

Toward an Empirical Approach to Evidentiary Ruling

Deborah Davis^{1,2} and William C. Follette¹

This paper responds to criticisms/misconstruals of our measure of the maximum probative value of evidence (D. Davis & W. C. Follette, 2002), and our conclusions regarding the potentially prejudicial role of “intuitive profiling” evidence, including motive. We argue that R. D. Friedman and R. C. Park’s (2003) criticisms and example cases are largely based on inappropriate violation of the presumption of innocence. Further, we address the merits of our absolute difference measure of probative value versus those of the Bayesian likelihood ratio championed by D. H. Kaye and J. J. Koehler (2003). We recommend methods for presentation of measures of evidence utility that convey complexities of interdependence between new and existing evidence. Finally, we propose a “probable cause” standard for admission of potentially prejudicial evidence, dictating that admissibility of such evidence should be contingent upon other substantial evidence of guilt.

KEY WORDS: evidence; probative value; motive; utility; profiling; empirical.

Participants in this exchange are among a growing number of social science and legal scholars who advocate probabilistic approaches to evaluation of evidence. Notwithstanding the existence of differences in specific approach among the participants, there exist fundamental areas of agreement that underscore the value of statistical approaches to assessing the utility of evidence. Hence, after some initial commentary on the crucial issue of the presumption of innocence and some broad misconstruals of our approach by our critics, we examine these areas of agreement. Following this, we state our position regarding how to construct and present indices of evidence utility, and in this context we explore areas of disagreement between our approach and those of our critics.

Although our original paper and those of our critics have devoted considerable attention to the Franklin case and other evidentiary issues specifically involving uxoricide, this paper will focus on broader issues of evidence utility. Our analysis of the Franklin case was based upon the specific assignment we were given—namely, to determine whether unfaithful men were more likely to murder their wives—and

¹Department of Psychology, University of Nevada, Reno, Nevada.

²To whom correspondence should be addressed at Department of Psychology, University of Nevada, Reno, Nevada 89557; e-mail: debdavis@unr.nevada.edu or follette@unr.edu.

the characterization of the evidence by our client earlier in the prosecution case and prior to my (Davis) testimony—which lacked (and specifically denied) much of the physical and other evidence reported in Keller’s (2001) book (of unknown accuracy) regarding the case. We considered only the motive of infidelity, and based our analyses primarily upon the assumptions given to us by our client of (1) lack of certainty regarding the issue of murder versus accident and (2) no useful evidence other than motive—assumptions that clearly warrant revision in light of the full evidentiary record. Our original paper (Davis & Follette, 2002) clearly illustrates the way in which the probative value of infidelity, a history of battery, or other evidence depends upon assumptions regarding the state of other evidence of guilt (or the presumed base rate of guilt given other evidence). Hence, we do not revisit the specific evidence in the Franklin case—or evidentiary issues involving uxoricide in particular. Instead, we focus upon principles we believe should underlie objective and empirical analyses of evidence utility, some of them undisputed and others hotly contested among the participants in this exchange.

THE ROLE OF THE PRESUMPTION OF INNOCENCE

In line with the assignment we were given by our client, the original application of our method for calculating maximum probative value (MPV) was intended to show the MPV of a single item of evidence alone. As noted earlier, the case as described to us centered upon motive as the primary, if not only, evidence of murder. Hence, we chose the population base rates of uxoricide and of infidelity as the basis of our calculations. That is, assuming we have no existing incident-specific evidence against the defendant, how much would evidence of infidelity enhance the likelihood of guilt?

Friedman and Park (2003; henceforth FP) have criticized our analysis essentially on the grounds that we did not adjust the base rate of guilt to reflect (a) the fact that a specific crime (or suspicious death) has occurred and (b) the fact that the defendant falls in a category (i.e., husband) deemed likely to commit the particular crime. Offering a variety of crime scenarios involving uxoricide, they reason that the husband’s probability of guilt is in actuality typically enhanced relative to the base rate on these two grounds alone. (Hence, the proper starting base rate of uxoricide is really quite large.) They further suggest that if a person is brought to trial, there will typically be other evidence that justifies enhancing the prior probability of guilt, and that it is rare that the issue is whether death was due to accident or murder. Essentially then, they suggest that there are no circumstances in which we should assume the population base rates, and none where the prior probability of guilt (PPG) should be assumed low. Hence, they imply that in the absence of low base rates (or PPG) our analysis is irrelevant, and that evidence such as battery, infidelity, profit, or other motives will clearly be relevant under all realistic circumstances.

As we have emphasized in our original paper and throughout this one, we agree with the basic premise that the probative value of evidence will change with the base rate (or prior probability) of guilt, and we have no quarrel with FP’s (2003) calculations, given their assumptions. However, in contrast to FP, we argue that the

use of population base rates as PPG should be adjusted only in response to incident-specific evidence—not in response to the very fact the crime (or potential crime) has occurred, or to the defendant's characteristics or category (such as "husband"). Although such evidence may be actuarially associated with greater probability of guilt, its use violates the presumption of innocence for the individual defendant—which is usually understood to mean that the trial begins with no prior probability against the accused (e.g., Scolnicov, 2000). Further, (as we discuss later) crime-related actuarial data is typically biased such that its utility, even if permissible, is compromised by invalidity.

Violation of the Presumption of Innocence

To understand how FP's illustrations violate the presumption of innocence, it may be helpful to consider a distinction that has been drawn by some evidence scholars (e.g., Sanchirico, 2001) between *trace* evidence (evidence resulting from the commission of the *specific crime* at hand) and *predictive* evidence (factors—often individual differences—statistically associated with commission of the *type of crime* in question). That is, trace evidence more typically is evidence that the person *did* commit the *specific* crime, whereas predictive evidence suggests the individual *might be more inclined* to commit the *type* of crime. Wells (1992) refers to this distinction as *trace* versus *naked statistical* evidence, noting that whereas naked statistical evidence would exist even if the event in question did not occur, trace evidence exists because (and only if) the event (crime) occurred.

Essentially, FP suggest that we should use predictive evidence to enhance PPG for use in our analysis. We believe this to be an inappropriate violation of the presumption of innocence, as well as inappropriate use of essentially inadmissible "profiling" evidence. Indeed, Friedman (2000) himself has argued against the use of prior odds in violation of the presumption of innocence as the basis of calculating posterior odds in other contexts, stating that although such adjustments might be rationally or empirically justified, *what is rational for a casual observer is not permissible for a juror* (p. 879). *A basic aspect of the presumption is that the jury may not treat the fact that the defendant is on trial as itself an indication of guilt . . . The answer, I believe, is that Juror should assess prior odds just as if she were implicitly assessing the odds outside the adjudicative system.* (pp. 879, 880). Should these principles be different when applied to evidentiary ruling?

More generally, an issue exists regarding the appropriate use of predictive evidence—and the way in which it should be combined with existing evidence. Some have argued, for example, that predictive evidence (including character evidence) should play *at best* a secondary role to trace evidence (e.g., Sanchirico, 2001). Consider the hypothetical example in which the defendant is accused of shooting another person in the course of a disagreement. Trace evidence has indicated posterior odds of guilt of 50% (1/1). Suppose the odds ratio of being male if guilty versus if not guilty is 7-1 (indeed, males are approximately seven times more likely than females to engage in violent crime). Hence, knowing that the defendant is male would increase the likelihood of guilt to 87.5%. Is it sensible to combine trace and predictive evidence in this way?

To take a more pernicious example, suppose trace evidence had indicated prior probability of guilt to be 80% (4/1 odds)—clearly below the threshold of beyond reasonable doubt. Now, suppose existing actuarial data to suggest that Blacks are four times more likely to commit the crime in question than Whites. The posterior odds of guilt would rise to 16/1, or 94.1%—and likely cross the reasonable doubt threshold for many jurors.

