

## The Misquantification of Probative Value

D. H. Kaye<sup>1</sup> and Jonathan J. Koehler<sup>2,3</sup>

---

*D. Davis and W. C. Follette (2002) purport to show that when “the base rate” for a crime is low, the probative value of “characteristics known to be strongly associated with the crime . . . will be virtually nil.” Their analysis rests on the choice of an arbitrary and inapposite measure of the probative value of evidence. When a more suitable metric is used (e.g., a likelihood ratio), it becomes clear that evidence they would dismiss as devoid of probative value is relevant and diagnostic.*

---

**KEY WORDS:** evidence; inference; probative value.

A man and a woman were found in a ditch by their snowmobile. The woman was lying face down in the water. The man was sitting face up, but apparently not breathing. CPR was applied to both. Only the man survived. The state charged him with murder. The prosecution’s theory was that he took the opportunity to kill his wife to recover as the beneficiary of a large insurance policy he had purchased that year. To support this charge, the state introduced evidence that the defendant had repeatedly been unfaithful to his wife (Davis & Follette, 2002; henceforth identified as “DF, 2002”).

The defendant hired two psychologists, Deborah Davis and William Follette, to evaluate the prosecution’s suggestion that the man’s infidelity made it more likely that he was a murderer. On the basis of some arithmetic involving conditional probabilities, they concluded that “[t]he fact of infidelity is not probative of whether a man murdered or will murder his wife” (DF, 2002, p. 139). In an article in this journal, they also conclude that the common belief that a history of wife-beating is probative of murder is “clearly false” (DF, 2002, p. 143). On the basis of their arithmetic, conjectures about the prejudicial impact knowledge of infidelity would have on jurors’ judgments, and statements of varying accuracy about the law of evidence, they call for the exclusion of broad categories of evidence that they label “intuitive profiling.”

No doubt, there are some characteristics that are thought to be probative of the propositions they are offered to prove that have little diagnostic value. Consider, for

<sup>1</sup>College of Law and Center for the Study of Law, Science and Technology, Arizona State University, Tempe, Arizona.

<sup>2</sup>McCombs School of Business and School Law, The University of Texas at Austin, Austin, Texas.

<sup>3</sup>To whom correspondence should be addressed at McCombs School of Business, University Station, B6500, Austin, Texas 78712; e-mail: koehler@mail.utexas.edu.

example, the probativity of the presence of nightmares in children in cases involving an allegation of sexual abuse of a child. Although one might expect that victims of abuse would be particularly likely to suffer from nightmares, it has been reported that nightmares occur at least as frequently among nonabused children as among sexually abused children (Lyon & Koehler, 1996). If so, testimony that an alleged victim of abuse has experienced nightmares that are “consistent with” abuse is irrelevant. Other putative symptoms may be more common among abused children than nonabused ones. Although these symptoms possess probative value, jurors may not assign the proper weight to this evidence. Data on the discriminating power of such evidence might well be admissible to assist the jury. Or, evidence of symptoms that are only mildly probative might be excluded entirely if the effort at educating the jury is not worth the time and resources it would consume. Thus, it is entirely appropriate to ask both whether proffered evidence has probative value and, if it does, to consider whether jurors are likely to assign the appropriate weight to that evidence.

Unfortunately, the methods proposed by DF (2002) to answer these questions fail. Regarding normative determinations of evidentiary value, the fatal flaw lies in an idiosyncratic definition of “probative value.” As we shall see in our discussion of weight versus sufficiency, this definition confuses evidentiary support with *sufficiency* of that support. This definitional confusion has important consequences for the evaluation of evidence. It implies that valuable evidence should be ignored when the event to be proved is either very probable or very improbable. It denies jurors information that, in some cases, would be the difference between a conviction and an acquittal. It also implies that the *order* in which evidence is proffered at trial affects its probative merit. These are undesirable features in a normative model of evidence evaluation. These features are not shared by the more conventional measures of probative value.

Regarding a descriptive account of the weight that jurors assign to evidence, DF (2002) offer little to substantiate their assertion that jurors attach more weight to intuitive profiling evidence than they should. Although DF (2002) recommend that “probable prejudicial impact of the evidence would be assessed through mock jury research” (p. 154), they assure readers that such evidence “is certain to have prejudicial impact” in most cases (p. 152) and “was sure to be prejudicial” in the snowmobile case (p. 153). Perhaps the authors are content to make such sweeping statements because their new definition of probative value finds that intuitive profiling evidence has zero value. Therefore, one might argue, *any* weight that jurors place on such evidence is unwarranted. The problem, of course, is that if the measure of probative value is flawed, the confident pronouncements about prejudice are suspect as well.<sup>4</sup>

<sup>4</sup>For cases in which it is less obvious that prejudicial impact outweighs probative value, DF (2002) imply that the two considerations should be evaluated by subtracting one percentage-point index from another. Specifically, they write that

Imagine, for example, that in our murder case we had been able to show (as we did) that the likelihood of spouse murder among unfaithful men is less than one-tenth of 1% greater than among faithful men. Thus, the fact of infidelity is not usefully probative of guilt. However, imagine in addition that we had conducted a mock jury involving 100 jurors, with two very

In this paper, we demonstrate that the measure of probative value in DF (2002) is indeed flawed. It defines probative value (evidentiary support) in terms of posterior probabilities that measure something quite different—namely, the sufficiency of that support. Despite an attempt to examine “sufficiency” separately, DF (2002) conflate the two concepts, falling into the trap that caused much confusion in the law of evidence in previous centuries. In what follows, we compare and contrast DF’s measure of probative value with a more conventional measure (DF, 2002). In the process, we apply the competing measures to specific scenarios to show that the conventional measure captures the meaning of probative value, while the Davis–Follette measure leads to undesirable, if not absurd, results.

