, the judge attempted to determine how likely it was that the experts' opinion was incorrect or misleading by assessing the strength of his analytical methods. Likewise, in discussing the *ipse dixit* problem, the *Joiner* Court cited *Turpin v. Merrell Dow Pharmaceuticals*,[98] a case that also involved extrapolation of animal studies.[99] In that case, the Sixth Circuit Court of Appeals similarly concluded that "[t]he analytical gap between the evidence presented and the inferences to be drawn on the ultimate issue of human birth defects is too wide."[100] Based on these studies, we conclude that the Supreme Court was referring to Category 1 when talking about *ipse dixit*—cases where there is some testing or scientific methodology, but the conclusion provided by the expert goes beyond what can be justified by the data and is therefore inadmissible.

95. *Id.* at 146.
96. *Id.* at 143.
97. *Id.* at 144.
98. 959 F.2d 1349 (6th Cir. 1992).
99. *Id.* at 1350.
100. *Id.* at 1360.

In Category 2 situations, the judge refers to the *ipse dixit* rule in a broader way.[101] No testing has been done and an expert is simply drawing a conclusion out of thin air. We did not consider this to be a technical *ipse dixit* problem, regardless of what the trial judge called it, because there cannot be "too great of an analytical leap" when there is no testing or data set to "leap" from. The problem instead is the more foundational one of testing.

Thus, we carefully distinguished between analysis that discussed testability (i.e., the judge stated that the expert, in the present case, did not conduct any testing and thus could not make a valid statement) and the *ipse dixit* problem that implicates implicit error analysis (the expert did conduct or use some type of test, but could not make the leap from that test to the opinion to be offered). In the above examples from *Joiner* and *Turpin*, other parts of the opinions indicated that the experts did conduct testing but could not validly arrive at their conclusions based on their results. The example below, however, would be coded as a testability analysis because the court opined that there was no testing conducted whatsoever: "[The expert] did not conduct any physical testing of the fryer's resistance to tipping over or other fryer models' resistance to tipping over. Edmondson employs nothing more than 'a subjective, conclusory approach that cannot reasonably be assessed for reliability.'"[102] As a whole, we only coded a passage as an *ipse dixit* analysis, and thus an implicit error rate analysis, where there was (1) explicit mention of *ipse dixit* or (2) a characterization of the expert's analysis as speculative, but based on at least some testing or data.

5. OTHER *DAUBERT* ANALYSIS

We also counted the number of words spent analyzing the expert testimony under the remaining four nonerror rate *Daubert* factors. We briefly discuss each factor below.

a. Testability

Though testability is only a single *Daubert* factor, we coded for two types of analysis under the factor in our sample: *measurement* and *testability*. Though not one of the *Daubert* factors, measurement analysis is highly related to the testability analysis, and the two often occurred together. While the testability analysis asks the question of whether a

101. *See, e.g.*, *Thomas v. Novartis Pharm. Co.*, 443 F. App'x 58, 61 (6th Cir. 2011).

102. *Cochran v. Brinkman Corp.*, No. 1:08-cv-1790-WSD, 2009 WL 4823858, at *9 (N.D. Ga. Dec. 9, 2009) (citations omitted).

method itself has been or can be tested to determine its validity, or whether the expert has conducted some analysis of data in the instant case to test a hypothesis, a measurement analysis is much more limited. When conducting a measurement analysis, a judge simply examines whether some data had been collected, such as whether an expert conducted physical measurement of the length, weight, or composition of an object relevant to the case.[103]

Frequently, a judge did a measurement analysis immediately before undertaking a testability analysis in which she examined whether the expert took the data collected and conducted some type of testing.[104] The hallmark of a measurement analysis is a discussion of whether the expert *collected* any sort of data. Such data could come in a variety of forms, such as photographs or observation and examination of a crime scene. The subsequent testability analysis, in contrast, examines whether the expert took the measurements or other data obtained and *did some analysis* to arrive at a conclusion.

b. Peer review and general acceptance

Peer review and general acceptance analysis frequently occurred together, as both examine data external to the methodology itself in order to assess evidentiary reliability. These factors were often explicitly

---

103. For example:

> The court will not exclude Mr. Dega's testimony because it is based in part upon general engineering principles—indeed, it would be of great concern if they were not. GM's contention that Mr. Dega simply "jettisoned analysis of facts for application of a general engineering principle" is also unavailing. Mr. Dega's investigation was based upon his own measurements of surface roughness and machine lead on the torquemeter shaft at issue, as well as extensive data gathered from the Navy Court of Inquiry Report and Warren Lieberman's analysis of the amount of oil which leaked from the crash airplane. "Analysis of facts" was clearly part of Mr. Dega's methodology.

Stecyk v. Bell Helicopter Textron, Inc., No. CIV.A. 94–CV–1818, 1998 WL 599256, at *3 (E.D. Penn. Sept. 9, 1998).

104. For example, this quote came immediately following the quote provided in *supra* note 103, and it was coded as a testability analysis:

> GM also contends that Mr. Dega's failure to perform independent tests to support his conclusion that a surface finish of 69 microinches would contribute to leakage requires exclusion of his testimony . . . . Here, Mr. Dega not only relies upon physical evidence of improper seal installation, excessive surface roughness and machine lead on the torquemeter shaft, and other documentary evidence compiled by the Navy Court of Inquiry and Warren Lieberman, but the results of Mr. Dega's own investigation do not undermine his ultimate opinion that the seal and torquemeter shaft were defective. The "analytical gap" between data and opinion which was present in *Childs* does not exist in this case.

Id.

introduced by the courts and discussed in a clear, cabined analysis, making them easier to code than other factors.[105] We measured discussion of each factor separately. Judges frequently discussed publications under the peer review factor. Similarly, the general acceptance factor was often clearly labeled as general acceptance by the courts (e.g., "[the expert] himself admits he does not know if his methods are widely accepted or what other methods might be widely accepted").[106] We also considered any discussion of agreement from other related experts as general acceptance, including instances where the judge pointed out that the opposing expert accepted an expert's methods as reliable or used similar methods.

c. Maintenance of standards

Though rarely used (and often not even mentioned in the discussion of the law), we also measured discussion of the maintenance of standards factor (e.g., "the expert did not follow the standards accepted in his field").

6. OTHER ADMISSIBILITY ANALYSIS

In addition to the *Daubert* factors, we measured discussion of four other admissibility considerations: qualifications, relevancy, whether the testimony was generated for the purpose of the litigation, and FRE 403 balancing. For consistency in coding for qualifications analysis, we coded any time the judge listed an expert's qualifications, not just when he assessed those qualifications under FRE 702. We coded qualifications in this way because the qualifications "analysis" was frequently just a statement that the list of achievements was sufficient for the purpose of FRE 702.

a. Case weighting

Because our random stratified sampling approach disproportionately selected criminal cases and cases in the first few years after the *Daubert* decision, we calculated a weighting variable to reflect the actual frequency of each case type (civil vs. criminal) and year combination in the population. For most of our analyses, weighting the variables did not lead to any differences, so we present data and statistics on the

105. This is demonstrated by our high intercoder reliability scores for these two factors: .97 for the peer review factor and .99 for the general acceptance factor.

106. *Marting v. Crawford & Co.*, No. 00 C 7132, 2004 WL 305724 (N.D. Ill. Jan. 9, 2004).

unweighted values. Where the weighting did lead to different outcomes, we point out the difference in the footnotes.

b. Intercoder reliability check

After developing and finalizing the coding rubric on the initial 15 pilot cases, we randomly selected 20 cases from our sample to be independently coded by a second coder in order to assess the reliability of coding for each of our dependent variables.[107] We evaluated the results using the Smith index[108]—twice the number of agreements in a category divided by the sum of the frequency that each rater used that category.[109] The reliability ranged from 0.73 to 1.00, averaging 0.86 across the 20 cases.

107. The coding of the full sample of 208 cases was done by the first author; the 20 reliability cases were also coded by the second author.

108. Charles P. Smith, *Content Analysis and Narrative Analysis*, *in* HANDBOOK OF RESEARCH METHODS IN SOCIAL AND PERSONALITY PSYCHOLOGY 313–35 (Harry T. Reis & Charles M. Judd eds., 2000). In general, reliability indicators at the levels achieved here are viewed as having "almost perfect" reliability. *See, e.g.*, J. Richard Landis & Gary G. Koch, *The Measurement of Observer Agreement for Categorical Data*, 33 BIOMETRICS 159, 165 (1977) (characterizing the strength of different agreement values).

109. In calculating this measure, we accounted for whether the coders applied the measure to the same point in the text. So, for example, if each coder coded 60 words of qualifications analysis, but only 40 of those words overlapped, the analysis would only consider the coders as having agreed on 20 words. Thus, under the Smith index, the reliability for such a scenario would be (40*2)/(60+60) = .66.

TABLE 1. PRIMARY DEPENDENT VARIABLES AND CORRESPONDING RELIABILITY SCORES

Dependent Variable	**Category**	**Smith Index**
Circuit	Case Identifier	N/A
Case Type	Case Identifier	N/A
Offering Party of Expert	Expert Information	N/A
Expert Type	Expert Information	N/A
Admissibility Decision	Expert Information	N/A
Testability Factor	Legal Standards	1.00
Peer Review Factor	Legal Standards	1.00
General Acceptance Factor	Legal Standards	1.00
Error Rate Factor	Legal Standards	1.00
Error Rate Type	Legal Standards	1.00
Standards Factor	Legal Standards	1.00
Kumho Tire Citation	Legal Standards	1.00
Explicit Error Rate Analysis	Error Rate Analysis	0.98
Implicit Error Rate Analysis	Error Rate Analysis	0.89
Testability Analysis	Other *Daubert* Analysis	0.89
Measurement Analysis	Other *Daubert* Analysis	0.87
Peer Review Analysis	Other *Daubert* Analysis	0.98
General Acceptance Analysis	Other *Daubert* Analysis	0.97
Standards Analysis	Other *Daubert* Analysis	0.89
Qualifications Analysis	Other Admissibility Analysis	0.99
Relevancy Analysis	Other Admissibility Analysis	0.79
Generated for Litigation Analysis	Other Admissibility Analysis	0.73
403 Balancing Analysis	Other Admissibility Analysis	1.00

IV. OUR HYPOTHESES

A. Hypothesis 1: Judges Will Allocate More Discussion to Implicit Error Rate Analysis than to Other Daubert *Factors in Assessing the Evidentiary Reliability of Expert Evidence*

Our chief aim in this study was to examine the extent to which judges *implicitly* use a broad conception of the error rate factor of *Daubert* by scrutinizing evidence in an attempt to identify the potential rate of error of an expert in giving an opinion to the trier of fact. Past literature seems to imply that judges either generally fail to understand the error rate factor as compared to more peripheral factors like peer review and general acceptance[110] or may simply not find error rates to be important criteria, and thus they may spend less time evaluating them.[111] However, we suspected that while the vague nature of the "known or potential rate of error" as defined in *Daubert* may lead judges to make little *explicit* use of the factor (because of confusion as to its importance and breadth, as well as uncertainty as to how to apply it), we anticipated that judges would extensively discuss error rates *implicitly*. Because the rate of error is the primary *Daubert* factor that speaks to the substantive quality of the scientific evidence itself, we expected that it would be heavily used through critiques of the expert's methods in an effort by the judge to assess the likely rate of error when one is not explicitly provided. We expected such discussion to be more prevalent than discussion regarding more peripheral factors such as qualifications, peer review, and general acceptance.[112] Essentially, we expected that judges would spend more time assessing the quality of the science itself than they would assessing proxies for the quality of the science such as peer review and general acceptance.