Juror verdicts are known to reflect the operation of stereotypes associating social categories with criminal behaviors (Saks & Kidd, 1980–81, see a review in Koski, 2001). Is this a proper application of Bayesian reasoning to evidence, or is the reasoning that *Members of this social group are more likely than others to commit such crimes, therefore it is more likely that this individual member of this group committed this crime* a generally inappropriate inference, and an inappropriate combination of trace and predictive (profiling) evidence? Although some would view such reasoning as *intelligent Bayesianism* (e.g., Armour, 1994), we heartily disagree.

Biases in Predictive Data

Finally, even if one believed use of predictive evidence should modify the presumption of innocence, it should do so only if reasonably accurate. However, crime statistics indicating associations of predictive evidence with guilt are necessarily inaccurate—overwhelmingly in the direction of overestimation. Generally, the data collection process is subject to the *confirmation bias* (e.g., Copeland & Snyder, 1995; Jonas, Schulz-Hardt, Frey, & Thelen, 2001; Snyder & Thomsen, 1988), whereby perceivers are prone to search the information pool and interpret information in a way that tends to confirm expectations, whether justified by the information or not. For example, the statistic that at least 50% of murdered women are murdered by their husbands represents only solved—and not unsolved—cases, *in which someone was convicted*. Hence, the high rate of conviction of husbands may be the result of biasing assumptions reflected in police selective targeting of husbands for investigation (and failure to identify and investigate alternative suspects), selective collection of expectation relevant facts and evidence, and biased interpretation of evidence, as well as the effects of such assumptions on decisions to prosecute, evidentiary rulings, and finally juror judgments (see Yin, 2000, for an excellent discussion of the contribution of such biases to the establishment of criminal profiles). Similar biases may taint statistics regarding relative likelihood of battery among murderers and nonmurderers. Once a woman is murdered, police may actively inquire about battery. Further, the woman's relatives will likely be happy to report battery if it existed. Hence, if battery occurs it is highly likely to be detected once the husband is suspected of murder. On the other hand, population base rates are collected via methods where respondents may choose to deny actual battery—leading to underestimates of the true rate. As a result, the apparent difference between $p(M|B)$ and $p(M|B-)$, or between $p(B|M)$ and $p(B|M-)$, will be exaggerated. These and others kinds of problems with collection of pertinent data for predictive evidence render their use to adjust the presumption of innocence, and even posterior probabilities of guilt in light of other evidence, highly suspect.

In summary, we believe it is inappropriate to use predictive crime statistics as the basis of adjusting the presumption of innocence. Further, because of their inherent inaccuracy and biases toward exaggeration, their use would more often than not result in overestimation of probative value. We expressed both concerns in our original paper, to explain why we viewed adjustments based on predictive evidence as “arguable.”

MISCONSTRUAL OF OUR APPROACH TO ASSESSMENT OF EVIDENCE UTILITY

Our original paper (Davis & Follette, 2002) focused upon two broad areas. First, we offered a definition of probative value in terms of differences in likelihood of a proposition ($p(H)$) such as guilt, given an item of evidence (E) versus given the lack of the evidence ($E-$), and a proposal for how to estimate *maximum probative value* for that evidence using the base rate of the evidence and the base rate of the proposition (or prior probability of the proposition). This was done essentially by maximizing the probability of the evidence given the proposition (i.e., $p(E|H)$), given the two population base rates. In other words, we assumed the maximum possible association between the evidence and the crime (or other propositions linked to the crime). Second, we applied our proposed measure of MPV to the issue of motive evidence, offering the conclusion that it will possess little to no probative value when the base rate of the evidence far exceeds that of the crime, but great potential for prejudicial impact. Our critics have misconstrued aspects of our position in both of these areas.

Generalizations Regarding the Utility of Evidence in Light of Low Base Rate Crimes

Perhaps the most pervasive misconstrual of our work on the part of the other contributors to this exchange is the repetitive offering of irrelevant examples to suggest that we would regard clearly relevant evidence as irrelevant. Our original illustrations clearly showed the probative value of evidence to depend upon the base rate (or prior probability) of the proposition (in our examples, the PPG). Hence, evidence such as infidelity or battery that may be minimally probative when it is not clear whether death was due to accident or murder (*and* other evidence has not significantly elevated the PPG), may become highly probative when murder is certain and other evidence indicates guilt—and we offered illustrations and calculations to illustrate this difference. Nevertheless, our critics continue to suggest that we would broadly consider infidelity, battery, DNA, and other forms of evidence as irrelevant, irrespective of the prior odds of guilt.

In their showcase example to illustrate what they believe to be the absurdity of our position, Kaye and Koehler (2003, henceforth KK) set up the following problem. A bloodstain found under a murder victim’s fingernails is assumed to come from the killer. The suspect’s DNA is found to match that of the bloodstain. The likelihood that this would happen by chance is 1 in 10 billion. They then go on to suggest that we would not regard such evidence as probative of the suspect’s guilt, because the base rate of murder is 6 per 100,000. This example, however, is irrelevant to our analysis. Why?

First, they set up the problem to *assume* that the bloodstain came from the killer. In light of this assumption, the DNA match is clearly definitive. Second, DNA match statistics are typically offered to prove that the suspect is the source of bodily products—not to directly prove murder. Hence, the base rate of murder is not relevant. The LR measure of probative value KK offer is the odds of a DNA match, given that the two samples come from the same person versus that they come from different people. This yields the probative value of 10 billion to 1. Then they suggest that our measure should be calculated as the probability of murder given a DNA match versus the probability of murder given no match. This is comparing apples and oranges—calculating their measure to indicate probative value for the conclusion that the blood came from the suspect, and ours to indicate probative value for the conclusion of murder. And in doing so, they assume we would use the base rate of murder as the PPG—an assumption that would not reflect other evidence that led police to target the suspect.

FP (2003) similarly offer a misleading example under the heading “The Case of Disputed Murder” (p. 636). They describe murder as uncertain because the husband claims accident. However, they also describe prosecution evidence that in fact she was beaten to death. They then go on to state that we would claim low MPV for evidence of battery under these incriminating evidentiary circumstances. This, like KK’s showcase example, is a straw man. Given evidence that the woman was beaten to death, the PPG would be adjusted upwards before calculating the MPV of battery. It is not analogous to our claim for lack of probative value in the absence of case-specific incriminating evidence.

Generalizations Regarding Low Probative Value and High PPG

As illustrated in our earlier paper, and as noted by our critics, when PPG is very high, probative value again becomes low. This makes sense, as the information that *can* be gained by the addition of evidence is necessarily less. KK (2003) suggest that we would wish to exclude evidence possessing little probative value, but which might reasonably tip a juror over the edge of the threshold of guilt. In fact, we would not, and did not, make such recommendation.

Instead, we are concerned with the relationship between probative value and prejudicial impact. At low levels of PPG *and* low probative value, the potential for prejudicial impact is great. In theory, evidence could actually increase the probability of guilt from 1 to 10% but have such impact as to move the juror to 80 or 90% (which might occur under conditions where a very confident witness has identified the suspect, but where (s)he witnessed the crime under conditions known to yield accuracy rates of 10% or less, for example). The potential for prejudicial impact is much less if existing evidence indicates 90% likelihood of guilt. Hence, lower levels of probative value are acceptable both at higher prior probabilities of guilt and at lower PPG—provided there is no reason to expect undue prejudicial effect.

KK mischaracterize us in suggesting, for example, that given a 98% chance of a defendant’s guilt of sex abuse, we would recommend against admission of evidence of sexually transmitted disease in the child. Granted, our index of probative value

would be greater than zero, but low. However, given low potential for prejudicial impact, we would not recommend against admitting the evidence.

Confusion of Indices of Probative Value With Admissibility

As the previous discussion and KK's (2003) child abuse example suggest, KK apparently assume that our recommendations regarding admissibility would correspond precisely to our index of probative value. Nothing could be farther from the truth. Our index of probative value can be considered to be an index of how much increase in likelihood of guilt is objectively justified by the evidence. This actual utility must be balanced against other considerations specified in Federal Rules of Evidence 402-412, such as prejudicial impact, waste of time, and conflict with specific exclusionary principles, to yield a decision.