## MEASURES OF PROBATIVE VALUE

### Absolute Difference Measures

Although they call for a “rethinking” of probative value, DF (2002) do not provide an explicit mathematical expression for probative value. Instead, they define it through an example as “the difference between the probability of murder given infidelity and the probability of murder given no infidelity” (p. 137). According to the statistics they adduce, this difference is at most  $.000923 - 0 = .000923$ . Because this arithmetic difference is so small, they conclude that “infidelity is *not* usefully probative of the likelihood of uxoricide.” (p. 137)

To clarify and generalize the analysis, let  $E$  stand for (undisputed) evidence of a trait (such as infidelity) purported to be diagnostic of a condition or hypothesis  $H$  (such as uxoricide). DF (2002) posit—with no explanation or justification—that the probative value of  $E$  is captured by the following absolute difference score (which we label  $AD_{DF}$ , for Absolute Difference, Davis–Follette measure):

$$AD_{DF} = P(H | E) - P(H | -E), \tag{1}$$

where  $P(H | E)$  is the probability of  $H$  given the existence of the trait, and  $P(H | -E)$  is the probability of  $H$  given the absence of the trait. According to this measure, the higher the difference score, the greater the probative value of the evidence  $E$ .

This definition of the probative value differs from that used in many disciplines<sup>5</sup> though it resembles another absolute difference measure proposed by Friedman

complete presentations of the case (one with and one without the evidence of infidelity), and that we had found a rate of 40% guilty votes among the 50 jurors not seeing the evidence of infidelity and a rate of 70% guilty votes among the 50 jurors who did see it. This contrast—of less than one-tenth of 1% actual greater likelihood of guilt among unfaithful men, as opposed to a 30% greater likelihood of guilty votes among those who hear of the infidelity—clearly indicates that the prejudicial impact of the evidence is greater than its probative value.

We do not believe that this is an appropriate way (1) to measure probative value, (2) to measure unfair prejudice, and (3) to compare the two. In this comment, however, we limit our analysis to the first point.

<sup>5</sup>We do not contend that there is a single, scalar measure of probative value that provides *the* correct quantification. However, any policy analysis that relies on a quantified measure should be robust. The measure itself should have desirable properties, and the policy analysis should show that the conclusions apply under all reasonable representations of probative value (Kaye, 1986a).

(1986):

$$AD_F = P(H | E) - P(H). \quad (2)$$

Here,  $P(H)$  stands for the “prior probability”—the probability of  $H$  before considering any evidence as to the presence or absence of the trait.<sup>6</sup> This definition of probative value is directly motivated by the wording of Federal Rule of Evidence 401, which defines relevant evidence as having “any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence.”<sup>7</sup>

### Ratio Measures

Federal Rule of Evidence 401 is equally consistent with a ratio measure of probative value. If the evidence  $E$  that establishes the trait does *not* tend to make the existence of  $H$  more probable or less probable than it would be without  $E$ , then the “posterior probability”  $P(H | E)$  is identical to the prior probability  $P(H)$ , and the ratio of these quantities is unity, as is the ratio of the corresponding odds. This ratio of the posterior and prior odds is known as the “Bayes factor.” Rule 401, in other words, states that evidence is relevant when the Bayes factor deviates from 1. In some situations, the Bayes factor can be estimated directly. In other circumstances, it is easier and more appropriate to estimate by forming a likelihood ratio. We focus our attention here on the likelihood ratio ( $LR$ ).

$LR$ s are especially important in forensic science because the scientist normally is not in a position to estimate the prior probability (Aitken, 1995; Ellman & Kaye, 1979; Evett, 1991). Indeed, it is *inappropriate* for forensic scientists and others who provide factfinders with an index of the strength of an evidentiary item (e.g., a DNA match between a crime scene stain and a reference sample from the suspect) to allow this index to reflect their beliefs about the prior probability in the legal case (Evett & Weir, 1998). As numerous commentators have pointed out, the assignment of prior probabilities is a task for factfinders, not for experts (Ellman & Kaye, 1979; Evett, Jackson, Lambert, & McCrossan, 2000; Finkelstein & Fairley, 1970; Risinger, Saks, Thompson, & Rosenthal, 2002; Robertson & Vignaux, 1995; Wagenaar, 1988).

The  $LR$  does not involve prior probabilities because it is based on the probability  $P(E | H)$  of the evidence  $E$  conditional on the hypothesis and the probability  $P(E | \neg H)$  conditional on its negation. These conditional probabilities are called

<sup>6</sup>It can be proved that  $P(H)$  must fall in the interval  $[P(H | E), P(H | \neg E)]$ . Hence, as  $P(H) \rightarrow 0$ ,  $AD_{DF} \rightarrow AD_F$ .

<sup>7</sup>“Without the evidence”  $E$  that defendant possesses a trait, a juror’s assessment of the hypothesis is simply  $P(H)$ . “With the evidence,” the assessment changes to  $P(H | E)$ . Hence, Eq. (2) follows directly from the text of the federal rule. The rule, as worded, does not track Eq. (1), which, as applied by DF (2002, p. 137), calls for a comparison between (a) the posterior probability of  $H$  given the fact that a trait is present and (b) the posterior probability given the fact that trait is absent. “Without the evidence” of infidelity, for example, a juror *does not know* whether the defendant was faithful or unfaithful to his wife. This juror is in a different position than one who knows that the defendant was unfaithful. Yet, DF (2002) implementation of Eq. (1) focuses on the implication for  $H$  of knowing that a man is unfaithful as contrasted to knowing that he is faithful. After all, it compares uxoricide rates among unfaithful men and faithful men. Federal Rule 401, on the other hand, suggests a comparison of uxoricide rates among unfaithful men and all men.

likelihoods, and their ratio is the *LR*:

$$LR = P(E | H) / P(E | -H). \tag{3}$$

The *LR* is a common measure of probative value in law (Finkelstein & Levin, in press; Kaye, 1986a, 1995; Koehler, 1996a; Strong, 1999), forensic science (Aitken, 1995; Evett & Weir, 1998; Kadane & Schum, 1996; Robertson & Vignaux, 1995), medicine (Black & Armstrong, 1986; Jaeschke, Guyatt, & Sackett, 1994; Lloyd, Talbot, & Lawson, 1998), and psychology (Edwards, Lindman, & Savage, 1963; Fischhoff & Beyth-Marom, 1983; Meehl & Rosen, 1955).<sup>8</sup> In some fields, the *LR* components for assessing probative (or diagnostic) value are present, but referred to by other names. For example, in medicine the terms “true positive rate” and “sensitivity” describe the numerator of the likelihood ratio (i.e.,  $P(E | H)$ ), and the terms “false positive rate” and “specificity” describe the denominator of the *LR* ( $P(E | -H)$ ), and one minus the denominator of the *LR* ( $1 - P(E | -H)$ ), respectively (Gastwirth, 1987). That is,  $LR = (\text{true positive rate}) / (\text{false positive rate}) = \text{sensitivity} / (1 - \text{specificity})$ . The sensitivity and specificity are used to “characterize the performance of a test” (Fink & Galen, 1982, p. 10), to indicate its “accuracy” (Gastwirth, 1987, p. 214), and to compare the validity of two tests (Goldman, Cook, Brand, Lee, et al., 1988, pp. 799–800).<sup>9</sup>