---

110. *See, e.g.*, Gatowski et al., *supra* note 48, at 445.

111. *See, e.g.*, Groscup et al., *supra* note 56; Haug & Baird, *supra* note 18.

112. The implicit error rate analysis can be conceptualized as central processing within the dual-process elaboration likelihood model of persuasion first described by Richard Petty and John Cacioppo. RICHARD E. PETTY & JOHN T. CACIOPPO, COMMUNICATION AND PERSUASION: CENTRAL AND PERIPHERAL ROUTES TO ATTITUDE CHANGE (1986). Central processing involves careful scrutiny of the message of the evidence itself. *Id.* In contrast, qualifications, peer review, and general acceptance analyses involve primarily peripheral processing, relying on the perceived credibility of the source of the testimony. *See id.* Likewise, implicit error rate analysis would also be considered systematic processing under the heuristic-systematic model of information processing proposed by Shelly Chaiken, while qualifications, peer review, and general acceptance analyses more closely align with heuristic processing. *See* Shelly Chaiken, *Heuristic Versus Systematic Information Processing and the Use of Source Versus Message Cues in Persuasion*, 39 J. PERSONALITY & SOC. PSYCHOL. 752 (1980).

B. Hypothesis 2: The Length of the Judge's Implicit Error Rate Discussion in an Opinion Will Predict the Outcome of the Admissibility Inquiry: When Judges Devote More Discussion to Implicit Error Rate Analysis, They Will Be More Likely to Reject All or Part of the Expert's Evidence than to Fully Admit It

Prior empirical work on *Daubert* has found a relationship between the amount of scrutiny placed on evidence by a judge and the admissibility decision made by the judge: as more *Daubert* and related factors are mentioned, the evidence is more likely to be ruled inadmissible.[113] One potential explanation is relatively straightforward: when judges identify methodological issues that raise questions about the reliability of the evidence, they are more likely to reject the evidence. An increased number of factors discussed is likely to mean an increased number of problems raised and a resulting increased likelihood of rejection. We expected to find a similar pattern in our word counts: as the length of discussion of the *Daubert* factors increased, we expected to find a reduced likelihood that the evidence would be admitted. Additionally, in accordance with our theory that judges recognize the importance of error rates in judging the adequacy of expert testimony (at least implicitly), we expected that the pattern would be particularly prominent for implicit error rate discussion. Because the implicit error rate task is central to the goal of *Daubert* (identifying the evidentiary reliability, or validity, of the method itself), we expected that this type of discussion would be especially predictive of outcomes. That is, extensive discussions of error rates should occur when the court analyzes flaws in the scientific methodology, and the length of such discussions should be predictive of the admissibility of the evidence.

C. Hypothesis 3: Judges Will Show Confusion as to Whether the Error Rate Standard Incorporates a Threshold Requirement

The Supreme Court in *Daubert* was ambiguous (perhaps intentionally so) in identifying precisely what is necessary to satisfy the error rate factor.[114] Critically, the Court did not make a clear statement as to whether scientific evidence must produce an error rate that is below a particular acceptable threshold as decided by the trial judge (termed here a "threshold" standard), or whether it is enough that the expert can simply provide a valid error rate for the trier of fact to incorporate in assessing the value of the expert's testimony (termed here a "simple provision" standard). Complicating matters further, in *Kumho Tire* the

113. *See, e.g*, Merlino, Murray & Richardson, *supra* note 68, at 200 tbl.5.

114. *See supra* notes 30–36 and accompanying text.

Court phrased the error rate standard differently, requiring the judge to question whether "there is a high known or potential rate of error,"[115] a clear threshold standard. However, the Court did not signal any intent to change the substance of the factor and left lower courts to notice the updated version of the formulation on error rates.[116] Based on this murkiness, we expected some confusion among the trial courts. Because most courts are likely to cite *Daubert* itself in explaining the factors, we expected that the majority of trial courts would use the ambiguous standard from that decision, though we anticipated that a substantial number of courts would also refer to the factor using either a threshold standard or a simple provision standard. Additionally, we expected that more judges would turn to a threshold standard instead of a simple provision standard following the *Kumho Tire* decision, given its threshold characterization.

D. Hypothesis 4: Implicit Error Rate Discussion Will Be More Prevalent When the Judge Is Assessing Social Science and Business Experts as Compared to Natural Science and Medical Experts

Daubert assigns a difficult task to the trial judge: assessment of the evidentiary reliability (i.e., validity) of unfamiliar expert methodology in fields in which the trial judge likely has no experience.[117] Implicit error rate analysis requires the judge to engage with the evidence at a fairly high level—the judge must be able to examine the methodology for logical and scientific flaws and assess how critical those flaws are in causing potential error. While the adversarial system no doubt helps to educate the judge through briefs and oral argument regarding admissibility of opposing experts, we expected that judges would be more comfortable engaging in implicit error rate analysis in contexts in which they feel more qualified to assess the quality of the methodology itself. Specifically, we expected that judges would devote more discussion to implicit error rates when assessing business and social science testimony (including economics) because judges are likely to have (or feel they have) greater familiarity with these fields than with fields involving more basic scientific or medical methodology, in which most judges likely have little training or experience.

---

115. *Kumho Tire Co. v. Charmichael*, 526 U.S. 137, 149 (1999) (internal quotation marks omitted).

116. *See supra* notes 34–35 and accompanying text.

117. *See supra* note 8 and accompanying text.

E. Hypothesis 5: Analysis of the "External" Factors—Peer Review and General Acceptance—Will Not Vary By the Type of Expert Evidence Presented

In contrast with Hypothesis 4, we anticipated that for "external" factors—those that do not require the judge to directly examine the expert's methodology for flaws—judicial discussion would not vary across type of expert evidence. Assessing whether an expert's methodology has undergone peer review should be (or should appear to be) no more difficult for natural science or engineering evidence than it is for social science or business evidence. It is possible that some types of expert evidence are more likely to undergo peer review, which may affect the amount of discussion devoted to that factor, but we had no *a priori* hypotheses regarding which categories of expert testimony were most likely to have been peer reviewed. Likewise, conducting an assessment of general acceptance requires an examination of the related scientific community regardless of the type of evidence offered, and thus it should not be any more or less demanding across different evidence categories.

F. Hypothesis 6: Explicit Error Rate Discussion Will Be More Prevalent in Response to Forensic Testimony, Where Individuation/Identification Testimony Is Common, as Compared to Other Types of Expert Testimony

Forensic testimony is more likely to elicit explicit error rate discussion simply because it often involves specific identification methods that have a clear truth value that could potentially be assessed and verified through testing, albeit not in the case at hand. For example, fingerprint examiners typically make individuation statements, declaring that the latent fingerprint left at a crime scene matches the defendant's fingerprint to the exclusion of other individuals' fingerprints. Such a statement is clearly true or false, and in theory, an expert's ability to make such discrimination could be observed through systematic testing. Similar statistical information, such as a random match probability in DNA testing, may be more frequently present for forensic testimony. This type of testimony naturally leads to the question of accuracy since the critical testimony is often a binary match/mismatch statement. Thus, we anticipated greater explicit error rate discussion for forensic experts as compared to experts in other domains.

V. RESULTS

A. Descriptive Characteristics of the Cases and Experts

Table 2 presents the total number of cases coded in our sample, including cases coded for the reliability check. Many cases in our sample contained two or more experts. Each of these experts was coded uniquely in our analysis. Our 208-case sample contained analyses of 272 experts.[118] 32% of the total cases were criminal and 68% civil. 29% of the total experts were derived from criminal cases and 71% from civil cases. Fewer cases were drawn from the years between 1994 and 1996 due to the low number of eligible criminal cases during those years. We coded fewer civil cases in those years to maintain our approximate one-third ratio of criminal cases per year.

118. On several occasions, a judge analyzed two separate experts simultaneously and in a way in which the analysis of each expert could not be parsed. In those cases, we coded all experts that were assessed simultaneously as a single expert. *See, e.g.*, *Gruener v. Ohio Cas. Ins. Co.*, No. 1:03–cv–780, 2005 WL 5988665 (S.D. Ohio Aug. 17, 2005).