Confusion of Indices of Probative Value With Sufficiency

KK (2003) argue that our index of probative value confuses evidentiary support with *sufficiency* of that support to prove guilt. This is certainly not true. Such an assertion would mean essentially that probative value would equal the likelihood of guilt given the evidence. Our definition does not imply this identity at all. Recall, we define *probative value* as the *difference* in likelihood of guilt with versus without the evidence. As long as the likelihood of guilt without the evidence is greater than zero, and there is some difference between the likelihood of guilt with and without the evidence, this index will not be identical to the likelihood of guilt with the evidence.

KK make this mistake by focusing on our method for estimation of *maximum probative value*, which is designed to maximize the difference between the two probabilities. This is done by maximizing $p(G|E)$ and minimizing $p(G|E-)$. When the latter is zero (which can be true only as long as the base rate of the evidence exceeds that of the base rate of guilt), the index of probative value is equal to the sufficiency of the evidence. Under all other conditions, it is not. There is no reason to suggest that the two must never converge.

Summary

The persistent misconstruals of our position detract from more reasoned consideration of the principles that should guide construction of indices of evidence utility. Our selection of a relatively extreme set of hypotheticals involving little evidence in addition to motive led to correspondingly extreme conclusions regarding the lack of probative value of motive evidence. This relatively extreme example has led to unfortunate overgeneralization of our assertions regarding limitations in probative value.

Although, unlike our critics, we believe cases do go to trial on the basis of very limited evidence, which may have little to no probative value alone (e.g., a single witness who had very poor opportunity to observe), we agree with their assertions that more commonly other evidence exists to create a higher PPG. Hence, the focus on the extreme case has tended to obscure significant areas of agreement between the participants in this exchange.

TO WHAT DO WE ALL AGREE?

The Relevance of Statistical Assessments of the Utility of Evidence

There has been long-standing and general agreement between scientific and legal scholars that the utility of information may be assessed via probabilistic analyses. Such analyses are commonplace among medical scholars, for example, to assess the utility of medical tests in populations with varying rates of disease (e.g., Sox, Blatt, Higgins, & Marton, 1988). Within the legal community, probabilistic analyses have been offered to support general theories of the proper (rational or objective) use of evidence (e.g., Friedman, 1986; Garbolino, 2001; Kaye, 1986; Lempert, 1977, 2001; MacCrimmon & Tillers, 2002; Posner, 1999; Reagan, 2000), including analyses of probative value such as those central to this exchange, as well as rational rules for combination of evidence.

In addition, a large body of literature, particularly among psychologists, has addressed the actual utility of specific forms of evidence. Koehler and Kaye have participated in such debates surrounding the utility of DNA evidence (e.g., Kaye, 1993, 1995; Koehler, 1996), polygraph results (e.g., Kaye, 1987), and clinical symptoms of sex abuse (e.g., Lyon & Koehler, 1996). Friedman (1991) and Park (1996, 1998) have offered statistical arguments regarding the implications of character evidence and the relevance of prior acts. Wells has offered such analyses with respect to the “information gain” provided by eyewitness identifications (Wells & Olson, 2002), and others have discussed probability-based analyses of clinical assessments ranging from Rorschach results to predictions of dangerousness (see reviews in Litwack & Schlesinger, 1999; Melton, Petrila, Poythress, & Slobogin, 1997).

To our knowledge, our paper (Davis & Follette, 2002) was the first to apply probabilistic analyses involving base rates to motive evidence, although others have provided such analyses with respect to explicit rather than “intuitive” profiling evidence (e.g., Yin, 2000). Nonetheless, all parties appear to agree upon the appropriateness of such an analysis, notwithstanding disagreement regarding such issues as selection of base rates, and the specific index of probative value.

The Contingent Utility of Evidence

Just as medical scholars have shown the informational properties of medical tests to depend upon the base rate of the target disease within the relevant population (e.g., Sox et al., 1988), all parties to this exchange agree that the informational properties associated with evidence are a function of the PPG. Our own analysis (Davis & Follette, 2002) and Friedman’s earlier analyses (1986, 1994) suggest that both probative value and posterior odds of guilt are dependent upon base rates, whereas Bayesian LR analyses suggest that posterior odds, but not probative value, will depend upon base rates.

Notwithstanding these differences, both positions clearly show that the degree to which evidence will change the likelihood of a proposition such as guilt (which, in our view, reflects the utility of the evidence) is dependent upon prior probability of the proposition. Using the Bayesian LR, Wells (2003) has illustrated nicely how the

“information gain” associated with evidence of a particular probative value depends upon the prior odds of guilt (essentially the concept of base rate of guilt without the evidence). Similarly, our previous paper (Davis & Follette, 2002) illustrated the dependence of probative value (defined in terms of differences in probability of guilt with and without the evidence) upon base rates. Just as the utility of medical tests is dependent upon the base rate of the criterion disease in the population tested (e.g., Gigerenzer, 2000), so is the utility of evidence dependent upon the base rate of guilt in the relevant population (or the prior odds of guilt given other evidence).

Legal scholars have also recognized and debated the concept of *conditional relevance* (e.g., Ball, 1980; Crump, 1997; Nance, 1990) or *conditional probative value* (e.g., Friedman, 1994). In fact, Friedman (1994) offered a conceptualization of “conditional probative value” (defined as the absolute difference in conditional probabilities rather than in terms of the Bayesian LR) that captures the degree to which the information gain resulting from one piece of evidence is dependent upon the nature of other available evidence. For example, evidence of one person’s warning to another may be relevant if the other heard it, but not if it went unheard.

Together, these two bodies of literature have clearly articulated the contingent utility of evidence. That is, the degree to which a specific item of evidence will support an increase in probability of guilt (or in the probability of another proposition relevant to the determination of guilt) will depend upon the nature of other evidence—whether the overall prior odds of guilt, or the existence of specific other items of evidence. All parties to this exchange agree to this principle. Our differences lie in whether *probative value* should be defined in terms of an information gain index of utility versus a context-free LR.

Proper Identification of Relevant Probabilities

Clearly, any analysis of the utility of evidence is dependent upon accuracy in identification/estimation of relevant probabilities. We included a section on the importance of this issue in our original paper, and much of the discussion among our critics in this exchange has focused upon the implications of differing base rate assumptions. We examine implications of these differences in other sections of this paper. However, it is important to note that two levels of accuracy are of interest.

The first level involves accuracy in what specific base rates we deem relevant (i.e., in determination of the relevant population). In the context of this exchange, for example, much is made of the issue of the appropriate base rate (or prior odds) of murder. Clearly, as FP (2003) so eloquently illustrate, the information gain resulting from evidence of infidelity will vary with the chosen base rates. Although selection of relevant base rates that will be generally accepted is challenging, it is our hope that this initial dialogue will serve to apprise others of the importance of the issue, and to stimulate further efforts to find objective base rates to facilitate empirically based evidentiary rulings.

The second level of accuracy involves accurate determination of the actual relevant probabilities. That is, once we have chosen the base rates of interest, we must then estimate them with reasonable accuracy. Our method of estimation of the MPV of evidence requires only base rates—although conditional probabilities would be

necessary to calculate actual probative value. However, the Bayesian LR requires accurate estimation of the probability of evidence given guilt versus given innocence. In practice, these conditional probabilities are rarely available—and unlikely to become available—for nonscientific evidence, as discussed in the earlier section on presumption of innocence. For example, police *may* have such contingent probabilities for those brought to trial or convicted, but will not have data for those who commit unsolved crimes in the same category (making any data on $p(E|G)$ inherently inaccurate). Hence, those who seek to estimate such probabilities must rely on intuitive assumptions—the very practice we are seeking to avoid.

Information Is Often Used Incorrectly (Improperly Weighed) in Practice

Although not specifically addressed by our critics in this exchange, all parties would agree that evidence may be perceived improperly by both judges and jurors. The judge may view evidence as more probative than warranted, whereas jurors may overadjust perceptions of guilt in response to it. In the latter situation, the evidence would be deemed to have prejudicial impact that may or may not outweigh its probative value. We have argued (Davis & Follette, 2002) that such mistakes are particularly likely to occur in situations where the base rate of the evidence is high and the base rate of guilt is low; and when the evidence in question fits intuitive assumptions regarding the type of person, cause, or other circumstance that tends to be associated with the crime in question.