The *LR* (or its components) has achieved its status as a measure of probative value because the *LR* expresses exactly how much *E* shifts the prior odds in favor of *H*.<sup>10</sup> If the probability of *H* prior to considering *E* is  $P(H)$ , then the odds of *H* are  $\text{Odds}(H) = P(H) / P(-H)$ . Once *E* is considered, the odds change to

$$\text{Odds}(H | E) = \text{Odds}(H) \times LR. \tag{4}$$

Equation (4) is the version of Bayes’ rule that motivates the *LR* as a measure of probative value.<sup>11</sup> If probative value is defined to be the *LR*, then (4) means that the posterior odds in favor of hypothesis *H* are the prior odds of *H* times the probative

<sup>8</sup>A variation, which has strong roots in statistics and information theory, uses the log-likelihood ratio to measure probative value (*PV*):  $PV = \log[LR] = \log[P(E | H) / P(E | -H)] = \log[P(E | H)] - \log[P(E | -H)]$  (3’) (Edwards, 1972). Evidence is irrelevant when  $\log[LR] = 0$  (i.e., when  $LR = 1$ ). Peirce (1878) referred to the  $\log[LR]$  as the weight of the evidence in favor of *H* provided by *E*.

<sup>9</sup>The sensitivity  $P(\text{Test+} | \text{Disease})$  and specificity  $P(\text{Test-} | \text{No Disease})$  are the operating characteristics of the test. Like the false positive and false negative error probabilities,  $P(\text{Test+} | \text{No Disease})$  and  $P(\text{Test+} | \text{Disease})$ , they are independent of the prevalence of the disease in a particular population. A valid test has a large sensitivity and specificity (and correspondingly small conditional error probabilities). Another quantity, the “predictive value” of a positive test result, does not measure the extent to which a test is a valid indicator of the presence or absence of a disease. Instead, this measure relates to the “utility” or “practical usefulness” of the test as applied to a population with a particular fraction of individuals who have the disease (Vecchio, 1966; Viana & Farewell, 1990). As explained in the section on “weight versus sufficiency,” even a highly accurate (very probative) test will not perform well in identifying the cases of disease when the disease is extremely rare. This low base rate problem is not a reason to dismiss the test as invalid, but it is a reason not to rely on the positive test result alone when making a diagnosis in that population. In legal terminology, the test always has probative value, but it may not be sufficient evidence of the disease.

<sup>10</sup>See Lavine and Schervish (1999) (discussing the nature of the Bayes factor as a measure of the support the data give to one hypothesis over both simple and composite alternative hypotheses).

<sup>11</sup>It is a maxim of modern statistics that the likelihood function reflects all the information in the data relevant to choosing between two hypotheses (Birnbaum, 1969; Good, 1983; Kyburg, 1974).

value of  $E$ .<sup>12</sup> For example, if  $E$  is eight times more probable under  $H$  than  $-H$  (i.e.,  $LR = 8:1$ ), the evidence yields an eightfold increase in the odds. One would think that such evidence deserves attention. It is true, of course, that the revised probability in favor of  $H$  for a moderately probative 8:1  $LR$  will continue to be very high or very low for those whose initial beliefs in  $H$  are very strong or very weak, respectively. But it would be an error to suggest that evidence that carries an 8:1  $LR$  that appears in contexts where prior beliefs are strong has *no* tendency to prove  $H$  as opposed to  $-H$ . Yet, as the next section indicates, this is precisely what DF (2002) argue.

## ANALYSIS

The  $AD_{DF}$  differs from the  $LR$  in two respects: (1) it is framed in terms of posterior probabilities rather than likelihoods (i.e., the conditional probabilities are transposed), and (2) it involves the arithmetic difference between these posterior probabilities. Both of these features have mathematical implications that render  $AD_{DF}$  unacceptable as a measure of probative value.

The latter feature means that evidence that is powerful enough to multiply the prior odds many times over can be dismissed as essentially irrelevant. Consider, for example, an item of evidence  $E$  and a hypothesis  $H$  for which  $P(H | -E) = .00001$ ,  $P(H) = .0001$ , and  $P(H | E) = .01$ . Even though  $LR = 101$ ,<sup>13</sup> the  $AD_{DF}$  is only .00999,<sup>14</sup> indicating that  $E$  is not probative. More generally, if we characterize evidence as “not probative” whenever  $AD_{DF} < 0.01$ , then the trivial mathematical fact that  $AD_{DF}$  is bounded above by  $P(H | E)$ <sup>15</sup> means that

$$E \text{ is not probative if } P(H | E) < 0.01. \quad (5)$$

In other words, whenever  $P(H | E)$  is small, evidence will be regarded as essentially irrelevant—even when the evidence arises many times more often under  $H$  than  $-H$ .

Although we find this most peculiar,<sup>16</sup> it is not itself an argument against the difference measure. At this stage, we offer it simply to highlight the very different

<sup>12</sup>No such simple relationship exists between DF’s measure (DF, 2002) of “probative value” ( $AD_{DF}$ ) and the posterior probability or odds.

<sup>13</sup> $P(H | E) = .01$  corresponds to  $Odds(H | E) = 1 : 99$ .  $P(H) = .0001$  corresponds to  $Odds(H) = 1:9999$ .  $LR = Odds(H | E)/Odds(H) = (1:99)/(1:9999) = 9999:99 = 101:1$ .

<sup>14</sup> $AD_{DF} = P(H | E) - P(H | -E) = .01000 - .00001 = .00999$ .