TABLE 2. CIVIL AND CRIMINAL CASES CODED BY YEAR

Year	Civil Cases	Criminal Cases	Total Cases	Civil Experts	Criminal Experts	Total Experts
1994	9	1	10	15	2	17
1995	6	4	10	8	4	12
1996	6	4	10	11	4	15
1997	8	4	12	8	6	14
1998	8	4	12	8	4	12
1999	8	4	12	15	4	19
2000	8	4	12	8	4	12
2001	8	4	12	12	4	16
2002	8	4	12	10	6	16
2003	8	4	12	14	7	21
2004	8	4	12	12	4	16
2005	8	4	12	16	4	20
2006	9	3	12	11	3	14
2007	8	4	12	8	6	14
2008	8	4	12	8	4	12
2009	8	4	12	8	4	12
2010	8	4	12	14	4	18
2011	8	4	12	8	4	12
Total	142	68	210	194	78	272
Percentage	68%	32%	-	71%	29%	-

Over half (56%) of the experts in the civil cases were offered by plaintiffs, and those experts were fully rejected and fully admitted at a roughly equivalent rate (see Table 3). Civil defendant experts constituted only 22% of the civil experts, and they were more likely to survive a *Daubert* challenge than were civil plaintiff experts.[119] As reported

119. $\chi^2 = 6.56$, $p = .01$. For the purposes of this analysis, we combined the "fully admitted" and "partially admitted" categories to form a single "admitted" category along with the "fully rejected" category, and we compared civil plaintiff experts and civil defendant experts across this measure.

elsewhere in the literature, criminal defendants had the worst admission rates of all parties, with 57% of all experts fully rejected, significantly greater than the rejection rate of criminal prosecution experts.[120]

TABLE 3. ADMISSIBILITY OUTCOMES OF EXPERT TESTIMONY BY PARTY[121]

Outcome	Civil Plaintiff	Civil Defendant	Criminal Prosecution	Criminal Defendant
Fully Rejected	43% (40%)	21% (17%)	26% (29%)	57% (67%)
Partially Admitted	15% (15%)	21% (23%)	21% (15%)	23% (20%)
Fully Admitted	42% (46%)	57% (61%)	53% (56%)	20% (13%)
Total Experts Offered by Party	153	42	47	30

The types of experts offered by the various parties differed as well (see Table 4). Notably, medical and engineering experts were commonly offered by civil plaintiffs, likely arising from the high number of personal injury and product defect cases. Medical experts were also frequently offered by criminal defendants (typically mental health experts). Forensic experts were almost entirely offered in criminal cases, usually by the prosecution.

120. $\chi^2 = 7.56$, $p = .006$. For the purposes of this analysis, we combined the "fully admitted" and "partially admitted" categories to form a single "admitted" category along with the "fully rejected" category, and we compared criminal prosecution experts and criminal defendant experts across this measure.

121. Parentheses indicate weighted averages.

TABLE 4. DISTRIBUTION OF EXPERT CATEGORIES OFFERED BY EACH PARTY TYPE

Expert Type	Civil Plaintiff	Civil Defendant	Criminal Prosecution	Criminal Defendant	Total
Medical	31% (21%)	17% (17%)	9% (12%)	50% (67%)	27% (21%)
Forensic	2% (1%)	2% (0%)	68% (57%)	23% (13%)	16% (10%)
Engineering	34% (34%)	36% (38%)	4% (7%)	0% (0%)	25% (29%)
Natural Science	5% (7%)	12% (11%)	4% (5%)	0% (0%)	6% (7%)
Social Science	13% (12%)	14% (14%)	9% (14%)	23% (20%)	13% (13%)
Business	16% (25%)	19% (20%)	6% (5%)	3% (0%)	13% (19%)

Admissibility outcomes across the expert categories also varied. Medical experts were the most frequently rejected category, with 52% of the experts' testimony fully rejected and only 30% fully admitted. In contrast, natural and social science experts were fully admitted in well over half of the cases in which they were offered, with forensic, engineering, and business testimony falling in the middle.

TABLE 5. ADMISSIBILITY OUTCOMES OF EXPERT TESTIMONY BY EXPERT CATEGORY

Outcome	Medical	Forensic	Engineering	Natural Science	Social Science	Business	Total
Fully Rejected	52% (47%)	30% (30%)	32% (25%)	20% (14%)	28% (33%)	50% (42%)	38% (34%)
Partially Admitted	18% (16%)	23% (22%)	26% (31%)	7% (10%)	11% (8%)	8% (8%)	18% (18%)
Fully Admitted	30% (38%)	47% (48%)	41% (44%)	73% (76%)	61% (58%)	42% (50%)	44% (49%)

To summarize our descriptive case data: we replicated the finding that civil defendants and criminal prosecutors tend to be more successful in having expert evidence admitted than do civil plaintiffs, and criminal defendants are by far the least successful party.[122] This pattern raises questions about the source of the differences. It is not clear whether this

122. *See* Risinger, *supra* note 67.

phenomenon is due to discrepancies between parties in the type and quality of evidence being offered, some type of bias against certain parties, or a combination of both. Which explanation is correct has important implications for conclusions about whether the legal system is achieving equal treatment for all litigants. This is particularly true with regard to forensic evidence: judges frequently noted instances in which the prosecution's expert evidence contained a number of flaws, but such evidence was still overwhelmingly admitted.

Table 6 presents mean word counts for all coded variables. For the dichotomous variables, we provide the percentage of cases in which the factor was present. For each of the word count variables, we also computed an index reflecting the number of words devoted to that factor divided by the number of words devoted to all of the factors discussed by the judge in that case. So, for example, if in Case A we coded 100 words dedicated to qualifications, 50 words to implicit error rate analysis, and 50 words to relevance analysis, each of those word counts would be divided by 200 to arrive at the proportion of total analysis dedicated to each variable (50% for qualifications, 25% for implicit error rate analysis, and 25% for relevance analysis). The purpose of this was to provide a measure that weights each case evenly regardless of length, whereas analyses based on raw word counts weight cases with longer discussion of all variables more heavily.

Some of our hypotheses focus only on the relative discussion of the *Daubert* factors themselves, so we also calculated a proportion measure using only the explicit error rate, testability, peer review, general acceptance, maintenance of standards, measurement (which is essentially a subset of testability analysis), and implicit error rate variables. To calculate the proportion of each of those variables in each case, we divided the number of words dedicated to the variable by the sum total of words dedicated to all seven variables to arrive at a percentage. We term this the "*Daubert* proportion." Cases in which none of these seven factors were discussed were not included in calculating the *Daubert* proportion.[123]

Finally, we included a dichotomous measure of whether each type of analysis was conducted at all in the opinion, regardless of word count. We termed this measure "frequency of use."

123. There was no discussion of our seven *Daubert* factors for 52 of our 272 total experts; thus, analyses conducted on *Daubert* proportions include 220 experts.

TABLE 6. WORD COUNTS (AND STANDARD DEVIATIONS) AND PROPORTIONS OF ALL VARIABLES[124]

Dependent Variable	**Average Word Count**	**Proportion**	***Daubert* Proportion**	**Frequency of Use**
Testability Factor	-	-	-	69.90%
Peer Review Factor	-	-	-	69.90%
General Acceptance Factor	-	-	-	69.90%
Error Rate Factor	-	-	-	66.03%
Standards Factor	-	-	-	32.54%
Kumho Tire Citation	-	-	-	22.49%
Explicit Error Rate Analysis	47.26 (156.83)	4.08%	6.62%	19.85%[125]
Implicit Error Rate Analysis	180.01 (290.47)	24.41%	43.75%	51.47%
Testability Analysis	91.21 (193.68)	12.55%	21.95%	39.34%
Measurement Analysis	14.31 (46.84)	3.05%	5.60%	13.24%
Peer Review Analysis	32.09 (148.42)	3.63%	7.09%	28.68%
General Acceptance Analysis	38.63 (92.13)	5.85%	12.25%	32.35%
Standards Analysis	20.03 (130.62)	1.32%	2.28%	8.82%
Ipse Dixit Discussion	19.57 (95.55)	1.99%	-	8.09%
Qualifications Analysis	123.19 (211.55)	21.04%	-	54.78%
Relevancy Analysis	53.93 (157.85)	8.30%	-	22.43%
Generated for Litigation Analysis	5.24 (25.88)	0.63%	-	4.78%
403 Balancing Analysis	27.57 (76.78)	4.39%	-	17.65%

124. The values provided in this table are unweighted. We present the same data after weighting in Table 7. Parentheses here indicate standard deviations.

125. We also note that at least one of the two error rate analysis factors was used by the judge in 61.76% of our cases.

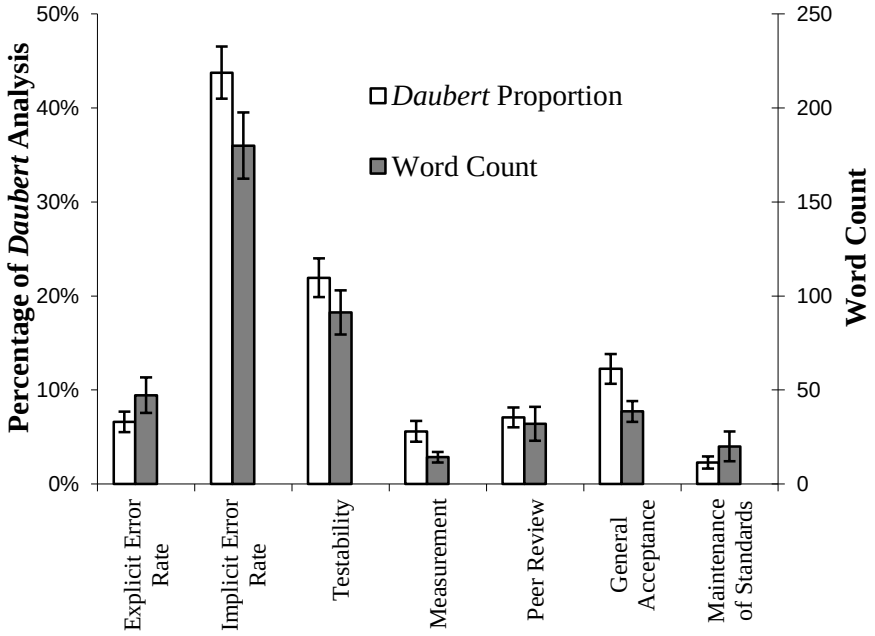
We now turn to data testing our specific hypotheses regarding the amount and type of judicial discussion in *Daubert* cases, discussing each of our hypotheses in turn.

B. Hypothesis 1: Judges Will Allocate More Discussion to Implicit Error Rate Analysis than to Other Daubert *Factors in Assessing the Evidentiary Reliability of Expert Evidence*

Figure 1 shows the word counts and *Daubert* proportions[126] of the five traditional *Daubert* factors (testability, peer review, general acceptance, error rate, and maintenance of standards) as well as our implicit error rate and measurement analysis variables. Both word counts and *Daubert* proportions show the same pattern.

126. As noted above, the *Daubert* proportion for each factor is calculated by dividing the number of words dedicated to the factor by the sum total of words dedicated to all seven factors.