Indeed, motive evidence fitting popular stereotypes concerning the causes of crime (Vanous & Davis, 2001, 2002) can exert powerful effects. We conducted a mock jury study in which the presence of motive evidence was varied (Davis, 2002; Vanous, 2002). An uxoricide case loosely resembling the Franklin case was constructed, such that forensic evidence was not useful for distinguishing between murder and accident. In the control condition the couple's marital unhappiness was depicted as the motive, whereas in the experimental condition the defendant was also depicted as having been involved in an affair at the time of the murder. Evidence of infidelity increased the proportion of guilty verdicts from 32 to 90%—an increase of fully 58% of jurors.

Although likely to disagree regarding the proper impact of motive evidence, all parties to this exchange would likely agree (1) that the way in which judges and jurors perceive and use evidence is unlikely to reflect the curves of probative value and posterior odds indicated by probabilistic analyses, and (2) that testimony regarding probabilistic analyses are potentially useful to judges and juries to inform them of the objective utility of evidence. Differences in opinion may exist, however, regarding what information should be given to judges and/or juries and how it should be expressed, as well as how that information should inform evidentiary rulings or verdicts. We offer our suggestions in later sections.

Understanding of Probabilistic Analyses Is Poor Among Lay and Professional Thinkers Alike

Although not explicitly addressed in their contributions to this exchange, the participants are aware of the poor understanding of statistical information among lay

jurors and professionals alike. This difficulty is reflected both in solving problems on the basis of statistical premises involving probabilities and frequencies oneself and in understanding of others' presentations of statistical analyses of similar problems. Gigerenzer (2000), for example, presents a series of studies of the ability of both professionals and laypersons to solve problems regarding implications of evidence (e.g., medical test results) presented in terms of probabilities versus frequencies. Generally, participants in the studies of Gigerenzer (2000) and others tend to overvalue the implications of evidence such as medical test results, particularly for low base rate diseases, by concluding that a positive test result is much more indicative of disease than warranted (i.e., to overestimate positive predictive value, or PPV). These studies show that performance on such problems is poor when presented in either probability or frequency format, but performance on probability-based problems is far inferior to performance on the same problem presented in frequencies.

A number of authors have documented failures among jurors and other laypersons to understand expert presentations of statistical concepts (e.g., Faigman & Baglioni, 1988; Smith, Penrod, Otto, & Park, 1996). Koehler (2001), for example, has shown that probability-based presentations regarding the implications of DNA match evidence are widely misunderstood, and that the degree of misunderstanding is a function of the manner in which the statistical information is presented. In contrast to Gigerenzer's findings regarding solving problems, Koehler found that results *presented* in terms of probability were better understood than those presented in terms of frequencies.

Given this undisputed difficulty in understanding statistical information, we may safely assume that all will agree that a prime consideration for those who wish to promote the use of objective indices of evidence utility is clarity in presentation of our analyses. Judges and jurors *must* understand our analyses for them to be useful.

Summary

Clearly, there is substantial agreement between the parties to this exchange concerning the value of statistical assessments of evidence utility, the contingent utility of evidence, and the difficulties of applying statistical reasoning in practice. Building upon these fundamental areas of agreement, we present an argument for how indices of evidence utility should be constructed and presented to judges and juries. Following this, we turn to remaining significant areas of disagreement between participants in this exchange—as they would apply to our proposal.

HOW SHOULD WE CONSTRUCT AND PRESENT INDICES OF EVIDENCE UTILITY?

Reliance on our knowledge of how humans solve problems seems very desirable, if not indispensable, in the formulation of rules of evidence.

—Callen (1996, p. 779)

To solve the problem of whether to admit evidence, the judge must first decide upon the relevance of the evidence, and then balance that degree of relevance

against his or her evaluation of competing exclusionary concerns such as prejudicial impact, time required to present and rebut the evidence, and conflict of the evidence with specific exclusionary rules. Juries, in turn, must weigh the importance of newly admitted evidence in light of existing evidence, to arrive at their verdicts.

We have argued (Davis & Follette, 2002) that susceptibility to the *representativeness heuristic* (Kahneman & Tversky, 1973) renders both judges and juries likely to err in these judgments in response to salient causal stereotypes, “intuitive profiles,” and other sources of bias in perceived associations between specific behaviors, characteristics, or other evidence and the likelihood of commission of the crime (guilt). Specifically, they will tend to overvalue evidence intuitively associated with the crime, and to undervalue evidence intuitively unassociated with the crime. Further, as suggested by the “Story Model” (Pennington & Hastie, 1993) of jury decision making, juries (and, in reality, judges as well) are affected by the extent to which evidence fits into a coherent story to explain what happened. Nesson (1985) pointed to the impact of this kind of narrative coherence to explain how a conjunction of items of evidence seems to imply guilt (or innocence) in excess of what would be logically implied by the combined separate items. Clearly, judges and jurors rely on particular categories of evidence (prominently including motive) to construct such stories, and thereby may give such story-relevant evidence excessive weight in proportion to its true utility. Hence, it is our position that accurate evidentiary ruling and juror use of evidence can be facilitated by the presentation of empirically based indices of evidence utility that can offset such biasing assumptions; and that can be presented in accessible format.

Given these concerns, we argue the following.

- (1) Absolute difference (AD) measures of probative value are superior to likelihood ratio (LR) measures.
- (2) Indices of probative value should be presented in the format of probative value curves reflecting values at different points of PPG, rather than single values, along with other graphical and numerical depictions of information gain, potential for prejudicial impact, likelihood of incorrect judgments based on the evidence, and others, as described in later sections.
- (3) Probative value should be calculated using objective rather than subjective values.

In subsequent sections, we consider these recommendations in light of areas of disagreement with the other contributors to this exchange.

WHAT ARE THE SIGNIFICANT AREAS OF DISAGREEMENT?

Disagreement between the parties to this exchange pertains largely to five broad areas: (1) what measure of probative value is appropriate and the related issue of (2) how measures should be presented, (3) criteria for selection of pertinent probabilities (including for the particular Franklin case example), (4) the particular status of motive and other “intuitive profiling” evidence, and (5) implications for legal procedure. Hidden among these explicit areas of disagreement are implicit differences that

influence them, including (a) emphasis on practical versus theoretical/mathematical concerns, (b) concern regarding the relationship between *actual utility* of evidence versus *actual use* by judges and jurors, (c) assumptions regarding independence of evidence, (d) preference for empirical versus subjective establishment of relevant probabilities, (e) the prevalence of circumstances in which (in light of prior probability of guilt) probative value of evidence will be small, (f) the importance of prejudicial value and false positives, (g) the role of the presumption of innocence in assessment of probative value and evidentiary ruling, and (h) the proper status of “predictive” (evidence indicating *greater likelihood* of committing the crime—e.g., character, motive, profiling evidence) versus “trace” (evidence resulting from *actual commission* of the crime—e.g., fingerprints, blood) evidence. Not surprisingly, the authors’ positions on each issue type tend to drive positions on the other. Our value of the presumption of innocence, for example, is reflected in reluctance to select prior odds reflecting an enhanced presumption of guilt to calculate probative value. Likewise, our concern with prejudicial impact is reflected in preference for a measure of probative value that reflects the actual amount of increase in probability of guilt, as opposed to a measure without reference to this specific amount. We consider these and other implicit influences in the context of our discussion of areas of explicit disagreement.

What Measure of Probative Value is Appropriate?

KK (2003) have put forth a strong argument favoring the LR measure of probative value, characterizing it as the commonly accepted measure among evidence scholars. This characterization is clearly false, however. Although Bayesian reasoning concerning combination of evidence is indeed common among evidence scholars and other scientists, LR measures of “probative value” per se are not commonly accepted as “the” appropriate measure. Indeed, the other participants in this exchange, Richard Friedman and Roger Park, both prominent evidence scholars, have themselves published analyses of probative value, using AD measures similar or identical to ours (Friedman, 1986, 1994, 1997; Park, 1996), which in turn have been endorsed by other prominent evidence scholars such as Peter Tillers (e.g., Tillers, 1994) and Dale Nance (e.g., Nance, 1995), among others. Our own analysis of the issue of probative value, as well as discussions of the nature of relevance by these and other prominent evidence scholars, reveals three vital weaknesses in the LR measure of probative value, to which we now turn.