<sup>15</sup>The difference measures (1) and (2) are of the form  $d = P(H | E) - x$ , where  $x$  is a conditional probability. Because  $x \geq 0$ , the value of  $x$  that maximizes  $d$  is 0. When  $x = 0$ ,  $d = P(H | E)$ . Hence, the maximum value of  $AD_{DF}$  is given by  $P(H | E)$ .

<sup>16</sup>This difference measure would have a devastating impact on the field of epidemiology if it were adopted as an indicator of the extent to which a condition is a risk factor for a disease. Assume, for instance, that the annual prevalence of oral cancers among nonsmokers is 1/10,000, while the prevalence among smokers is ten times larger, i.e., 10/10,000. According to the reasoning in DF, a person’s smoking behavior has little bearing on whether he or she will contract oral cancer. After all, the difference between  $P(\text{Oral Cancer} | \text{Smoker})$  and  $P(\text{Oral Cancer} | \text{Nonsmoker})$  is 9/10,000  $\approx 0$ . But in a city the size of Reno, Nevada (population = 180,000), about 18 people would contract this cancer if no one in the city smoked, whereas approximately 180 people would contract it if everyone smoked. In other words, some 90% of the smokers in Reno who are afflicted with this frequently fatal cancer would have been free of it had they not smoked. Clearly, smoking is a major risk factor for this cancer.

mathematical properties of  $AD_{DF}$  and  $LR$ <sup>17</sup> and to demonstrate that DF's conclusions (DF, 2002) as to lack of probative value are driven by the posterior probability  $P(H | E)$ . In the later section on "Weight versus Sufficiency," we argue that this driving probability captures the legal notion of the sufficiency of the evidence rather than probative value.

### PRIOR PROBABILITIES AND BASE RATES

The other feature of  $AD_{DF}$ , that it depends on posterior probabilities, means that it is affected by prior probabilities, which are informed by base rates and other evidence in the case.<sup>18</sup> In contrast, the  $LR$  is a constant—it does not change according to one's prior belief in  $H$ . Dependence on prior beliefs creates four problems for a measure of probative value, rendering  $AD_{DF}$  inferior to  $LR$ .

First, for jurors whose prior beliefs about  $H$  are such that they teeter at the edge of conviction, the failure to include such evidence could be the difference between a vote for acquittal and a vote for conviction. For example, a juror who is 99% sure of a defendant's guilt may not be sufficiently comfortable voting to convict. But he may change his mind if provided with evidence that is eight times more probable if the defendant is guilty than if he is not guilty (i.e.,  $LR = 8:1$ ). After all, the odds in favor of guilt have now changed from 99:1 to 792:1.<sup>19</sup> Thus, even though the new evidence should increase the juror's confidence in the guilt hypothesis by less than one percentage point (99.0% to 99.874%<sup>20</sup>), his verdict changes based on this evidence.

Second, dependence of probative value on prior probabilities implies that evidence that is probative in the presence of one set of unrelated evidentiary items may not be probative in the presence of a different set of unrelated evidentiary items. Consider evidence of postnatal sexually transmitted disease (STD) in a child who, according to the prosecution, was sexually abused by a defendant. Suppose that the nonmedical evidence in a trial (e.g., eyewitness testimony, incriminating statements by defendant) points to about a 98% chance that the defendant abused the child. By DF's measure (DF, 2002), the new STD evidence would not be probative on the

<sup>17</sup>Although simple difference measures have some intuitive appeal in the area of probabilistic reasoning, they can lead to undesirable outcomes. In the 1960s, the Supreme Court embraced a flawed difference-measure approach in jury discrimination cases. In *Swain v. Alabama* (1965), the Court construed earlier cases as demonstrating that "[w]e cannot say that purposeful discrimination based on race alone is satisfactorily proved by showing that an identifiable group in a community is underrepresented by as much as 10%" (*Swain v. Alabama*, 1965, pp. 208–209). Such an approach permits the exclusion of minorities from jury pools altogether as long as the difference between the expected proportion and observed proportion is small. A better approach is to use conventional statistical methods (such as confidence intervals and  $p$ -values) to determine whether the difference between the expected and observed proportions are beyond those that would be expected by chance, and to use a statistic such as the odds ratio for selection to ascertain whether the difference is of sufficient magnitude to suggest purposeful discrimination (Kaye, 1986b; Zeisel & Kaye, 1997, pp. 178–184).

<sup>18</sup>This section considers  $AD_{DF}$  as applied to a single item of evidence. This is how DF introduce the concept and use it to justify their conclusion that infidelity lacks probative value. A later section on "weight versus sufficiency" considers the possibility of using this difference measure to gauge the probative force of several items of evidence taken as a whole.

<sup>19</sup>As defined by (4),  $Odds(H | E) = 8:1 \times 99:1 = 792:1$ .

<sup>20</sup> $Odds(H) = 99:1$ ; therefore  $P(H) = 99/100 = 99\%$ .  $Odds(H | E) = 792:1$ ; therefore  $P(H | E) = 792/793 = 99.874\%$ .

question of abuse because the difference between  $P(\text{Abuse}|\text{STD})$  and  $P(\text{Abuse}|\text{No STD})$  is small.<sup>21</sup> In contrast, an *LR* approach finds the STD evidence to be probative regardless of the strength of the nonmedical evidence.<sup>22</sup> By an *LR* account, the evidence is probative because postnatal STDs are much more common among abused children than among nonabused children. Indeed, some postnatal STDs are virtually nonexistent among nonabused children. As such, we believe that it would be a mistake to suggest that the presence of STDs is uninformative on the question of child sexual abuse.

Third, DF's measure (DF, 2002) requires that the probative value of evidence varies depending on *when* it is presented at trial. If presented at a point in the trial when beliefs in the accused's guilt are about 50%, evidence with a small 3:1 *LR* will be judged enormously probative because it changes beliefs in guilt by 25 percentage points (from 50% to 75%). However, this same evidence will seem woefully lacking in probative value by DF's measure (DF, 2002) at another point in the trial if belief in the accused's guilt has fallen to 1%. In this case, the 3:1 *LR* should change beliefs in guilt by less than 2 percentage points (from 1% to 2.94%), and DF would likely treat the evidence as worthless.