FIGURE 1. WORD COUNTS AND PROPORTIONS OF THE *DAUBERT* FACTORS ACROSS ALL CASES



We conducted a 7 x 1 repeated measures ANOVA to examine the differences in *Daubert* proportion values across our seven *Daubert* factor measures. As predicted, we found a strong difference in the percentage of text devoted to each factor.[127] To assess the individual differences between the percentage of discussion across factors, we conducted Tukey post-hoc follow-up comparisons between each factor. As expected, implicit error rate was discussed significantly more than any other factor,[128] accounting for 44% of all analysis of the *Daubert* factors. Testability was the second most-discussed factor, discussed more than every other factor except implicit error rate,[129] followed by general acceptance, which also differed from all other factors.[130] Measurement, peer review, and explicit error rate analysis counts did not differ, and the maintenance of standards factor was used less frequently than all others,

127. $F(6, 1314) = 68.435$, $p < .001$, $\eta = .238$.
128. All p's $< .001$.
129. All p's $< .001$.
130. All p's $< .005$.

accounting for just 1.85% of the total *Daubert* analysis.[131] In sum, the evidence strongly supported Hypothesis 1—judges spent a great deal of time in opinions making implicit assessments of the potential error of an expert's method. We further explore the implications of this finding in Part VI.

C. Hypothesis 2: The Length of the Judge's Implicit Error Rate Discussion in an Opinion Will Predict the Outcome of the Admissibility Inquiry: When Judges Devote More Discussion to Implicit Error Rate Analysis, They Will Be More Likely to Reject All or Part of the Expert's Evidence Than to Fully Admit It

To examine the extent to which the variables we coded predicted the judge's eventual decision as to the admissibility of the expert testimony, we ran a multinomial logistic regression with admission decision (three levels: fully admitted, partially admitted, fully rejected) as the nonparametric dependent variable.[132] We included all 12 parametric variables that we coded for as independent variables in the analysis.[133] The overall model was highly significant.[134] Two of our variables significantly predicted the admissibility outcome: implicit error rate analysis[135] and qualifications analysis.[136] In cases in which an expert was fully rejected, the opinion averaged 278.4 words of implicit error rate analysis compared to 143.53 words when the expert was partially admitted and 109.05 words when the expert was fully admitted. Though the amount of qualifications analysis also predicted the admissibility outcome, the relationship was not linear among the three possible

131. All p's < .01. We also conducted the same analysis using raw word counts rather than *Daubert* proportions. The omnibus ANOVA was also highly significant, $F(6, 1626) = 36.87$, $p < .001$, $\eta = .12$. Most post-hoc comparisons yielded the same result, but there were several differences. Using raw word counts, there was a significant difference between explicit error rate analysis and measurement analysis, unlike in the *Daubert* proportion analysis. Also, when using raw word counts, there were no differences between explicit error rate and general acceptance, between measurement and maintenance of standards, or between peer review and maintenance of standards. Using the weighted version of the *Daubert* proportion, we found a significant omnibus ANOVA, $F(6, 1200) = 100.56$, $p < .001$, $\eta = .335$, but we also saw slightly different individual effects: measurement analysis was significantly greater than explicit error rate analysis and marginally smaller than peer review analysis, but did not differ from general acceptance analysis; and maintenance of standards analysis did not differ from either explicit error rate or peer review.

132. The "fully rejected" code was used as the reference category.

133. The 12 variables used in the analysis are the 12 parametric variables for which word counts can be found in Table 6.

134. $\chi^2(24) = 57.241$, $p < .001$.

135. $\chi^2(2) = 20.23$, $p < .001$.

136. $\chi^2(2) = 15.92$, $p < .001$.

outcomes: when the expert was fully rejected, the opinion averaged 86.93 words of qualifications discussion compared to 227.39 words when the expert was partially admitted and 111.97 words when the expert was fully admitted.

The significant predictive result for qualifications was unexpected and is especially odd considering the nonlinear nature of the relationship. One possible explanation for this is that in close cases where the reliability decision does not lead to a clear outcome, judges may turn to qualifications as a more critical factor in determining admissibility. This rests on several assumptions, most importantly that cases in which the evidence is partially admitted are closer cases than those in which the evidence is fully admitted or fully rejected. Another possibility is that judges may be reluctant to fully exclude experts with impressive credentials. It will take a comparison of credentials across cases to evaluate this explanation.

Additionally, when we conducted the same analysis on the weighted case data, both implicit error rate analysis and qualifications remained significant predictors of the admissibility decision, but testability analysis was also a significant predictor.[137] The pattern for testability analysis is similar to the pattern for implicit error rate analysis: in cases in which an expert was fully rejected, the opinion averaged 101.85 words of testability analysis compared to 59.91 words when the expert was partially admitted and 29.74 words when the expert was fully admitted. Because our sample was disproportionately inclusive of criminal cases and cases from the years immediately following the *Daubert* decision, the fact that this effect was only significant after weighting the data suggests that testability analysis may have become more important over time or may be more important in civil cases.

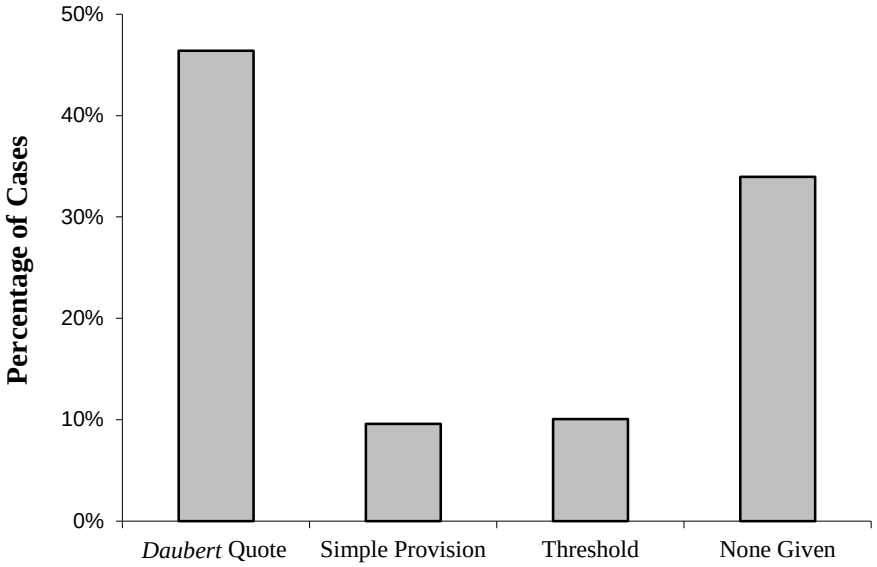
D. Hypothesis 3: Judges Will Show Confusion as to Whether the Error Rate Standard Incorporates a Threshold Requirement.

Figure 2 presents the frequency with which each of the various definitions of the error rate standard was observed across all cases. As can be seen in the Figure, the most common phrasing of the standard was a simple quote or paraphrase of *Daubert*, accounting for 46% of all cases. In 34% of cases, no error rate standard was used in the opinion. However, in 20% of cases, either the threshold standard or simple provision standard was given by the court (split evenly between the two). While the *Daubert* quote standard is ambiguous, the threshold standard and simple provision standard are in tension with one another.

137. $\chi^2(2) = 8.08$, $p = .004$.

Supporting our hypothesis, the split seen here indicates confusion in the lower courts as to what the correct standard is.

FIGURE 2. PROPORTION OF ERROR RATE STANDARDS PROVIDED ACROSS ALL CASES



To examine whether the type of error rate standard mentioned by the court was associated with the amount of explicit error rate discussion, we conducted a 4 x 1 ANOVA comparing the *Daubert* proportion of explicit error rate discussion for each of the possible error rate standards. We found that the percentage of explicit error rate discussion varied based on the standard given.[138] Unsurprisingly, when the explicit error rate standard was not mentioned in the outline of the law, there was less explicit error rate analysis than when a quote, simple provision, or threshold standard was given.[139] However, the standard given had no effect on the amount of implicit error rate discussion, even when the error rate factor was not mentioned at all in the outline of the law,[140] possibly indicating that judges do not consider their implicit error rate

138. $F(3, 166) = 3.11$, $p = .028$.

139. Comparing cases in which no standard was given with cases in which a *Daubert* quote standard was given, $p = .014$. Comparing with a simple provision standard, $p = .009$. Comparing with a threshold standard the effect was not significant, $p = .14$.

140. $F(3, 166) = 1.49$, $p = .219$.

analysis to fall under the error rate factor. We discuss this possibility further in the next Part.

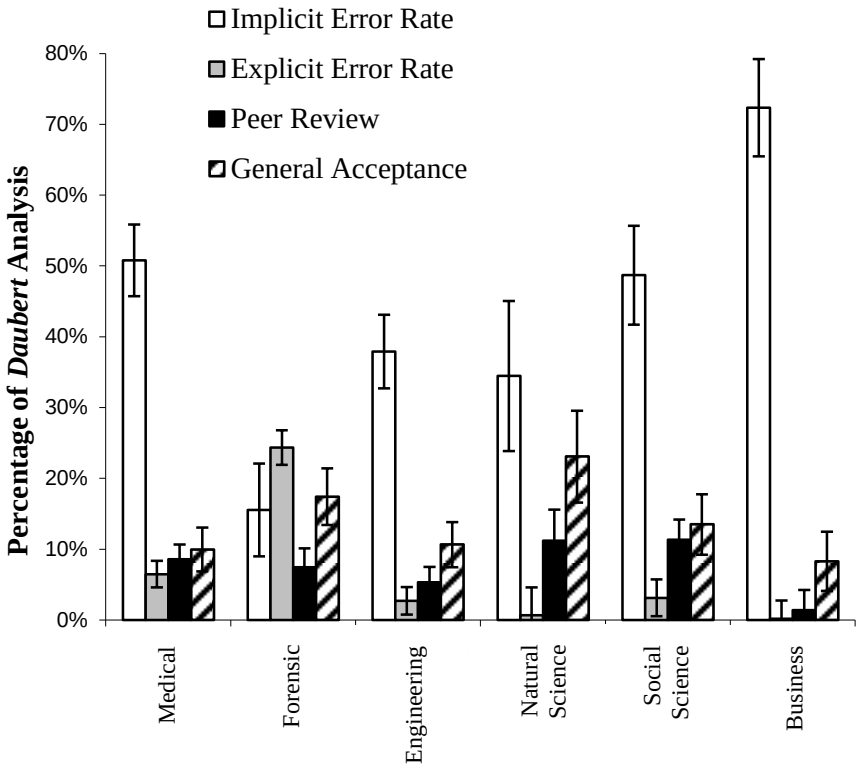
One would expect that given *Kumho Tire*'s threshold phrasing of the error rate standard, courts would begin to increasingly cite the threshold standard over time in the years following the decision. However, we did not find any evidence of an increase—dividing the sample between cases before and after January 1, 2000, revealed no increase in the mention of the threshold error rate standard given by the judge.[141] Eight of the 65 cases (12%) occurring before 2000 mentioned a threshold standard. Similarly, 13 of the 144 cases (9%) occurring after 2000 gave a threshold standard. Though this is a relatively small sample of cases, these data indicate continuing confusion regarding the error rate standard even after *Kumho Tire*, which we discuss further in the next Part.