The Interdependence of Evidence

Our original paper (Davis & Follette, 2002) illustrated the differences in probative value and posterior odds when combining correlated versus uncorrelated items of evidence, showing that the increment in posterior odds of guilt when adding an item of evidence becomes smaller as the correlation between the new item of evidence and existing evidence becomes larger. Thinking in terms of prediction via regression equation, when a new item of evidence sharing variance with existing items is entered into an analysis including existing evidence, it may carry less predictive weight than

when entered alone, because variance due to the existing predictors is extracted, and likewise the weight of existing predictors may decrease because of variance shared with the new predictor. Overall, the variance explained may increase very little when the additional item is added. It is also generally true that a set of independent items of evidence will yield greater combined probability of guilt than a set of correlated items of evidence (as also illustrated in our original paper). This point is vitally important, in that no reasonable estimate of either prior or posterior odds can be obtained without regard to the degree of interdependence between items of evidence.

KK (2003, p. XXX) apparently view a context-dependent measure of probative value as inherently defective, and argue that our AD measure of probative value is flawed due in part to its context dependence. They view the implication that context-sensitive measures of probative value would offer different values at different points in the trial (and hence in different evidentiary contexts) as unacceptable.

In contrast, we argue (1) that context dependence is rationally and empirically inevitable, (2) that it is fully compatible with existing treatments of relevance in the law and among evidence scholars (including some involved in this exchange), (3) that the LR as explicated by KK is context-insensitive and thus inappropriate, and (4) that trial procedure might be improved through advance consideration of all evidence rather than the current piecemeal approach to evidentiary ruling.

The context-dependent relevance of evidence is recognized both by the Federal rules of evidence and by prominent evidence scholars. The Federal Rules of Evidence (Rule 104 (b)) define the concept of *conditional relevance* as

Relevancy conditioned on fact. When the relevancy of evidence depends upon the fulfillment of a condition of fact, the court shall admit it upon, or subject to, the introduction of evidence sufficient to support a finding of the fulfillment of the condition.

For example, the DNA pattern or blood type of a defendant may be considered to have no relevance unless other DNA or blood is available from the crime scene or other pertinent sources with which the defendant's DNA or blood can be compared for matching. In such a case, the defendant's DNA and blood sample results may be considered conditionally relevant, subject to the availability of comparison evidence.

The utility of the concept of conditional relevance, as well as the difficulties of the concept as expressed under the existing Federal Rules, has been the subject of considerable discussion and debate among evidence scholars (e.g., Friedman, 1994, 1995; Morgan, 1929; Nance, 1990, 1995; Tillers, 1994). Among the criticisms of the current Rule 104 (b) is the fact that it is stated in absolute terms—i.e., that the evidence is relevant when a particular additional fact is true, but irrelevant when it is not (e.g., Friedman, 1994). Hence, Friedman (1994) proposed the broader concept of *conditional probative value*.

Friedman's concept of conditional probative value refers to differences in *magnitude* of probative value under varying states of existing evidence. In turn, *probative value* is defined in terms of the probability of a proposition (X), given NE (new evidence), OE (old evidence), and RFFK (additional relevant fact-finder knowledge) versus the $p(X)|OE, RFFK$ (i.e., the probability of the proposition without the new evidence). Conditional probative value refers to the difference in probative value of

NE (the new evidence) under varying conditions of OE and RFFK (old evidence, and relevant fact-finder knowledge). Friedman (1994) proposed modifications in the Federal Rules to accommodate this more flexible concept of conditional probative value. Other prominent evidence scholars have endorsed the general intent of Friedman's proposals (e.g., Nance, 1995; Tillers, 1994), in the process affirming the inherently contingent nature of relevance with such statements as *Indeed, the probative value of every piece of evidence is conditional upon other evidence*, (Nance, 1995), or *Try as we might, however, we cannot get rid of the phenomenon of conditional relevance—or, more broadly speaking, the phenomenon of conditional probative value . . . the force of any single inference is practically always contingent upon, or affected by, at least one other inference* (Tillers, 1994, p. 479; see also Nesson, 1985; Scolnicov, 2000). Further, another of our critics, Park (1996), has provided a graphical illustration of the context dependence of the probative value of motive evidence, noting that *As the incident-specific evidence increases, so does the probative value of the motive evidence. At some point the curve turns down because there is so much incident-specific evidence that the motive evidence starts becoming cumulative* (p. 765). Finally, FP (2003) argued strongly for context dependence.

Those who have argued for context dependence of probative value have generally also addressed the issue, using AD measures. KK (2003) instead argue that an appropriate measure of probative value should be context-free, and propose that the LR measure meets this criterion. Instead, we regard a context-free measure as both inaccurate and undesirable, and the LR measure as, in fact, context dependent.

The view of the LR as context independent is based upon the assumption that the LR does not in any way vary along with the prior odds (or base rate). This assumption is unlikely to be reliably true in the applied medical and legal contexts in which the LR is commonly used. The LR will not change simply because the base rate of the disease has changed, or simply because the prior odds of guilt have changed. Instead, the LR may change because the population with the new base rate is different from the population in which the LR was established. For example, suppose that for a particular diagnostic medical test the LR for an entire population (with a base rate of the disease of 1%) indicates that a person is nine times more likely to test positive if (s)he has the disease than if not. Will this ratio remain constant for a subsection of the population in which the base rate of the disease is 50%? We suggest that this assumption will often fail. The performance characteristics of the test (including the LR) may well be a function of factors that vary with the population base rates of disease. That is, any number of factors confounded with incidence of the disease, such as correlated medical conditions, use of substances that interact with the test, genetic variations due to race or susceptibility to the disease, tendency to comply with pretest instructions (such as to fast), or others may affect the performance of the diagnostic test. In such circumstances, the LR will not remain constant across prior odds (i.e., across the varying evidentiary conditions resulting from the differences between populations).

In similar fashion, the LR of $p(E|G)$ versus $p(E|G-)$ may vary with the existence of other evidence suggesting guilt. The LR for the presence of fingerprint match evidence (between defendant prints and those found at the crime scene) given guilt versus innocence will likely vary with whether other evidence suggests the

crime was planned in advance (circumstances in which guilty defendants are likely to avoid leaving prints—thereby suppressing the LR) versus occurred in the heat of passion (circumstances in which the defendant will typically exert less control over fingerprints—thus enhancing the LR), or that the crime was committed in the defendant's home (where prints will be found regardless of guilt) versus another location (where the defendant's presence at the scene may be more dependent upon guilt), among others.

Indeed, even DNA match statistics, commonly regarded as one of the strongest forms of evidence, have been repeatedly subject to legal challenges based upon context dependence of the LR. That is, critics argue that the LR of each match (and hence that of the total match) will be substantially reduced in populations characterized by unusual degree of inbreeding (see a review in Overstall, 1999). Koehler himself (Koehler et al., 1995) has previously acknowledged another form of context dependence in his discussions of misuse of DNA match statistics. That is, he correctly pointed out that the LR of a match given that the target is the source of the DNA versus given that (s)he is not is dependent upon evidence regarding lab error rates. Hence, we are unclear as to why KK (2003) advocate a context-free index of probative value.

Does our measure fully reflect interdependence? Neither of the proposed measures can reasonably reflect the full degree of interdependence between new and existing evidence (except in the extremely rare cases where there are few items of evidence and the needed relationships are known). Our index of MPV does, however, continually maximize $p(E|G)$ as well as $p(G|E)$ as the PPG changes, within the constraints of a constant base rate of E. Hence, it is context-sensitive in a manner that favors admissibility, in that probative value is always maximized, and rises as $p(G)$ rises.