Finally, there is something strange about linking probativity to prior beliefs when the only prior beliefs that count are those of the factfinders who have access to all of the evidence. Whose priors should be used to make probative value determinations? The expert's? The judge's? The jurors'? By DF's measure (DF, 2002), this means that the probative value of the evidence varies across people. Do we really want to conclude that the probative value of an item of scientific evidence varies across individuals depending on their predispositions and views of other evidence?

It is important to note that DF (2002) do not sidestep this set of criticisms by confining their examples to those in which they assume that decision makers will adopt the available base rate as their prior. There are three reasons this strategy fails. First, prior beliefs are not the same as base rates.<sup>23</sup> Base rates may *inform* priors, but priors may also be affected by personal experiences, biases, beliefs, and knowledge of other base rates (Koehler, 1996b). As a result, it is usually not reasonable to presume that a single base rate accurately identifies the priors of all decision makers. Nevertheless, DF treat base rate and prior probability as synonyms. For example, in the short section entitled "What if the prior probability of guilt is known?"

<sup>21</sup>The vast majority of both abused and nonabused children do not contract STDs (Lyon & Koehler, 1996, p. 72, and reference cited therein at p. 59, fn 58). Therefore, in a child sexual abuse case where the nonmedical evidence suggests a very high probability of guilt, a finding that the alleged victim lacks STDs has little impact on the probability of guilt. In other words, if  $P(\text{Abuse})$  is very high, then  $P(\text{Abuse}|\text{No STD})$  is also high. Furthermore, if  $P(\text{Abuse})$  is very high, then  $P(\text{Abuse}|\text{STD})$  is also very high. Consequently,  $P(\text{Abuse}|\text{STD}) - P(\text{Abuse}|\text{No STD})$ , is small.

<sup>22</sup>Although DF focus on probativity of evidence proffered in a low base rate context, their measure must seem sensible in moderate and high base rate contexts as well to provide a serious alternative to existing probativity measures. By this standard, their measure fails.

<sup>23</sup>Even if priors were the same as base rates, the 4 per 1,000,000 base rate offered up in DF (2002) paper as the uxoricide base rate is an inappropriate prior for cases in which a woman died under mysterious circumstances. The question is "Did this husband kill his wife?" The reference class from which to compute a base rate is not the broad population of "husbands." We can do much better than this. A more informative reference class—one that incorporates the central feature of the case—is the population of "husbands whose wives died under mysterious circumstances." Surely the base rate for uxoricide in this select group is *much* larger than 4 in 1,000,000.

(pp. 143–144), they describe an example in which they use “base rate” four times and “prior probability” three times. Second, and more fundamentally, it is hard to imagine the underlying justification for defining the probative value of evidence one way when a base rate statistic is available and another way when the prior probability is partially ascertained by reference to the base rate. Finally, if DF arbitrarily choose to confine their analysis to the limited case in which priors really *are* identical to base rates, then their analysis has nothing to say about the vast preponderance of cases in which *any* information other than base rates exists for assessing the prior probability of guilt.

Readers already familiar with modern theories of statistical inference may find these conclusions obvious.<sup>24</sup> For those who are not yet convinced, we offer a final example that involves a common form of scientific evidence. The evidence would be regarded as quite valuable by a conventional *LR* analysis but rejected as worthless by the  $AD_{DF}$  measure. These interpretations cannot both be correct.

### AN EXAMPLE—DNA TESTING

A bloodstain recovered from the fingernails of a woman killed by a man who was seen fleeing came from this killer. DNA in the stain is analyzed at ten STR loci and compared to the STR alleles of a suspect.<sup>25</sup> There is a clear and accurate match at each locus. The alleles at each locus occur independently in the relevant population with a relative frequency of one-tenth.<sup>26</sup>

Under these stylized conditions, the likelihoods for the hypothesis *H* that the suspect is the killer (assuming that the killer is the source of the bloodstain) are easily computed. If *H* is true, then the match at the first locus occurs with probability  $P(E_1 | H) = 1$ . If *H* is false, the probability is  $P(E_1 | -H) = 1/10$ . The *LR* for this first datum is therefore  $LR_1 = 1/(1/10) = 10$ . The same result applies to every other locus, and the *LR* for the series of tests is

$$LR = LR_1 LR_2 \dots LR_{10} = (1/10)10 = 1/10, 000, 000, 000. \tag{6}$$

By an *LR* measure of probative value, the test result at each locus ( $LR = 10:1$ ) is moderately probative, and the series of results ( $LR = 10$  billion:1) is extremely probative.

It is not clear to us how Davis and Follette (2002) would treat this evidence. Using  $AD_{DF}$  as they did in their murder case, it would seem that the evidence should be excluded. According to the Bureau of Justice Statistics, the murder rate in the general population of the United States is about 6/100,000 (Bureau of Justice Statistics, 2002). Because some individuals commit more than one murder, the base rate for murderers

<sup>24</sup>For an earlier criticism of measuring probative value in a way that depends on the prior probability, see Edwards (1986, pp. 625–626).

<sup>25</sup>STR typing is described in, e.g., Butler (2001) and Kaye and Sensabaugh (2000).

<sup>26</sup>For simplicity, we assume that there are no relatives of the suspect in the relevant population.

<sup>27</sup>This expression illustrates a nice property of *LR* as a measure of probative value. The probative value of a series of conditionally independent items of evidence is the product of the probative value of each item. With difference measures of probative value, there is no comparable function that relates the probative value of the series to the probative value of the conjoined items.

is no more than this figure. Since the alleles at the STR loci used in forensic tests are unrelated to any behavior, let alone murder, the chance that an individual with the STRs of the perpetrator would commit a murder is also 6/100,000. So is the probability for an individual without these particular STRs. The  $AD_{DF}$  is therefore zero.