E. Hypothesis 4: Implicit Error Rate Discussion Will Be More Prevalent When the Judge Is Assessing Social Science and Business Experts as Compared to Natural Science and Medical Experts

To examine the difference in implicit error rate discussion across the six categories of experts in our sample (medical, forensic, engineering, natural science, social science, and business), we conducted a 6 x 1 between-subjects ANOVA using the *Daubert* proportion measure (see Figure 3).

141. $\chi^2(3) = .809$, $p = .847$.

FIGURE 3. IMPLICIT ERROR RATE ANALYSIS, EXPLICIT ERROR RATE ANALYSIS, PEER REVIEW ANALYSIS, AND GENERAL ACCEPTANCE ANALYSIS BY EXPERT CATEGORY (FOR HYPOTHESES 4, 5, AND 6)



As expected, there was a main effect of expert category, indicating differences in the proportion of implicit error rate analysis across expert types.[142] We conducted a Tukey post-hoc test to determine differences between individual expert types. Notably, implicit error rate analysis was most prominent in *Daubert* analysis of business experts, where it accounted for 72% of the *Daubert* factor analysis, greater than any other expert type.[143] We suspect that this difference is due to the comparative ease with which judges can assess business methods that are most similar to their own areas of expertise. However, contrary to our expectations, social science experts did not engender more implicit error rate analysis

142. $F(5, 213) = 8.05$, $p < .001$, $\eta = .159$.

143. All p's $< .05$.

than any category other than forensic experts;[144] implicit error rate analysis was equally frequent for medical,[145] engineering, and natural science experts. Implicit error rate discussion of forensic experts was rare, accounting for just over 15% of all *Daubert* analysis and lower than any other category except for natural science.[146] This may be due to a "grandfathering" effect in which forensic evidence is rarely questioned, as we explore below.[147]

We found a somewhat different pattern of differences when weighting our cases to account for our stratified sampling. We conducted a 6 x 1 ANOVA on the weighted *Daubert* proportion of implicit error rate discussion. As with the unweighted data, we found a significant main effect of expert category.[148] As in the unweighted analysis, business experts attracted a high rate of implicit error rate discussion, but we found that business, social science, and medical experts all received a roughly equally large amount of implicit error rate discussion, with implicit error rate accounting for over 60% of discussion for all three categories.[149] Engineering, forensic, and natural science experts received significantly less error rate discussion and did not differ from one another.[150] Aside from the high amount of implicit error rate discussion for medical experts, these results are more in line with our original hypothesis than the unweighted data: judges tended to engage in more methodological analysis when the discipline was more accessible and less technical.

---

144. For the difference between social science experts and forensic experts, $p = .001$.

145. There was a marginally significant difference between medical and engineering experts, $p = .078$.

146. All p's $< .001$.

147. As in analyses above, we chose the *Daubert* proportion measure for its usefulness in reducing the effect of outlier cases. The analysis on raw word counts yielded similar results, though when using raw word counts there were no significant differences between engineering and forensic testimony, and engineering expert cases produced significantly less implicit error rate analysis than social science expert cases. All other effects were the same.

148. $F(5, 193) = 3.38$, $p = .006$, $\eta = .081$.

149. Business and social science discussion was significantly greater than engineering, natural science, and forensic discussion (all p's $< .05$), while medical implicit error rate discussion was significantly greater than forensic discussion ($p = .033$) and moderately greater than engineering discussion ($p = .07$) and natural science discussion ($p = .088$).

150. All p's $> .3$.

F. Hypothesis 5: Analysis of the "External" Factors—Peer Review and General Acceptance—Will Not Vary by the Type of Expert Evidence Presented

To examine the difference in peer review and general acceptance analysis across the six categories of experts in our sample (medical, forensic, engineering, natural science, social science, and business), we conducted two 6 x 1 between-subjects ANOVAs using the *Daubert* proportion measure (see Figure 3): one for peer review analysis and one for general acceptance analysis. Neither ANOVA yielded a significant main effect,[151] indicating that discussion of the peer review and general acceptance factors did not differ across expert categories.

G. Hypothesis 6: Explicit Error Rate Discussion Will Be More Prevalent in Forensic Testimony, Where Individuation/Identification Testimony Is Common, as Compared to Other Types of Expert Testimony

To examine the difference in explicit error rate analysis across the six categories of experts in our sample (medical, forensic, engineering, natural science, social science, and business), we conducted a 6 x 1 between-subjects ANOVA using the *Daubert* proportion measure (see Figure 3). There was a significant main effect of expert category, indicating differences in explicit error rate discussion among expert types.[152] Tukey's post-hoc tests revealed that explicit error rate analysis was more prevalent for forensic experts as compared to all other experts.[153] Additionally, medical testimony generated a greater amount of explicit error rate analysis than business testimony.[154] Thus, our hypothesis was partially confirmed.

Overall, our hypotheses were mostly confirmed regarding differences in use of the *Daubert* factors across expert disciplines: more implicit error rate discussion was devoted to business experts than to other categories, more explicit error rate discussion was devoted to forensic experts, and the general acceptance and peer review discussion

151. For the peer review factor, $F(5, 213) = 1.66$, $p = .145$, $\eta = .038$. For the general acceptance factor, $F(5, 213) = 1.25$, $p = .288$, $\eta = .028$. As in analyses above, we chose the *Daubert* proportion measure for its usefulness in reducing the effect of outlier cases. The same results were yielded conducting the analysis on raw word counts and on weighted *Daubert* proportions.

152. $F(5, 213) = 13.5$, $p < .001$, $\eta = .241$.

153. All p's $< .001$.

154. $p = .049$. Conducting the same ANOVA using raw word counts yielded the same results, except there was no difference between the amount of explicit error rate discussion for medical experts as compared to business experts. Using weighted *Daubert* proportions yielded the same result as using raw word counts.

was roughly steady across all types of experts. The balance of implicit error rate analysis across expert disciplines provides some support for the notion that judges are less comfortable with the assessment of natural sciences than they are with social sciences. The most plausible explanation for these results is that judges are simply more comfortable in business disciplines and have more expertise in those areas most related to law. However, other expert categories did not show results that are consistent with this explanation: we expected that judges would also demonstrate more comfort analyzing social science evidence as compared to medical or natural science evidence, but there were no differences between those groups in the quantity of implicit error rate analysis. To fully understand the effect, more systematic examination of the type of discussion and the nuance with which judges make the implicit error rate analysis is necessary. In this area in particular, a reliable rating scale reflecting the competency of judges' assessments would be extremely helpful: if the elevated implicit error rate discussion of business experts is caused by judges' greater expertise in business, we would also expect to see more nuanced and competent discussion in those same cases as compared to experts in other disciplines.

VI. DISCUSSION AND LEGAL RESPONSES

In this study, we report several novel results that shed light on how judges evaluate expert evidence. On the whole, we found that judges faced with a *Daubert* challenge often undertake a detailed analysis of the quality of the methodology used by the expert rather than simply relying on proxies for the quality of the method such as peer review and general acceptance. This finding is somewhat in tension with much of the current literature in the area, which tends to report that judges are either unwilling or unable to directly assess the expert's methods.[155] We characterize much of this discussion as falling under the "known or potential rate of error" factor in the *Daubert* test, though we are not convinced that the Supreme Court had this type of analysis in mind when fashioning the error rate factor. In line with this, we find a great deal of judicial confusion regarding the error rate standard, notably confusion regarding whether an expert must simply present an error rate or whether he must stay below a certain threshold rate of error in order to satisfy the factor. Both the implicit error rate analysis we describe and the more traditional explicit error rate analysis varied across expert types, which sheds some light on the nature of the two types of analysis.

Our most important finding was the substantial amount of implicit error rate analysis undertaken by judges across all expert types. We

155. *See supra* Part II.

observed a rich and diverse set of judicial analyses assessing the internal, construct, and external validity of expert methodology—from confounds in experiments to carelessness in calculations or unjustified conclusions being drawn from researcher premises. Importantly, the distinguishing feature in all of these analyses is that by assessing the quality of the methods themselves, rather than relying on proxies of good science such as peer review and general acceptance, judges implicitly assessed the likelihood that the expert would make an error in his final testimony. To get a better sense of the implicit error rate discussion, it is helpful to look at a few examples in addition to the examples we provided in Part IV. In this first example, the judge focused on facts and data that an expert left out of his analysis which would likely lead to inaccurate results. After listing the numerous factors which a valuation expert left out of his analysis, the judge concluded:

> In computing Point's lost sales after the Agreement's termination, Churchill's failure in his expert report to address any of these events and failure to attempt to capture their impact on Point's ability to be a full player in the budget market is startling. Even more astounding is that instead of factoring into his analysis these real world facts and events, Churchill systematically adopts possibly speculative assumptions and predictions that are vital to his projections . . . . Standing alone, these assumptions might not render Churchill's report and anticipated testimony unreliable. But when his willingness to rely on these sometimes questionable assumptions is considered in light of the report's gaping omissions of real world events that were highly material to Point's vitality and unrelated to Sony's termination, Churchill's testimony is left irretrievably unreliable and indefensible. It is therefore excluded.[156]

Occasionally, a judge used implicit error rate analysis to conclude that an expert's argument was logically invalid from its own premises. Here, the judge noted that the expert's statement to the jury in a tort case could not have possibly been accurate if his earlier statements regarding his methods were true:

> The most striking testimony he gave during his deposition regarding this so-called testing was that the refrigerator door, when heavily loaded, could close with sufficient force to crush

156. *Point Prods. A.G. v. Sony Music Entm't, Inc.*, No. 93 Civ. 4001(NRB), 2004 WL 345551, at *7 (S.D.N.Y. Feb. 3, 2004) (footnote omitted).