It should also be noted that although, given our concern with prejudicial evidence, we constructed our original measure to represent what might be called *maximum inculpatory probative value*, one can just as easily minimize $p(E|G)$ or $p(G|E)$ to construct a measure of *maximum exculpatory probative value* (probative in the sense of making the conclusion of guilt *less* likely). This would also be context-sensitive, in that $p(E|G)$ and $p(G|E)$ would be constantly minimized instead of maximized.

Our methods, of course, do not constitute full context sensitivity. Further, neither our measure nor the LR can be used with noncategorical data (such as scores on various diagnostic tests, or number of incidents of infidelity, for example). Hence, future probabilistic approaches to evidence might more reasonably employ more complex modeling strategies employing such tools as regression or structural equation modeling. Although such approaches may require more complex assumptions (for hypothetical reasoning about evidence) and more extensive relevant data (for estimates of realistic utility of evidence), they may lead to more accurate results.

The Issue of Understandability Determines Usability

We take it as axiomatic that indices of evidence utility are more likely to be used—and to be used correctly—when they are easier to understand. Two aspects of

the LR compromise its understandability relative to AD measures of PV: (1) lack of correspondence to the wording of the Federal Rules of Evidence and (2) widespread difficulty with Bayesian reasoning.

An index that readily maps onto wording and concepts in which we are naturally inclined to think will be easier to understand—and to accept as valid. The Federal Rules of Evidence specifically define relevance in terms of an AD measure of probative value, i.e., *evidence 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.* (Rule 401). The AD measure of probability of guilt with versus without the evidence ($p(G|E) - p(G|E-)$) fits this wording, whereas the LR in itself does not. The LR is essentially an index reflecting the opposite concept of probability of the evidence given guilt versus innocence ($p(E|G)/p(E|G-)$). Granted, when applied to the prior probability, the LR does adjust the probability of the fact, but this quality is not inherent in the measure itself. KK (2003) dismiss the importance of this noncorrespondence. Although we agree that it may be mathematically unimportant, we believe it will affect the practical application of their LR index.

A large body of work has demonstrated that humans do not naturally use Bayesian reasoning nor do they readily understand presentations involving such reasoning (see Gigerenzer, 2000, for a review). In fact, FP specifically stated that they have avoided presentation of their scenarios in this exchange in LR terms *because some situations are easier to explain with hypothesized frequentist data rather than with, or supplementing, assessments of probability ranging from 0 to 1* (their footnote 6). If given appropriate frequencies, and asked to evaluate relevance under Rule 401, it is likely that an untrained group would use frequencies and percentages to come up with the AD measure more commonly than any measure resembling the LR. Such a result would not favor the measure if its other properties were undesirable. However, in the absence of clear superiority of a different measure on other grounds, it is highly preferable to use a measure compatible with natural understanding of the concept.

In Practice the LR Is Misleading

We believe the LR, when stated without reference to the prior probability, will typically mislead. Americans are confronted almost daily (if not more often) with news such as the following: *Those who eat a diet heavy in vegetables are 300% less likely to get colon cancer!*, *Use of HRT results in a 26% increase in the risk of breast cancer!*, and other statements of the percent or multiples of increase or decrease in risk of disease as a function of various habits, family history, treatments, and so on. However, in the absence of the base rates from which the change occurs this information is inherently next to useless for understanding the personal action implications of the information (i.e., how to apply the information to a single individual). Similarly, the LR has no implications for how many percentage points the evidence should adjust the probability of guilt.

Nevertheless, those exposed to such statistics are prone to overreact. Millions of Americans react to such medical news as if the *personal* impact will be great. Women

stopped HRT in droves in response to the news that HRT results in a 26% increase in incidence of breast cancer. They did not know to ask “26% of what!?” In fact, the actual personal risk of breast cancer rose only from 33/10,000 to 43/10,000 per year (i.e., from 0.0033 to 0.0041; Grady, 2003). Granted, the absolute number of women who get breast cancer from HRT may be large, but the personal risk resulting from HRT is miniscule, because of the low base rates of the disorder per year.

To use KK’s example in their footnote 13 (KK, 2003), if the risk of oral cancer in smokers is 10 in 10,000 but 1 in 10,000 among nonsmokers, the medical news might report the LR *Smokers are 10 times as likely to get oral cancer as nonsmokers!!* How different this would sound, and how different the perceived personal implications if the headline read *Smoking increases the risk of oral cancer from 1 100th of 1 percent to 1 10th of 1 percent*. Hearing this, who would rush out to quit smoking? We find the widespread reporting of such medical findings absent base rates to be uninformative, misleading, and frustrating—and the application of similar odds and percent increase statistics to evidence to be the same.

The way in which evidence increases the likelihood of individual guilt, or of the proposition in a specific instance, is much more analogous to the last news headline. The odds ratio may be huge, but in light of the base rate, the absolute increase in likelihood may be very small. When applied to the specific increase in likelihood of guilt (disease) for a specific individual, only the AD (not the LR) is relevant.

We find it highly unlikely that the general public is aware of the necessity of base rate information to interpret the information they receive on LR ratios, percent, rather than absolute, increase, and so on. We also find it highly likely that use of the LR measure of probative value will commonly result in overestimation of the utility of evidence (i.e., overestimation of the *actual absolute*, rather than relative, information gain resulting from the evidence). In the context of public health, such overestimations may be harmless to the individual, and promote behaviors that in turn promote aggregated beneficial outcomes. At the individual level, however, they may lead to needless suffering (of menopausal symptoms, for example). In the legal context, fact finders may be more often correct than not, but at the cost of an elevated number of false convictions of individuals.

The LR Cannot Be Used to Estimate Maximum PV

Because it is uncommon for actual contingent probabilities to be known, we believe it is desirable for a measure of PV to be able to estimate *maximum PV*, as we did with our measure. The LR is undesirable for this purpose, because maximizing the LR would sometimes require division by zero (if base rates permitted $p(E|G-)$ or $p(E|G)$ to be zero).

How Should Measures of Evidence Utility Be Presented?

As the entirety of the current exchange has made clear, there is no single maximum or actual probative value for specific evidence. Instead, it will vary with the PPG. Further, probative value per se is only one of a set of issues that should be

considered when ruling on admissibility. Of most concern, of course, is the potential for the evidence to lead to an incorrect judgment of either guilt or innocence (Type I and II errors, respectively). Hence, the utility of evidence is best presented in terms that reflect information gain (utility) in the context of potential for prejudicial impact, and errors of each kind.

If we were asked today to offer testimony in the Franklin case (or others), we would not choose to present a single value for probative value on the basis of the assumption of no other evidence. Instead, we would present the following.

- (a) A probative value curve at a constant base rate of the evidence as illustrated in our original paper, and by Wells (2003) to show how maximum probative value will vary across varying PPG. We would not adjust the base rate of the evidence unless clearly justified by data (in which case we would present probative value curves illustrating the impact of this variation as well), and would generally avoid subjective assumptions regarding contingent probabilities and their relationship to the base rate of guilt.
- (b) An information gain curve (as in Well's Fig. 2; well, 2003) to illustrate the amount of increase in probability of guilt justified by the evidence at different levels of PPG.
- (c) Above the information gain curve, we would include a shaded area to indicate the maximum possible prejudicial impact of the evidence. Essentially, this would indicate that at very low PPG, the potential for prejudicial impact is essentially to move the juror all the way from rationally finding innocence even with the evidence to an inappropriate finding of certain guilt. It would also show, of course, that the potential for prejudicial impact from the evidence steadily declines as the PPG escalates.
- (d) As recommended in our original paper, the *odds ratio* of number of incorrect to correct decisions, and the *percent* of incorrect decisions, if the person concluded guilt on the basis of this evidence. Again, unlike our original recommendation, we would present graphical depictions of these values at varying levels of prior probability of guilt.

As noted earlier, these presentations can reflect either maximum *inculpatory* probative value (reflecting increased likelihood of guilt given the evidence) or maximum *exculpatory* probative value (reflecting decreased likelihood of guilt given the evidence). Also, where evidence exists to support actual values, actual probative value rather than MPV can be presented.