Or is it? Perhaps DF (2002) would argue that, with such a rare genotype, the probability of an individual with the genotype being the murderer is nearly one, while the probability of someone with a different genotype being the murderer in this case is zero. However, even if this were consistent with the approach taken in the spousal murder case, it would not work for less extreme but still clearly probative items of evidence. Imagine, for instance, that only two loci were suitable for testing. Then the genotype frequency would be 1/100. In a population of 1 million people, there would be approximately 1,000 individuals with the perpetrator's genotype. Even though the suspect is one of these people, the match itself does not establish that it is practically certain that he is the source. To the contrary,  $P(H | E)$  is close to zero. Hence, the difference between  $P(H | E)$  and  $P(H | -E)$  is approximately zero, and this implementation of  $AD_{DF}$  suggests that evidence that narrows the class of potential suspects by a factor of 100 is logically worthless.

### WEIGHT VERSUS SUFFICIENCY

Small posterior probabilities, as estimated from a small base rate (when base rates and prior probabilities are roughly identical) and one item of evidence, do not mean that evidence is irrelevant or lacking in probative value. That  $P(H | E)$  is only slightly larger than  $P(H | -E)$  simply means that the evidence is insufficient to swamp the base rate. One should not expect or require every item of admissible evidence to be persuasive in and of itself. As Judge Learned Hand explained nearly 60 years ago, “[a]ll that is necessary . . . is that each bit may have enough rational connection with the issue to be considered a factor contributing to an answer” (*United States v. Pugliese*, 1945). Were the law otherwise, few cases could get off the ground.

In presenting their new measure of probative value, however, DF (2002) isolate a single item of evidence (proof of infidelity), compare it to a single, ill-chosen base rate, and reach the conclusion that the evidence has negligible probative value. This reasoning is riddled with errors. First, the defendant's infidelity was not the only evidence suggesting guilt. There was, for example, the additional proof of a possible motive to recover a large life insurance payment. Second, the authors mistakenly suggest that the low base rate for uxoricide provides strong evidence against the hypothesis that the defendant murdered his wife. This base rate ignores the crucial fact that the defendant's wife died violently. As we suggested earlier, the base rate for uxoricide among men whose wives died violently is much larger.

But even putting these issues aside, DF (2002) have not shown that proof of infidelity fails the “factor contributing to an answer” test for probative value outlined by Judge Hand (and endorsed in the notes of the Advisory Committee that drafted the Federal Rules of Evidence). At best, they show that the evidence, standing alone, is insufficient to support a verdict. As we have noted,  $AD_{DF}$ , as applied to a single

item of evidence  $E^{28}$  with a low prior probability for  $H$ , is essentially the posterior probability  $P(H | E)$ . This probability tells us where a juror stands after considering the evidence. It addresses the question of sufficiency: Is the case, with this evidence included, convincing? It does not address the issue of probative value: Is the evidence such that it should *move the juror* a bit farther in the direction of being convinced? The law appropriately recognizes that evidence that is insufficient to prove guilt can still be quite probative.

The situation is analogous to the low base rate problem in medical screening tests, in psychiatric predictions of dangerousness, and in many other contexts. As DF (2002) correctly note, “for populations with an extremely low base rate of the criterion (whether behavior or disease), even using the most accurate medical tests or most predictive psychological assessments, the rate of false positive predictions far exceeds the rate of true positive predictions” (p. 135). But this phenomenon occurs not because “the most accurate medical tests” and the “most predictive psychological assessments” lack probative value. It occurs because the prior probability for a condition is so low that the test for the condition leads to only a small increase in belief that the condition exists in the target case.

In these circumstances, the positive result of a probative test has a small “probative value positive” (PVP).<sup>29</sup> PVP is the proportion of true positives in all positives (Vecchio, 1966). It is a well-known and useful concept in fields such as medicine and clinical psychology where base rates are often low and cue diagnosticity is not extremely high. DF’s absolute difference measure (DF, 2002) essentially rediscovers PVP and mistakenly refers to it as probative value. In the spousal murder case, DF

<sup>28</sup>  $AD_{DF}$  also can be applied to a conjunction of items of evidence. In a section entitled “sufficiency evaluation through base rate analyses of multiple predictors,” DF seem to say that when there is other, independent evidence indicative of guilt, the evidence that was to have been excluded as “not probative” somehow can become informative in that the  $AD_{DF}$  can be large for the entire body of evidence. They write that:

[O]ne might find that although each single predictor or item of evidence may not be probative of guilt (or sufficient proof of guilt) alone, they are informative when combined together. In other words, if the defendant has multiple characteristics believed to be associated with the to-be pre(post)dicted behavior, then the combined predictors are probative, or highly suggestive of guilt.

Does this mean that an individual “predictor,” such as infidelity, that was said to have “miniscule” probative value and to lead to false inferences of guilt “99.907% of the time” should be admitted after all? Instead of starting with a base rate as a prior probability, should one compute marginal values of the  $AD_{DF}$  for each item of evidence assuming that every other piece of evidence already has been introduced, and adopt these as the appropriate measure of probative value? We fear that this line of reasoning would lead nowhere, and are inclined to dismiss the entire effort to use a proposed measure of probative value as a device to ascertain sufficiency as needlessly confusing.

<sup>29</sup> When the base rate is low, the vast bulk of individuals tested are actually negative. As a result, even with a highly accurate test—one that correctly classifies almost 100% of the people with the condition as positives and that misclassifies only a few percent of the people who are free from the condition as positives—most of the positives will be false. The reason is that the pool of positives consists of (a) a low percentage of the large number of condition-free individuals tested plus (b) a high percentage of the small number of affected individuals tested. For instance, suppose that the test correctly classifies 99% of all individuals with the condition as positive and incorrectly classifies only 1% of individuals free of the condition as positive, and that the base rate is such that for every rare individual with the condition, there are 100 individuals free from the condition. Then for a batch of 100 affected individuals and 10,000 unaffected ones, there will be  $99\% \times 100 = 99$  true positive and  $1\% \times 10,000 = 100$  false positives. Thus, the PVP for this highly accurate test is only 1/101.

have shown that the PVP for scoring positive on the “infidelity test” is small when the test is used on the general population.<sup>30</sup> Because the PVP is small, they have concluded that the information is useless. Similar reasoning has been used to suggest that the polygraph is not a valid lie detector and that psychiatric predictions of dangerousness are invalid (Kaye, 1987a, 1987b, 1987c). A leading example is *Barefoot v. Estelle* (1983), where, quoting Ennis and Litwak (1974, p. 737), three Supreme Court Justices wrote that psychiatric predictions of dangerousness are “less accurate than the flip of a coin.” In reality, however, coin flipping is plainly invalid as a test for future violence, but clinical and statistical predictions have some claim to validity (Slobogin, 1984, pp. 111–112). Because the PVP is a function of both the operating characteristics of a test and the base rate in the population on which it is used, it is a fallacy to conclude that because the PVP is low, the test is invalid (Kaye, 1997, p. 272). The risk of succumbing to this fallacy is enhanced if one conflates “not proved” with “not probative.” A low PVP means “not proved.” It does not mean that the use of infidelity in classifying mysterious deaths is, like flipping coins, not probative of uxoricide.