> a carrot. Most significantly, and contradictory, Leshner's own notes jotted down that day state that the door would not swing shut by itself! He also admitted that the door could only be closed by the application of sufficient external force, i.e., by pushing it closed. Despite his acknowledgment that the refrigerator door had to be pushed closed manually because it would not close by itself, Leshner would opine to the jury that, in July of 1992, the refrigerator door closed by itself on plaintiff's thumb with enough force to crush her thumb. If there is any method in this madness, the Court cannot find it.[157]

Often, judges attacked the internal validity of studies conducted by experts or cited by experts at a general level, stating that the studies lacked control or were likely to be erroneous for a number of reasons, as in this toxic tort case:

> The case reports upon which [the experts] rely make little attempt to isolate or exclude possible alternative causes, lack adequate controls, and lack any real analysis. Granted, an overwhelming amount of case reports of a temporal proximity between a very specific drug and a very specific adverse event might, as [the opposing expert] admits, be enough to make a general causation conclusion sufficiently reliable. In this case, however, we have a scant number of case reports indicating that Parlodel is temporally associated with all types of adverse events. There is not the volume of or specificity within these case reports to reliably show that [the plaintiff's drug caused the defendant's injuries].[158]

Sometimes the implicit error rate analysis was derived from the *ipse dixit* rule of *Joiner*, with the judge opining that the expert's conclusion could be inaccurate because his testimony could not be justified by his data:

> Does an ability to appreciate wrongfulness only at the level of a child between 8 and 12 years of age make one insane? The court has found no authority for such a sweeping generalization. Courts have long allowed children as young as six years old to testify because "there is no precise age which determines the question of competency. This depends on the

157. *Belofsky v. Gen. Elec. Co.*, 980 F. Supp. 818, 823 (D.V.I. 1997) (citation and footnote omitted).

158. *Caraker v. Sandoz Pharm. Corp.*, 172 F. Supp. 2d 1046, 1050 (S.D. Ill. 2001) (citation omitted).

> capacity and intelligence of the child, his appreciation of the difference between truth and falsehood, as well as of his duty to tell the former."
>
> . . . .
>
> The analytical gap between tests which show "low normal" functioning and an immature thought process on one hand and a conclusion of insanity on the other is just too great. The gap between the evidence concerning Klinefelter Syndrome and a diagnosis of insanity is even greater. This factor weighs heavily against admission of the testimony.[159]

While we characterize implicit error rate analysis under the "known or potential error rate" factor of *Daubert*, we note that some of our data call into question whether the trial judges consider their discussion on this topic an error rate analysis—when judges do not mention the error rate factor in their description of the law, they are less likely to conduct extensive explicit error rate analysis, but equally likely to conduct implicit error rate analysis. This is not surprising in our view, given the absence of much explanation of the factor in *Daubert* and *Kumho Tire*. We suspect that judges are not likely to take the broad interpretation of the factor as we described in Part I of this Article; they are more likely to consider their error rate analysis as cabined by (or limited to) situations in which quantitative error rates are discussed. However, this does not mean that judges are not interested in error rates; it simply means that they do not characterize their error rate analysis under the framework laid out by *Daubert*. Our chief aim in presenting these data is to demonstrate that judges are actually quite interested in the likelihood of an expert's error due to methodological weakness, though their discussion of it is not framed in terms of the language of the traditional factor.

We do not, based on this analysis, conclude that judges are methodologically sophisticated in their discussions of error rate, known or potential. The fact, however, that they are sufficiently motivated to engage in such discussion is important. It suggests that a methodologically informed judiciary can be depended upon to play the gatekeeping role that *Daubert* and the FRE require them to play in reaching admissibility decisions involving experts. The follow up, of course, is to assure that that judges are in fact methodologically informed and competent to do this work.

This stance is bolstered by the apparent importance of implicit error rate analysis in the final admissibility decision: we found that the length

159. *United States v. Eff*, 461 F. Supp. 2d 529, 535 (E.D. Tex. 2006) (quoting *Beausoliel v. United States*, 107 F.2d 292, 293 (D.C. Cir. 1939)).

of implicit error rate analysis predicted the judge's admissibility decision better than any other factor.[160] The association between implicit error rate analysis and admissibility took the expected form, with the most implicit error rate analysis present when the evidence was ultimately rejected, a moderate amount present when the evidence was partially admitted, and the least amount present when the evidence was fully admitted. Though it is tempting to conclude that this predictive value shows that implicit error rate analysis is more important to judges in assessing scientific evidence, we cannot be sure that it is the only reason for this relationship. It is possible that other factors are equally important or more important in the admissibility decision of the judge but their level of importance remains high regardless of length.

Yet as Table 6 indicated, the judges engaged in at least some implicit error rate analysis for over half of the experts (51.4%), but engaged in general acceptance analysis for only 32.3% of the experts and in peer review analysis for only 28.7% of them.[161] While we cannot make strong comparative statements about the importance of the various factors based on this measure, we would expect to find no relationship between the length of the discussion and the decision outcome if the implicit error rate analysis played no role in the final admissibility decision.

Significantly, the prominence of implicit error rate analyses revealed in our research contrasts sharply with reports from prior studies that judges are relatively uninterested in error rates[162] and unable to use or understand the error rate factor[163] or make clear assessments of scientific validity.[164] Past research has concluded that "judges simply lack understanding of [the *Daubert*] criteria and of scientific reliability in

160. *See supra* notes 132–36 and accompanying text.

161. *See supra* Table 6.

162. *See supra* Part III.

163. *See supra* notes 49–55 and accompanying text.

164. *See, e.g.*, Margaret Bull Kovera, Melissa B. Russano & Bradley D. McAuliff, *Assessment of the Commonsense Psychology Underlying* Daubert*: Legal Decision Makers' Abilities to Evaluate Expert Evidence in Hostile Work Environment Cases*, 8 PSYCHOL. PUB. POL'Y & L. 180 (2002) (arguing that "judges, attorneys, and jurors are not particularly skilled in identifying flawed research"); Lora Levett & Margaret Bull Kovera, *The Effectiveness of Opposing Expert Witnesses for Educating Jurors About Unreliable Expert Evidence*, 32 LAW & HUM. BEHAV. 363, 363–65 (2008) (finding that opposing expert testimony merely caused mock jurors to be skeptical of all expert testimony rather than sensitizing them to flaws in such testimony); Bradley D. McAuliff, Margaret Bull Kovera & Gabriel Nunez, *Can Jurors Recognize Missing Control Groups, Confounds, and Experimenter Bias in Psychological Science?*, 33 LAW & HUM. BEHAV. 247, 248 (2009) (demonstrating that laypeople have difficulty recognizing confounds in psychological evidence).

general to apply them to their admission decision making."[165] We can sympathize with this conclusion at least in part because we understand that some would not characterize our implicit error rate analysis as falling under the *Daubert* criteria. *Daubert* did not suggest that the five listed factors were exhaustive, so another way to view what we have labeled as implicit error rate might be the larger methodological evaluation that *Daubert* called on judges to perform. If so, it appears that federal judges generally have not been treating the *Daubert* factors as an exhaustive checklist. Our analysis reveals that many judges go much further, adopting the spirit as well as the letter of *Daubert*.

Another novel finding we report here is the extent of confusion regarding the nature of the error rate factor: the judges in a full 20% of the cases in our sample characterized the factor as either a threshold test or a simple provision test, and of those 20%, judges were nearly evenly split on either side. Even after the *Kumho Tire* Court phrased the error rate factor as a threshold standard, confusion has remained essentially the same at the trial court level. The fact that in both *Daubert* and *Kumho Tire* the error rate factor is given essentially a single sentence of direct discussion likely contributes to the inconsistency in the lower courts;[166] the Supreme Court to date has not felt it necessary to further define the standard, and so lower courts have been left to guess. In that regard, the confusion among the district courts should not be surprising. Although we know that the implicit error rate discussion was not longer when a judge cited the threshold standard, an interesting question for future analysis is whether the nature of the error rate analysis differs depending on which error rate standard is cited by the court. One might expect generally more critical error rate analysis coming from a court citing a threshold standard as compared to one citing a simple provision standard; logically, for a judge to conduct a threshold analysis, the expert must have presented an error rate, which presumably should meet the simple provision standard. Future study of the nature of this discussion, preferably with a larger sample size of cases citing the two conflicting standards, would be valuable.

There are several drawbacks to the word-counting method we employ that deserve discussion. We were able to control for some, but not all of them. First, the number of words spent discussing a particular topic in an opinion does not necessary reflect the importance of that topic; a number of factors could lead to implicit error rate discussion being lengthier than other *Daubert* factor discussions without it being more important in the admissibility calculus. One concern may simply be variability in the overall length of the opinion, which could overweight

165. Groscup et al., *supra* note 56, at 367.

166. *See supra* Part I.

lengthier opinions—a concern that was also noted in the Groscup et al. (2002) word-counting study.[167] We controlled for this problem by conducting our analyses using proportion measures, which control for opinion length. Importantly, however, one reason that implicit error rate discussion may be more extensive than discussion of the other factors is that it may be more *complex* than other types of analysis. In order to properly conduct an implicit error rate analysis, the judge must fully understand the expert's methods and evaluate them for flaws, which may require significant explanation, as seen in the above examples of implicit error rate analysis. In contrast, discussion about peer review or general acceptance may be simpler, especially in the case of peer review, where the judge may be able to simply state whether or not the research has undergone the publication process:

> Where proffered expert testimony is not based on independent research, the party must come forward with other objective, verifiable evidence that the testimony is based on "scientifically, valid principles," *e.g.*, peer review and publication. Here, however, [the expert] concedes he has not published any article about the valuation of trademarks. Thus, his opinions and analysis regarding trademark valuation have not been subjected to the rigors of peer review.[168]

The fact that this analysis is shorter may not necessarily mean that it is less important; the length of the implicit error rate analysis may simply stem from necessity. However, the fact that implicit error rate discussion tends to appear in more opinions than do peer review and general acceptance discussion shows its importance, regardless of length.

We do suspect, moreover, that word counts are a good proxy for the relative importance of the various reliability considerations, especially with respect to the implicit error rate analysis. We found that our word count analysis was highly predictive of the admissibility outcome of an expert.[169] If the length of analysis of the various reliability concerns was entirely unrelated to the judge's consideration of the relative importance of the factors, one would expect no relationship between word counts and admissibility outcome. Moreover, our implicit error rate analysis measure was more predictive of admissibility outcome than any other factor; when the amount of implicit error rate analysis increased, the

---

167. Groscup et al., *supra* note 56, at 370 ("An additional confound on this measure includes the writing style of the individual judges, which could be quite brief or verbose.").

168. *United Phosphorus, Ltd. v. Midland Fumigant, Inc.*, 173 F.R.D. 675, 686 (D. Kan. 1997) (citation omitted).

169. *See supra* note 134 and accompanying text.

expert was much more likely to be rejected. This strongly suggests that the amount of implicit error rate reflects the extent of implicit error rate problems with the evidence, which makes word counts a valuable measure.