This more complete presentation circumvents the problem of incorrect assumptions regarding the state of the evidence underlying selection of base rates or PPG. It allows the judge and/or jury to reason with the evidence in question in light of all other evidence of which they are aware. Although not fully empirically based, and not fully context-sensitive, such a presentation provides a mechanism for apprising judge and jury of the rather nonintuitive curves of evidence utility, cautioning them against inappropriate admission or weighing of evidence, and perhaps dampening the heuristic influences that would otherwise exert greater impact on their judgments.

Criteria for Selection of Pertinent Probabilities

FP (2003) have illustrated the difficulties inherent to identification of fully accurate relevant probabilities in the absence of a complete database incorporating a variety of circumstances and forms of evidence. Just as the unaided judge and jury must reason with the most reasonable assumptions possible, those attempting to construct objective indices of PV must attempt to identify the best values within the limits of available resources. On this, all parties can agree. Differences arise in two notable areas, however: (1) preference for empirical versus subjective probabilities and (2) the implications of the presumption of innocence for adjusting prior odds in the absence of case-specific evidence. We began the paper with discussion of the former issue. Here we focus upon the latter.

Subjective Versus Empirical Probabilities

As reflected in the title of this paper, it is our position that the legal system and the scholars who influence its course should strive to move toward ever more objective empirically based evidentiary rulings. The ultimate goal of the system, of course, should be more accurate determinations, whether of the admissibility (utility) of evidence or ultimately of actual guilt. It is our position that for both determinations empirical estimates of relevant probabilities are superior to subjective estimates.

Bayesian legal scholars have adopted the approach of switching freely between analyses based on actual and subjective probabilities. KK, for example, have employed strictly empirical approaches to analyses where necessary data is available (both in this debate, and in other contexts, such as those involving DNA match statistics (e.g., Kaye, 1995; Koehler, 1996) or the relationship of symptoms to sexual abuse (e.g., Lyon & Koehler, 1996). However, these authors and Bayesian legal scholars in general reason freely on the basis of subjective estimates of prior odds, as well as relevant contingent probabilities. Perhaps more often than not, such subjective estimates will be subject to stereotype/expectation/intuition/heuristic-based sources of error.

Subjective assumptions regarding motives for crime, for example, can be shown to be quite in error. In one study (Vanous, 2002; Vanous & Davis, 2002), we asked a large sample of psychology students and community residents to list the seven most frequent motives for various crimes. For uxoricide, for example, profit was offered by almost twice as many participants (60%) as “he was abusing her” (35%), “occurred during an argument” (34%), or his jealousy (34%), and more than seven times as frequently as she threatens to leave him (8%), all of the latter of which are empirically substantially more strongly associated with uxoricide (e.g., Bixenstine, 1999; Websdale, 1999). “Infidelity” (Unspecified) was mentioned by 73% and husband’s infidelity by 23%. Hence, the actual association of motives with crimes (and specifically with uxoricide) can be seriously misunderstood.

Should Motive or Other Commonly Admitted Evidence Be “Presumed Probative/Admissible”?

Our critics have argued here and elsewhere (FP, 2003; KK, 2003; Park, 1996) that should motive evidence be excluded from trial, the prosecution will be unfairly

hindered, in that the jury will be left with no explanation of why the defendant may have committed the crime, and thus likely to draw the mistaken inference that no motive exists. Our position, in contrast, is that no evidence lacking in substantial probative value when presented alone should be presumed admissible. Indeed, we find the use of predictive evidence to be generally problematic, although full discussion of this issue is beyond the scope of this exchange.

Notwithstanding the importance of motive in the minds of police, judges, and juries, and for the narrative stories they construct to explain the crime, it can and will be far more prejudicial than probative under some circumstances of very low PPG. Similarly, although eyewitness testimony seems intuitively compelling, it can in fact be highly inaccurate under many conditions, but very prejudicial, in that it will enhance juror perceptions of guilt far beyond what is justified by its likely accuracy.

We recognize the controversial nature of this suggestion. However, one way to avoid the false convictions that can result in such situations is to require, in effect, “probable cause” to admit potentially powerful (in terms of impact on jurors) evidence such as motive, eyewitness identifications, contested confessions, and so on. By this we mean significant additional evidence in support of guilt (i.e., at least moderate PPG based upon all other expected evidence). Similar suggestions regarding probable cause have been offered to reduce the likelihood of mistaken identifications (i.e., by requiring other evidence against the defendant before conducting a line up—e.g., Wells, 2002), and the risk of false confessions (i.e., requiring other evidence to provide probable cause for interrogation—e.g., Ofshe & Leo, 1997).

In fact the law does recognize such logic, deeming motive evidence as inappropriate under some circumstances. For example, Park (1996) analyzed the issue of character evidence in the O.J. Simpson trial, noting that the prosecution argued against admission of L.A.P.D. detective Mark Furman’s racist motivations using analogy to cases in which the defendant attempts to blame a third party. In such cases, evidence that the third party has motive may be excluded in the absence of sufficient additional evidence of guilt. In other words, motive is not regarded as probative absent sufficient PPG.

Park then offered an example in which a defendant accused of murdering another man’s wife argues that he was framed by the husband, who was having an affair, and thus presumed to have motive to murder his wife. Park opined that *the judge might reasonably make a conditional probative value ruling that, in the absence of other evidence to support the defense theory, the affair is just not worth going into, despite the fact that it might furnish a motive for murder.* (p. 764). One must question the basis on which the same logic is considered inappropriate when applied to the defendant, or indeed to other evidence lacking in sufficient probative value absent other evidence.

As our previous discussion has clearly shown, the utility of evidence is context-dependent no matter its current subjective role in judge and juror judgments. We should not be bound by current procedural and cognitive constraints in our thinking regarding procedures that can best facilitate a “best evidence” principle. Rules of evidence and trial procedures can and should be modified to take context dependence into account in a way that will further minimize the admission of evidence lacking in sufficient utility to outweigh considerations such as prejudicial impact and waste of time, regardless of how they violate current juror expectations. Such modification

should include something like our probable cause test for all potentially prejudicial information. As our critics suggest, unapprised of such changes, jurors will certainly make inferences regarding motive or its lack. However, just as they become aware that certain kinds of evidence are inadmissible, they can learn that evidence is contingently admissible.

CONCLUSIONS

The application of statistical reasoning to evidence provides the potential to bring greater empirical objectivity to evidentiary ruling, and to juror use of evidence. At present, however, this potential is starkly limited by the extent of relevant available data, as well as by the complexity of the body of evidence. Hence we are forced to limit testimony on such issues to more scientific areas of evidence, and to a limited number of other evidentiary areas where appropriate data is available. The obstacles are indeed formidable. However, we do not believe these obstacles demand that we cease attempts to bring empiricism to evidentiary issues. Instead, we hope to offer the challenge to others, and to fuel interest in others who can contribute productively to the effort.

ACKNOWLEDGMENTS

The authors would like to thank Gary L. Wells for his comments on this manuscript, and Richard Friedman, Roger Park, David Kaye and Jonathan Koehler for furthering the discussion of evidence utility through their contributions to this exchange.