In contrast to the  $AD_{DF}$  measure for probative value, Bayes’ rule and the  $LR$  definition make the legal distinction between weight and sufficiency crystal-clear (Lempert, 1977). On the one hand, probative value, or weight, pertains to the shift in the odds brought about by the evidence. That quantity is the  $LR$  for that evidence. On the other hand, the sufficiency of a body of evidence  $E_1$  through  $E_n$  depends not merely on the total weight,  $LR = LR_1 LR_2 \dots LR_n$  (for independent evidence), but also on the starting odds,  $Odds(H)$ . When the posterior odds, as given by  $Odds(H) \times LR$ , of the prosecution’s version of the events exceeds some threshold value, the body of evidence is sufficient to support a verdict of guilty. The threshold value, in turn, is a function of the losses (in terms of utilities) wrought by false convictions and false acquittals (Kaye, 1987b, 1999). These utilities affect the verdict, but not the probative merit of individual items of evidence.

### WHAT SHOULD AN EXPERT TELL THE COURT?

We have argued that the probative value of evidence is more appropriately described by the  $LR$  than as the absolute difference measure ( $AD_{DF}$ ) that DF (2002) offer. Neither base rates nor prior probabilities associated with the ultimate issue play any role in the determination of likelihood ratios. Therefore, experts who are charged with helping a court understand the probative value of  $E$  should confine their testimony to matters related to the chance of observing  $E$  under various theories of the case. We side with the many scholars who have shown that neither base rates nor prior probabilities are relevant for a determination of the strength of  $E$  (Ellman & Kaye, 1979; Evett et al., 2000; Evett & Weir, 1998; Risinger et al., 2002; Robertson

<sup>30</sup>Because  $P(H | -E)$  is effectively zero in the spousal murder case, the term  $P(H | E)$  dominates:

$$AD_{DF} = P(H | E) - P(H | -E) \approx P(H | E). \quad (7)$$

When  $P(H | E)$  is based on the prevalence of the condition and a positive test result, it is simply the PVP.

& Vignaux, 1995; Wagenaar, 1988).<sup>31</sup> Therefore, an expert who wished to provide insight into the probative value of infidelity on uxoricide should provide the court with the ratio of the proportion of men known to have killed their wives who were unfaithful and the proportion of men known not to have killed their wives who were unfaithful. The base rate for uxoricide is irrelevant for this purpose.<sup>32</sup>

## CONCLUSION

In a revealing passage, DF (2002) note that “[o]ur experience in discussions of these ideas—even with intelligent colleagues and students—has been that they understand the math, but nevertheless find it difficult to accept” (p. 153). These students and colleagues are correct in sensing that something is amiss. It is the use of the difference between two posteriors as a measure of probative value. To justify a “rethinking of probative value,” DF would need to demonstrate why their posterior-probability difference is superior to the conventional likelihood measures that they inexplicably overlook.

DF (2002) are right about one thing—“there will be great resistance to the presentation or acceptance of such analyses” (p. 156). This is as it should be. Their analysis of probative value provides no assistance to a court attempting to weigh probative value against prejudicial impact or to jurors deliberating on the significance of the evidence.

## ACKNOWLEDGMENTS

The authors are grateful to Colin Aitken, David Bernstein, Craig Callen, Craig McKenzie, Molly Mercer, Michael Redmayne, and Steven Penrod for their comments on the issue discussed in this paper, and to Deborah Davis, Susan Ehrlich, David Freedman, Peter Killeen, and Gary Wells for comments on a draft of the paper.

## REFERENCES

- Aitken, C. G. G. (1995). *Statistics and the evaluation of evidence for forensic scientists*. Chichester, United Kingdom: Wiley.
- Barefoot v. Estelle*, 463 U.S. 880 (1983).
- Birnbaum, A. (1969). Concepts of statistical evidence. In S. Morgenbasser et al. (Eds.), *Philosophy, science, and method* (pp. 112–143). New York: St. Martin's Press.
- Black, W. C., & Armstrong, P. (1986). Communicating the significance of radiologic test-results: The likelihood ratio. *American Journal of Roentgenology*, *147*, 1313–1318.
- Bureau of Justice Statistics. (2002). *Criminal victimization: Summary findings*. Retrived March 19, 2002 from <http://www.ojp.usdoj.gov/bjs/cvictgen.htm>

<sup>31</sup>“It is not the expert’s task to tell a court what its prior odds are” (Wagenaar, 1988, p. 502).

<sup>32</sup>Of course the base rate for uxoricide does provide relevant information for a computation of the posterior probability that an unfaithful defendant is guilty of uxoricide. It does so by informing the prior odds, Odds( $H$ ). An expert may also provide testimony about various base rates, not as indices of probative value, but as background information.