A second limitation to our method is that simply because judges spend words in their opinions conducting implicit error rate analysis does not necessarily mean that they can analyze the methodology *competently*. We certainly did come across examples in our coding in which judges conducted an analysis that led them to an incorrect conclusion or the judge misunderstood the nature of the analysis. For example, in this forensic case, a judge assessed the error rate of a DNA testing method, but made the assumption that the lab conducting the analysis applied the methodology perfectly, ignoring the possibility of individual lab error (arising, for example, from a mislabeled DNA sample):

> The FBI protocol for performing PCR/STR analysis has been designed to eliminate any potential technological errors and establish an acceptable range of measurement error. The FBI methodology has been developed to result in a zero error rate within acceptable measurement error conditions (error being understood as yielding an incorrect result), if the methodology is followed and properly calibrated instruments are used.[170]

While we did not code for the correctness or competence of a judge's analysis, we do note that examples like this were the exception rather than the rule. Nevertheless, some past literature has documented how nonscientists may struggle with tasks that are central to the implicit error rate analysis: recognizing confounds and faulty conclusions in science.[171] Additionally, the Gatowski et al. (2001) survey data imply that judges have particular difficulty with the error rate factor, as less than 5% of all judges in that study demonstrated a "clear understanding" of the factor.[172] One thing that these studies do not account for, however, is the fact that judges in actual cases have input from the adversarial system to assist them as they make admissibility determinations. While it may very well be difficult for a judge to recognize the problems with an expert's methods based on a blank slate, judges have resources to aid them, most notably the parties' briefs on the *Daubert* motion, which can help bring the relevant competing arguments to the forefront. Federal judges in particular have the benefit of several clerks to do additional

170. *United States v. Trala*, 162 F. Supp. 2d 336, 347 (D. Del. 2001) (citation omitted).

171. *See* sources cited *supra* note 164.

172. *See supra* note 49 and accompanying text.

research, which may mean that judges have better performance in assessing the validity of science in actual cases than in experiments and surveys.

Of course, despite these safeguards, errors are likely to be made even by judges at the highest levels who possess the greatest resources with which to aid them in their analysis. One example of such an error occurred in *Exxon Shipping Co. v. Baker*,[173] a 2008 Supreme Court case arising from a supertanker oil spill.[174] In part, the case called on the Court to determine whether the proper balance had been struck between compensatory damages and punitive damages.[175] In reducing a $2.5 billion punitive damages award to $500 million where the compensatory damages were $500 million, the Court relied in part on an empirical analysis by Eisenberg and colleagues.[176] Based on that study, the Court noted that there was little evidence to support the notion that "punitive damages [have] mass-produced runaway awards,"[177]—a conclusion clearly demonstrated in the paper—but the Court also asserted that "[t]he real problem, it seems, is the stark unpredictability of punitive awards."[178] This latter conclusion, however, resulted from a relatively unsophisticated examination of the data—it relied only on the mean and standard deviation of the data set *as a whole* to demonstrate the point that "the spread is great, and the outlier cases subject defendants to punitive damages that dwarf the corresponding compensatories."[179] But the Court missed an important point of the paper: the mean and standard deviation of punitive damages varied greatly depending on the size of the compensatory award—the standard deviation dramatically decreased for cases with compensatory awards greater than $10,000.[180] As Eisenberg and his coauthors later described, "[l]umping [low- and high-value] cases together to make policy or doctrine based on a single mean or a single standard deviation is . . . statistically questionable."[181] Ironically, the Court's misinterpretation of the data led it away from its own initial

173. 554 U.S. 471 (2008).

174. *Id.*

175. *Id.* at 489–90.

176. *Id.* at 497–500 (citing Theodore Eisenberg et al., *Juries, Judges, and Punitive Damages: Empirical Analyses Using the Civil Justice Survey of State Courts 1992, 1996, and 2001 Data*, 3 J. EMPIRICAL LEGAL STUD. 263, 278 (2006)).

177. *Id.* at 497 (alteration in original).

178. *Id.* at 499.

179. *Id.* at 499–500.

180. Theodore Eisenberg, Michael Heise & Martin T. Wells, *Variability in Punitive Damages: Empirically Assessing* Exxon Shipping Co. v. Baker, 166 J. INSTITUTIONAL & THEORETICAL ECON. 5, 18 (2010).

181. *Id.* at 20.

intuition about the existence and appropriateness of higher levels of variability in low compensatory award cases.[182]

Though this case is not an example of a judge assessing the quality of an *expert witness* at trial, it demonstrates our point that even the most talented and diligent of judges may run into problems when assessing scientific or technical evidence outside of their area of expertise, and even the best of intentions may not always lead to the best of results.[183] While we argue that we demonstrate here judges' good intentions in applying *Daubert*, we cannot yet say anything about their results. Clearly, this is an area to focus on in the future.

Lastly, we note that the sample of cases we use here, comprised entirely of federal district cases, limits the conclusions we can make. Other studies have focused instead on appellate opinions, which allow for different conclusions.[184] We reasoned that if the goal is to understand how judges are most likely to apply the *Daubert* factors, it makes the most sense to study the trial judges who will be conducting the analysis on a day-to-day basis. We did not include any state-court cases in our sample, in part because of the difficulty of controlling for the various differential standards in state law. Such state-to-state differences may provide a useful background for studying how differences in the law impart differences in the analysis, if at all.

We conclude with a few brief responses to the legal system that we would consider positive developments in light of our findings. We find that in some ways, trial courts are not conducting the analysis that *Daubert* has instructed them to, but in other ways they are actually conducting a much more wide-ranging analysis. As we have stated above, the critical message of *Daubert* is that admissibility of scientific evidence should be based on the validity of such evidence, which can be ascertained in a variety of ways. We argue that what we have here termed implicit error rate analysis is the most direct way to assess validity: by examining the expert's methods for confounds, flaws, and mistakes in reasoning that are likely to lead to errors. Based on our findings, federal trial judges seem to agree—they conduct more implicit error rate analysis than any other type of analysis. Ideally, however, this instruction should be clearly defined in the law; while we have characterized the analysis under the "potential" half of the "known or potential error rate" factor, we are not convinced that the *Daubert* Court

182. *Id.* at 21.

183. *See also* Paul S. Miller, Bert W. Rein & Edwin O. Bailey, Daubert *and the Need for Judicial Scientific Literacy*, 77 JUDICATURE 254 (1994); Michael I. Myerson & William Myerson, *Significant Statistics: The Unwitting Policy Making of Mathematically Ignorant Judges*, 37 PEPP. L. REV. 771, 774 (2010).

184. *See supra* note 72 and accompanying text.

intended this broad interpretation of the error rate factors, as argued above.[185] Of course, that does not mean that the Court thought that implicit error rate analysis should not be a part of the reliability framework, but simply rather that it was not clearly described in *Daubert*. While this ambiguity has not stopped most trial judges from conducting implicit error rate analysis, ideally the law would clarify that the analysis is an important part in assessing scientific reliability. That clarity could come either from the Supreme Court or from the FRE.

Second, the lack of clarity regarding whether the error rate factor is intended as a threshold standard or a simple provision standard needs further attention. While the Supreme Court may feel that the threshold standard was clearly delineated in *Kumho Tire*, the cases do not reflect success in changing understanding of the factor.[186] This is a fundamental and critical distinction: differences in interpretation of the standard could easily lead to opposite outcomes in the same case—experts with high but well-specified error rates might pass muster by definition under the simple provision standard, while their high error rates would likely be seen as cutting against admissibility under the threshold standard. In addition to these two possible interpretations of the standard, we also consider a third possibility: the *Daubert* Court may have left the standard ambiguous because it preferred to leave trial judges to decide whether a threshold standard or a simple provision standard is appropriate on a case-by-case basis. This would certainly not be an unreasonable stance—we can think of some contexts in which the concern of unfair prejudice is low, and probative expert evidence would be useful regardless of its error rate so long as there is an error rate for the jury to assess, while in other contexts the trial judge might want to assess evidence based on the magnitude of the error rate. However, we do not think that flexibility is what the Supreme Court intended, given its characterization in *Kumho Tire*. If the Court did intend such a flexible test, it would be helpful to make it clear with more exposition than a single sentence on the factor.

On the whole, our findings fit well with Justice Blackmun's statement of confidence in *Daubert* that "federal judges possess the capacity to undertake" the *Daubert* requirement of "assess[ing] whether the reasoning or methodology underlying the [expert] testimony is scientifically valid." [187] Our data indicate that federal judges take the validity assessment seriously, spending more words in their opinions directly assessing validity of the evidence (though implicit error rate analysis) than they do assessing external factors like peer review or

185. *See supra* notes 22−27 and accompanying text.

186. *See supra* notes 138–41 and accompanying text.

187. *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579, 592–93 (1993).

general acceptance. That is, they engage substantially in central processing in making methodological evaluations rather than merely relying on the peripheral cues of peer review and general acceptance. While their analyses may not always fit within the exact terms of the *Daubert* factors, our findings indicate that they take the spirit of *Daubert* to heart.

APPENDIX: FULL EXPLANATION OF THE THREE VALIDITY THREATS DISCUSSED IN IMPLICIT ERROR RATE ANALYSIS

Our implicit error rate analysis variable was primarily composed of analysis of an expert's methods in terms of its construct validity, external validity, and internal validity. In this Appendix, we describe in detail how we define these terms and how we identified them in our case sample for coding.

I. CONSTRUCT VALIDITY

Judges frequently assessed construct validity without explicitly saying so through objections that the expert's conclusions cannot be drawn from the data they gathered—they made the statement that the variable measured does not reflect or fully capture what it purported to measure. Thus, sometimes the *ipse dixit* problem may be a criticism of construct validity (e.g., a judge might say, "The expert makes too great of an analytical leap in stating that his survey of individual preferences actually captures the true preference of those individuals").[188] In some cases, judges simply made a statement that the method of measurement could not be relied upon to produce a trustworthy measure, as in this example:

> [The expert] states that his opinion regarding the turkey fryer's stability is based on his analysis of its design. His testimony discloses that he did not have the information an expert requires to calculate the fryer's resistance to tipping over. Edmondson testified that he did not use precise values, but elected instead to estimate the turkey fryer pot's volume and diameter. Volume and diameter values are necessary to determine the center of gravity of the fryer, which affects the stability of the fryer.[189]

In the passage, the judge stated that while the expert's estimates purported to correspond to actual volume and diameter measurements, because they were estimates, they did not accurately capture the volume

188. *See, e.g.*, *Gen. Elec. Co. v. Joiner*, 522 U.S. 136, 146 (1997) ("But nothing in either *Daubert* or the FRE 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. *Cochran v. Brinkman Corp.*, No. 1:08-cv-1790-WSD, 2009 WL 4823858, at *9 (N.D. Ga. Dec. 9, 2009) (citation omitted).

and diameter of the turkey fryer.[190] Thus, the judge was concerned that the expert's opinion was not likely to be accurate because the methods lacked construct validity, and we coded this as an implicit error rate analysis.