REFERENCES

- Armour, J. D. (1994). Race ipsa loquitur: Of reasonable racists, intelligent Bayesians, and involuntary Negrophobes. *Stanford Law Review*, *46*, 781–790.
- Ball, V. C. (1980). The myth of conditional relevancy. *Georgia Law Review*, *14*, 435.
- Bixenstine, V. E. (1999). Spousal homicide. In H. V. Hall (Ed.), *Lethal violence: A sourcebook of fatal domestic, acquaintance and stranger violence* (pp. 231–257). New York: CRC Press.
- Callen, C. R. (1996). Proving the case: Character and prior acts. *Colorado Law Review*, *67*, 777–787.
- Copeland, J., & Snyder, M. (1995). When counselors confirm: A functional analysis. *Personality and Social Psychology Bulletin*, *21*, 1210–1221.
- Crump, D. (1997). On the uses of irrelevant evidence. *Houston Law Review*, *34*, 1–53.
- Davis, D. (2002, October). *Toward empirical standards for evaluation of the admissibility of evidence*. Paper presented at the Society of Experimental Social Psychology, Columbus, OH.
- Davis, D., & Follette, W. C. (2002). Rethinking probative value of evidence: Base rates, intuitive profiling and the *postdiction* of behavior. *Law and Human Behavior*, *26*, 133–158.
- Faigman, D. L., & Baglioni, A. J. (1988). Bayes' theorem in the trial process: Instructing jurors on the value of statistical evidence. *Law and Human Behavior*, *12*, 1–17.
- Friedman, R. D. (1986). A close look at probative value. *Boston University Law Review*, *66*, 733–759.
- Friedman, R. D. (1991). Character impeachment evidence: Psycho-Bayesian[!] analysis and a proposed overhaul. *UCLA Law Review*, *38*, 655–662.
- Friedman, R. D. (1994). Conditional probative value: Neoclassicism without myth. *Michigan Law Review*, *93*, 439–477.
- Friedman, R. D. (1995). Refining conditional probative value. *Michigan Law Review*, *94*, 457–465.
- Friedman, R. D. (1997). Irrelevance, minimal relevance, and meta-relevance. *Houston Law Review*, *34*, 55–71.

- Friedman, R. D. (2000). A presumption of innocence, not of even odds. *Stanford Law Review*, 52, 873–887.
- Friedman, R. D., & Park, R. C. (2003). Sometimes what everybody thinks they know is true. *Law and Human Behavior*, 27 (6), 629–644.
- Garbolino, P. (2001). Explaining relevance. *Cardozo Law Review*, 22, 1503–1521.
- Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. London: Oxford University Press.
- Grady, D. (2003, June 25). Study finds new risks in hormone therapy. *New York times*.
- Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology*, 80, 557–571.
- Kahneman, D., & Tversky, A. (1973). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Kaye, D. H. (1986). Quantifying probative value. *Boston University Law Review*, 66, 761–766.
- Kaye, D. H. (1987). The polygraph and the PVP. *Statistical Science*, 2, 223–226.
- Kaye, D. H. (1993). DNA evidence: Probability, population genetics, and the courts. *Harvard Journal of Law and Technology*, 7, 101–172.
- 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., & Koehler, J. J. (2003). The misquantification of probative value. *Law and Human Behavior*, 27(6), 645–659.
- Keller, D. (2001). *An almost perfect murder: The saga of Michael Franklin*. McLean, VA: IndyPublish.com.
- Koehler, J. J. (1996). Proving the case: The science of DNA. On conveying the probative value of DNA evidence: Frequencies, likelihood ratios, and error rates. *Colorado Law Review*, 67, 859–886.
- Koehler, J. J. (2001). The psychology of numbers in the courtroom: How to make DNA-MATch statistics seem impressive or insufficient. *Southern California Law Review*, 74, 1275–1305.
- Koehler, J. J., Chia, A., & Lindsey, J. S. (1995). The random match probability in DNA evidence: Irrelevant and prejudicial? *Jurimetrics Journal*, 35, 201.
- Koski, D. D. (2001). Jury decisionmaking in rape trials: A review and empirical assessment. *Criminal Law Bulletin*, 38, 21–159.
- Lempert, R. (1977). Modeling relevance. *Michigan Law Review*, 75, 1021.
- Lempert, R. (2001). The economic analysis of evidence law: Common sense on stilts. *Virginia Law Review*, 87, 1619–1712.
- Litwack, T. R., & Schlesinger, L. B. (1999). Dangerousness risk assessments: Research, legal, and clinical considerations. In A. K. Hess & I. B. Weiner (Eds.), *The handbook of forensic psychology* (pp. 171–217). New York: Wiley.
- 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.
- MacCrimmon, M., & Tillers, P. (Eds.). (2002). *The dynamics of judicial proof: Computation, logic and common sense* (Vol. 94). Heidelberg, Germany: Physica-Springer-Verlag.
- Melton, G. B., Pettila, J., Poythress, N. G., & Slobogin, C. (Eds.). (1997). *Psychological evaluations for the courts: A handbook for mental health professionals and lawyers*. New York: Guilford Press.
- Morgan, E. M. (1929). Functions of judge and jury in the determination of preliminary questions of fact. *Harvard Law Review*, 43, 165–192.
- Nance, D. A. (1990). Conditional relevance reinterpreted. *Boston University Law Review*, 70, 448–507.
- Nance, D. A. (1995). Conditional probative value and the reconstruction of the Federal Rules of Evidence. *Michigan Law Review*, 94, 419–456.
- Nesson, C. (1985). The evidence or the event? On judicial proof and the acceptability of verdicts. *Harvard Law Review*, 98, 1357–1392.
- Ofshe, R. J., & Leo, R. A. (1997). The decision to confess falsely: Rational choice and irrational action. *Denver University Law Review*, 74, 979–1122.
- Overstall, R. (1999). Mystical infallibility: Using probability theorems to sift DNA evidence. *Appeal*, 5, 28–37.
- Park, R. C. (1996). Proving the case: Character and prior acts. Character evidence issues in the O. J. Simpson case. *Colorado Law Review*, 67, 747–776.
- Park, R. C. (1998). Character at the crossroads. *Hastings Law Journal*, 49, 717–779.
- Pennington, N., & Hastie, R. (1993). The story model for jury decision making. In R. Hastie (Ed.), *Inside the juror: The psychology of juror decision making* (pp. 192–221). New York: Cambridge University Press.
- Posner, R. A. (1999). An economic approach to the law of evidence. *Stanford Law Review*, 51, 1477–1546.
- Reagan, R. T. (2000). Supreme court decisions and probability theory: Getting the analysis right. *University of Detroit Mercy Law Review*, 77, 835–873.

- Saks, M., & Kidd, R. (1980–81). Human information processing and adjudication: Trial by heuristics. *Law and Society Review*, *15*, 123–125.
- Sanchirico, C. W. (2001). Character evidence and the object of trial. *Columbia Law Review*, *101*, 1227–1311.
- Scolnicov, A. (2000). On the relevance of “relevance” to the theory of legal factfinding. *Israel Law Review*, *34*, 260–301.
- Smith, B. C., Penrod, S. D., Otto, A. L., & Park, R. C. (1996). Jurors’ use of probabilistic evidence. *Law and Human Behavior*, *20*, 49–82.
- Snyder, M., & Thomsen, C. J. (1988). Interactions between therapists and clients: Hypothesis testing and behavioral confirmation. In D. C. Turk & P. Salovey (Eds.), *Reasoning, inference, and judgment in clinical psychology* (pp. 124–152). New York: Free Press.
- Sox, H. C., Jr., Blatt, M. A., Higgins, M. C., & Marton, K. I. (1988). *Medical decision making*. Boston: Butterworth.
- Tillers, P. (1994). Exaggerated and misleading reports of the death of conditional relevance. *Michigan Law Review*, *93*, 478–484.
- Vanous, S. (2002). *Prejudicial nature of motive evidence*. Unpublished Master’s Thesis, University of Nevada, Reno.
- Vanous, S., & Davis, D. (2001, April). *Motive evidence: Probative or just prejudicial?* Paper presented at the Rocky Mountain Psychological Association, Reno, NV.
- Vanous, S., & Davis, D. (2002, April). *Murder scripts: Perceived motives and means for spouse murder*. Paper presented at the Rocky Mountain Psychological Association, Salt Lake City, UT.
- Websdale, N. (1999). *Understanding domestic homicide*. Boston: Northwestern University Press.
- Wells, G. L. (1992). Naked statistical evidence of liability: Is subjective probability enough? *Journal of Personality and Social Psychology*, *62*, 739–752.
- Wells, G. L. (2003). Murder, extramarital affairs, and the issue of probative value. *Law and Human Behavior*, *27*(6), 623–627.
- Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied*, *8*, 155–167.
- Yin, T. (2000). The probative values and pitfalls of drug courier profiles as probabilistic evidence. *Texas Forum on Civil Liberties and Civil Rights*, *5*, 141–190.