- Butler, J. M. (2001). *Forensic DNA typing: Biology and technology behind STR markers*. San Francisco: Academic Press.
- Davis, D., & Follette, W. C. (2002). Rethinking the probative value of evidence: Base rates, intuitive profiling, and the "postdiction" of behavior. *Law and Human Behavior*, 26, 143–158.
- Edwards, A. W. F. (1972). *Likelihood*. Cambridge, MA: Cambridge University Press.
- Edwards, W. (1986). Comment. *Boston University Law Review*, 66, 623–628.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193–242.
- Ellman, I. M., & Kaye, D. H. (1979). Probabilities and proof: Can HLA and blood test evidence prove paternity? *New York University Law Review*, 55, 1131–1162.
- Ennis, B. J., & Litwak, T. R. (1974). Psychiatry and the presumption of expertise: Flipping coins in the courtroom. *California Law Review*, 62, 693–752.
- Evetts, I. (1991). Interpretation: A personal odyssey. In C. G. G. Aitken & D. A. Stoney (Eds.), *The use of statistics in forensic science* (pp. 9–22). New York: Ellis Horwood.
- Evetts, I. W., Jackson, G., Lambert, J. A., & McCrossan, S. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science and Justice*, 40, 233–239.
- Evetts, I. W., & Weir, B. S. (1998). *Interpreting DNA evidence: Statistical genetics for forensic scientists*. Sunderland, MA: Sinauer Associates.
- Fink, D. J., & Galen, R. S. (1982). Probabilistic approaches to clinical decision support. In B. T. Williams (Ed.), *Computer aids to clinical decisions* (Vol. 2, pp. 1–65). Boca Raton, FL: CRC Press.
- Finkelstein, M. O., & Fairley, W. B. (1970). A Bayesian approach to identification evidence. *Harvard Law Review*, 83, 489–517.
- Finkelstein, M. O., & Levin, B. (2003). On the probative value of evidence from a screening search. *Jurimetrics Journal*, 43, 265–290.
- Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, 90, 239–260.
- Friedman, R. D. (1986). A close look at probative value. *Boston University Law Review*, 66, 733–759.
- Gastwirth, J. L. (1987). The statistical precision of medical screening procedures: Application to polygraph and AIDS antibodies test data. *Statistical Science*, 2, 213–238.
- Goldman, L., Cook, E. F., Brand, D. A., Lee, T. H., Rovin, G. W., Weisberg, M. C. (1988). A computer protocol to predict myocardial infarction in emergency department patients with chest pain. *New England Journal of Medicine*, 318, 797–803.
- Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. Minneapolis: University of Minnesota Press.
- Jaeschke, R., Guyatt, G., & Sackett, D. L. (1994). Users' guides to the medical literature III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-based Working Group. *Journal of the American Medical Association*, 271, 389–391.
- Kadane, J. B., & Schum, D. A. (1996). *A probabilistic analysis of the Sacco and Vanzetti evidence*. New York: Wiley.
- Kaye, D. H. (1986a). Quantifying probative value. *Boston University Law Review*, 66, 761–766.
- Kaye, D. H. (1986b). Statistical analysis in jury discrimination cases. In D. H. Kaye & M. Aickin (Eds.), *Statistical methods in discrimination litigation* (pp. 13–32). New York: Marcel Dekker.
- Kaye, D. H. (1987a). The validity of tests: Caveat omnes. *Jurimetrics: The Journal of Law, Science and Technology*, 27, 349–361.
- Kaye, D. H. (1987b). Apples and oranges: Confidence coefficients and the burden of persuasion. *Cornell Law Review*, 73, 54–77.
- Kaye, D. H. (1987c). The polygraph and the PVP. *Statistical Science*, 2, 223–226.
- Kaye, D. H. (1995). The relevance of "matching" DNA: Is the window half open or half shut? *Journal of Criminal Law and Criminology*, 85, 676–695.
- Kaye, D. H. (1997). *Science in evidence*. Cincinnati, OH: Anderson.
- Kaye, D. H. (1999). Clarifying the burden of persuasion: What Bayesian decision rules do and do not do. *International Journal of Evidence and Proof*, 3, 1–28.
- Kaye, D. H., & Sensabaugh, G. F. (2000). Reference guide on DNA evidence. In *Reference manual on scientific evidence* (Federal Judicial Center ed., pp. 485–576). Washington, DC: Federal Judicial Center.
- Koehler, J. J. (1996a). On conveying the probative value of DNA evidence: Frequencies, likelihood ratios and error rates. *University of Colorado Law Review*, 67, 859–886.
- Koehler, J. J. (1996b). The base rate fallacy reconsidered: Normative, descriptive and methodological challenges. *Behavioral and Brain Sciences*, 19, 1–53.
- Kyburg, H. E. (1974). *The logical foundations of statistical inference*. Dordrecht, The Netherlands: D. Reidel.

- Lavine, M., & Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, 53, 119–122.
- Lempert, R. (1977). Modeling relevance. *Michigan Law Review*, 75, 1021–1057.
- Lloyd, J. J., Talbot, P. R., & Lawson, R. S. (1998). Quantifying the value of diagnostic tests. *Nuclear Medicine Communications*, 19, 999–1004.
- Lyon, T. D., & Koehler, J. J. (1996). The relevance ratio: Evaluating the probative value of expert testimony in child sexual abuse cases. *Cornell Law Review*, 82, 43–78.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52, 194–216.
- Peirce, C. S. (1878). The probability of induction. *Popular Science Monthly*, reprinted In J. R. Newman (Ed.), *The world of mathematics* (1956 ed., pp. 1341–1354). New York: Simon and Schuster.
- Risinger, D. M., Saks, M. J., Thompson, W. C., & Rosenthal, R. (2002). The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *California Law Review*, 90, 1–56.
- Robertson, B., & Vignaux, G. A. (1995). *Interpreting evidence: Evaluating forensic science in the courtroom*. Chichester, United Kingdom: Wiley.
- Slobogin, C. (1984). Dangerousness and expertise. *University of Pennsylvania Law Review*, 133, 97–174.
- Strong, J. W. (Ed.). (1999). *McCormick on evidence* (5th ed.). St. Paul, MN: West Publishing Group.
- Swain v. Alabama*, 380 U.S. 202 (1965).
- United States v. Pugliese*, 153 F.2d 497, 500 (2d Cir. 1945).
- Vecchio, T. J. (1966). Predictive value of a single diagnostic test in unselected populations. *New England Journal of Medicine*, 274, 1171–1173.
- Viana, M., & Farewell, V. (1990). A test for diagnostic utility. *Canadian Journal of Statistics*, 18, 289–295.
- Wagenaar, W. A. (1988). The proper seat: A Bayesian discussion of the position of expert witnesses. *Law and Human Behavior*, 12, 499–510.
- Zeisel, H., & Kaye, D. (1997). *Prove it with figures: Empirical methods in law and litigation*. New York: Springer-Verlag.