In addition to a lack of precision in operationalizing or measuring variables, sampling bias or other sampling problems are construct validity issues that may bring about an implicit error rate analysis. For example, in the following passage from a trademark case, an expert attempted to extrapolate data regarding consumer perception to make a claim that a particular set of artists were well known by the general public:

> [The expert] does not explain the significance of Media Guide data, how it is compiled, what it reflects, and/or whether it is typically (or ever) utilized as a proxy for consumer perception. He does not explain why he limited his analysis to the 2007–2011 time frame. Moreover, it appears that he relies on an incomplete data set even for that time frame.[191]

Here, the judge opines that the expert arbitrarily selected the time frame, which may lead to potential bias, and thus, an increased likelihood of an error from the expert.[192] Thus, this is coded as an implicit error rate analysis.

Finally, confounds in a study that make it impossible to determine which variable caused a result also fall under construct validity, as well as treatment artifacts such as experimenter bias, demand characteristics, or order effects. Many of judges' methodological criticisms questioning the validity of a methodology were related to experimenter bias—ways that the expert designed the methodology that bias the result in favor of what the expert would like to find. For example, in the following passage, a judge expressed skepticism of a forensic ballistics expert:

> In addition, the standards employed by examiners invite subjectivity. "The AFTE theory of toolmark comparison permits an examiner to conclude that two bullets or two cartridges are of common origin, that is, were fired from the same gun, when the microscopic surface contours of their toolmarks are in sufficient agreement. . . ." [B]allistic comparisons "involve subjective qualitative judgments by

190. *Id.*

191. *Moore v. Weinstein Co.*, No. 3:09–CV–00166, 2012 WL 1884758, at *6 (M.D. Tenn. May 23, 2012).

192. *See id.*

> examiners and that the accuracy of examiners' assessments is highly dependent on their skill and training. . . . [There is a] lack of a precisely defined process."[193]

We coded such discussion as an implicit error rate analysis. The judge's reasoning was that because the method is biased, and thus likely to lead to conclusions that the expert favors even when those conclusions are incorrect, the method is likely to lead to error and thus not valid.

II. EXTERNAL VALIDITY

A very common type of implicit error rate analysis was a critique of a method's external validity. We defined an external validity threat as a threat from the method's generalizability outside of the unique setting of, or beyond the subjects included in, the study itself. Thus, as with construct validity, some *ipse dixit* analysis may also be an issue of external validity—that is, whether the expert can "bridge the gap" between the mere existence of his principles in theory and his invocation of them on the specific facts of the case. The classic external validity analysis in a *Daubert* case is one that challenges the use of animal research to draw conclusions about humans:

> First, the Court follows numerous other decisions by holding that Chinese animal studies are inadmissible due to the uncertainties in extrapolating from effects on mice and rats to humans. The Chinese animal studies are short term, high-toxicity studies of effects on animals that took place outside the United States government's regulatory supervision. First, the nature of the Chinese animal studies requires extrapolation from animals to humans, from high doses to low doses, and from short to long-term exposures. Difficulties in such extrapolation has led to controversy concerning the admissibility of such studies.[194]

Such a statement stems at least in part from a question of population validity—whether the results based on the sampling population can be generalized to the larger population of interest (though the dose and exposure issues raise other external validity questions). Other types of

193. *United States v. Sebbern*, No. 10-CR-87-SLT, 2012 WL 5989813, at *5 (E.D.N.Y. Nov. 30, 2012) (citing *United States v. Otero*, 849 F. Supp. 2d 425, 431–32 (D.N.J. 2012); COMM. ON IDENTIFYING THE NEEDS OF THE FORENSIC SCI. CMTY., STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES 153, 155 (2009)).

194. *Metabolife Int'l., Inc. v. Wornick*, 72 F. Supp. 2d 1160, 1169 (S.D. Cal. 1999) (citations omitted).

population validity concerns (such as a worry that an experiment performed on college students could not be externalized to the general population) were also considered implicit error rate analyses, as they assessed the expert's statements for the likelihood that they would produce error.

Similar implicit error rate analyses may also be made based on external validity statements that instead draw from ecological validity concerns—concerns that a method will not generalize in a different setting. For example, in this passage from a product liability case, a judge made an implicit error rate statement by saying that because an expert analyzed a video of a vehicle crash different than the plaintiff's crash, he could not extrapolate his conclusions to the plaintiff:

> [The expert] depicts an accident that differs in several respects from the collision in which the plaintiff was involved. In particular, the videotape shows a van striking a fixed, immovable barrier. Plaintiff's accident involved a collision with the rear of a pick-up truck that was in operation on a highway and was neither fixed nor immovable. In the crash depicted in the videotape, the van hit the barrier at 31 miles per hour. Kelsey is not certain how fast the plaintiff's van was traveling at the point of impact but calculates its speed at between 20 and 25 miles per hour. Third, the angle of impact in the crash test differed from the angle at which plaintiff's van struck the pick-up truck. Fourth, as plaintiff's van struck the pick-up truck there was some degree of underride as the nose of the van went under the rear of the pick-up truck; the crash test videotape did not depict any underride.[195]

Similarly, external validity concerns came into play when a judge argued that the sample of a study was or was not relevant to the sample at hand in the case: "First, Microsoft's assertion that Dr. Sukumar used a non-representative sample does not appear well-founded. Microsoft suggests that the relevant universe is Xbox owners, users, or individuals likely to purchase an Xbox. Dr. Sukumar's survey only surveyed Xbox owners."[196]

Other external validity concerns, such as temporal validity issues, were coded similarly. For example, if, in a trademark case, one of the parties' experts testified regarding a survey five years prior to the case to test whether consumers perceived the product of a brand name, despite

195. *Pillow v. Gen. Motors Corp.*, 184 F.R.D. 304, 307 (E.D. Mo. 1998).

196. *Microsoft Corp. v. Motorola, Inc.*, 904 F. Supp. 2d 1109, 1120 (W.D. Wash. 2012).

the fact that the market had changed considerably in the interim, a judge might consider the study less likely to lead to an accurate conclusion because of poor temporal validity. All of these statements have in common a worry that the expert's testimony on the stand is likely to be inaccurate because his methods are not valid, and thus we code them as implicit error rate analyses.

III. INTERNAL VALIDITY

Judges critiqued the internal validity of an expert's methods in a number of ways. Internal validity is scientifically defined as the extent to which research can determine that changes in an independent variable as operationalized caused changes in a dependent variable as operationalized.[197] That is, internal validity is the extent to which a methodology can accurately determine cause-effect relationships. Analysis of a methodology's internal validity by definition critiques the ability of the study to come to an accurate conclusion about causation, so we considered internal validity discussion a part of implicit error rate analysis.

A common way that judges implicitly assessed internal validity was by discussing general incoherence of an expert's methods or lack of thoroughness, which may increase the potential error rate. Sometimes, judges opined that methods simply were not scientific—they did not appear well reasoned or thorough. While this is not one of the traditional threats to experimental internal validity, we still considered it an internal validity issue because it is an assessment of whether the methods of the study are likely to make accurate cause-effect conclusions. For example, in the following passage, the judge in a toxic tort case made a general statement that the expert did not design his methods carefully:

> [I]t is clear that [the expert] did not follow the accepted toxicology methodology in formulating his opinion of causation in this case. At bottom, his opinion is founded primarily on the temporal connection between the spill and the development of [plaintiff's] symptoms, as well as on his subjective, unverified, belief that [defendant's product] can cause the types of injuries from which [plaintiff] suffers. This is not the method of science.[198]

197. THOMAS D. COOK & DONALD T. CAMPBELL, QUASI-EXPERIMENTATION: DESIGN & ANALYSIS ISSUES FOR FIELD SETTINGS 39 (1979).

198. *Cavallo v. Star Enter.*, 892 F. Supp. 756, 773 (E.D. Va. 1995) (footnote omitted).

We considered analysis like this to be part of the internal validity of the study—the expert's conclusions about causation may not be validly drawn from his data. Additionally, we considered this to be an implicit error rate analysis because the natural conclusion of the analysis is that the expert's opinions were not valid, and thus less likely to be accurate, because of his simplified methodology.

We also encountered other more traditional types of internal validity threats. For example, a history threat to internal validity may come into play when an expert is unable to rule out competing potential causes of the outcome of interest. Other internal validity threats may be discussed also, such as selection bias, which may come up where a social scientist testifies regarding differences between populations and there is potential that the two populations sampled are different in some way other than the way claimed by the expert.

TABLE 7. WEIGHTED WORD COUNTS (AND STANDARD DEVIATIONS) AND PROPORTIONS OF ALL VARIABLES

Dependent Variable	**Average Word Count**	**Proportion**	***Daubert* Proportion**	**Frequency of Use**
Testability Factor	-	-	-	65.7%
Peer Review Factor	-	-	-	65.5%
General Acceptance Factor	-	-	-	64.2%
Error Rate Factor	-	-	-	62.8%
Standards Factor	-	-	-	31.0%
Kumho Tire Citation	-	-	-	27.3%
Explicit Error Rate Analysis	27.24 (116.63)	2.36%	3.83%	12.5%
Implicit Error Rate Analysis	208.74 (311.80)	27.73%	52.29%	56.5%
Testability Analysis	70.23 (168.85)	9.61%	18.71%	31.2%
Measurement Analysis	15.63 (51.10)	3.87%	6.67%	12.9%
Peer Review Analysis	18.88 (96.13)	2.18%	4.43%	18.6%
General Acceptance Analysis	27.41 (75.17)	5.30%	11.08%	23.4%
Standards Analysis	20.13 (162.19)	1.10%	2.14%	6.7%
Ipse Dixit Discussion	22.04 (94.56)	2.23%	-	10.3%
Qualifications Analysis	147.42 (223.97)	23.19%	-	58.8%
Relevancy Analysis	43.32 (135.03)	7.48%	-	18.9%
Generated for Litigation Analysis	3.37 (21.80)	0.47%	-	3.3%
403 Balancing Analysis	20.83 (62.37)	3.68%	-	15.70%