The Problematic Value of Mathematical Models of Evidence

Ronald J. Allen and Michael S. Pardo

ABSTRACT

This paper discusses mathematical modeling of the value of particular items of evidence. We demonstrate that such formal modeling has only limited use in explaining the value of legal evidence, much more limited than those investigators who construct and discuss the models assume, and thus that the conclusions they draw about the value of evidence are unwarranted. This is done through a discussion of several recent examples that attempt to quantify evidence relating to carpet fibers, infidelity, DNA random-match evidence, and character evidence used to impeach a witness. This paper makes the following contributions. Most important, it is another demonstration of the complex relationship between algorithmic tools and legal decision making. Furthermore, at a minimum it highlights the need for both analytical and empirical work to accommodate the reference-class problem and the risk of failing to do so.

1. INTRODUCTION

Legal scholarship that explores the nature of evidence and the process of juridical proof has had a complex relationship with formal modeling. As demonstrated in so many fields of knowledge, theory formation and formal modeling have the potential to increase our comprehension of and our ability to predict and control aspects of complex, ambiguous,

and unruly nature. And when theory formation and formal modeling are applied to the legal system, they perhaps also increase the accuracy of fact finding, which is a tremendously important goal. The hope that knowledge could be formalized within the evidentiary realm generated a spate of papers that attempted to use probability theory to explain aspects of trials (see, for example, Kaplan 1968; Finkelstein and Fairley 1970; Lempert 1977; Friedman 1987; Tillers and Schum 1992). This literature was both insightful and frustrating. It shed much light on the legal system by bringing the tools of probability theory to bear upon it, but it also quickly became evident that the tools were in many ways ill constructed for the task. Fundamental incompatibilities between the structure of legal decision making and the extant formal tools were identified, and a number of the purported explanations of legal phenomena turned out to be internally inconsistent.[1] As a consequence, interest in this type of formal modeling declined, and attention was directed toward different kinds of explanations of the phenomena (Pardo 2005; Allen and Leiter 2001; Allen 1994, 1991; Cohen 1977).

Interestingly, a number of recent papers have attempted to apply mathematical models to quantify the probative value[2] of various items of evidence in ways consistent with the formal features of probability theory and then to study decision making from that perspective (Nance and Morris 2002, 2005; Finkelstein and Levin 2003; Davis and Follette 2002, 2003; Friedman and Park 2003). For example, the value of evidence is taken to be its likelihood ratio, that is, the probability of discovering or receiving the evidence given a hypothesis (for example, the defendant did it) divided by the probability of discovering or receiving the evidence given the negation of the hypothesis (somebody else did it) (Nance and Morris 2005, 2002; Finkelstein and Levin 2003, pp. 268–69; Koehler 2001; Kaye 1995). Or alternatively, the value of evidence is (more contextually) taken to be the information gain it provides, which is defined as the increase in probability it provides for a hypothesis above the probability of the hypothesis based on the other available evidence

1. For example, there are attempts to defend an expected-utility approach to burdens of persuasion with an argument that is valid if, but only if, burdens of persuasion apply to cases as a whole (the defendant is liable or not, guilty or not), but this is false; they apply to individual elements (Allen 2000).

2. Probative value is a relational concept that expresses the strength with which evidence supports an inference to a given conclusion. It is a crucial concept both for determining admissibility (see Fed. R. Evid. 403, which instructs judges to exclude evidence when its probative value is substantially outweighed by its prejudicial, confusing, or duplicative effect) and for determining whether parties have satisfied their burdens of proof.

(Davis and Follette 2003, pp. 668–69; Friedman 1994b).[3] Both conceptions further assume that all of the various probability assessments conform or ought to conform to the dictates of Bayes's theorem, a formal probability theorem that maintains consistency among such assessments (Finkelstein and Levin 2003; Davis and Follette 2003; Nance and Morris 2005, 2002); empirical studies are then done that test the extent to which this is so and propose how the law can increase the probability that it is so (Nance and Morris 2005, 2002).

As with the first wave of interest in the application of probability theory to juridical proof, this recent scholarship is interesting, instructive, and insightful. However, it also suffers from a deep conceptual problem that makes ambiguous the lessons that can be drawn from it— the problem of reference classes. The implications of this problem are considerable. To illustrate the problem, consider the famous blue bus hypothetical. Suppose a witness saw a bus strike a car but cannot recall the color of the bus; assume further that the Blue Company owns 75 percent of the buses in the town and the Red Company owns the remaining 25 percent. The most prevalent view in the legal literature of the probative value of the witness's report is that it would be determined by the ratio of Blue Company buses to Red Company buses, whether this is thought of as or plays the role of a likelihood ratio or determines information gain (including an assessment of a prior probability) (see Finkelstein and Levin 2003, pp. 268–69; Koehler 2001; Kaye 1995).[4] But suppose the Red Company owns 75 percent (and Blue the other 25 percent) of the buses in the county. Now the ratio reverses. And it would do so again if Blue owned 75 percent in the state. Or in the opposite direction: it would reverse if Red owned 75 percent running in the street where the accident occurred (or on that side of the street) and so on. Or maybe the proper reference class has to do with safety standards and protocols for reporting accidents. Each of the reference classes leads to a different inference about which company is more likely liable, and nothing determines the correct class, save one: the very event under discussion, which has a likelihood of one and which we are trying to discover.

Now consider tests of rationality given to decision makers that employ a problem akin to the blue bus hypothetical. To critique the ratio-

3. We include here as well the use of data based on relative frequencies to inform prior probabilities, which is a form of information gain (it changes belief states).

4. As our discussion in the text indicates, the reference-class problem is ubiquitous.

nality of fact finders requires that one compare the answers they give to a correct answer. However, as in so much of life, even if every step one takes is perfectly rational—perfectly consistent with Bayes's theorem, for example—where one starts determines where one comes out. One has to have a correct starting point to critique an endpoint different from one's own, yet often no such objectively correct starting point exists. If experimenters get results different from what they believe to be appropriate, it may reflect on the rationality of the subjects, but it may just as readily be attributable to the influence of different, but equally appropriate, reference classes than those thought to be appropriate by the experimenters. Differences in outcome in such cases cannot readily be construed as reflecting on rationality, which, as we say, makes ambiguous the lessons of this renewed interest in formal modeling within the field of evidence.

We examine the implications of the reference-class problem for recent evidence scholarship that deals with a wide range of topics from carpet fibers (Finkelstein and Levin 2003), to infidelity (Davis and Follette 2003, 2002), to DNA random-match evidence (Nance and Morris 2005, 2002), to character evidence (Friedman 1991), to drug smuggling (*United States v. Shonubi*, 103 F.3d 1085 [2d Cir. 1997]; *United States v. Shonubi*, 895 F. Supp. 460 [E.D.N.Y. 1995]). We also try to demonstrate where the lessons to be drawn from applying formal methods are less ambiguous. This paper thus makes three contributions. First, and most important, it is a further demonstration of the problematic relationship between algorithmic tools and aspects of legal decision making. Second, it points out serious pitfalls to be avoided for analytical or empirical studies of juridical proof. Third, it indicates when algorithmic tools may be more or less useful in the evidentiary process. At the highest level of generality, this paper is another demonstration of the very complex set of relationships involving human knowledge and rationality and the difficulty in attempting to reduce either to a set of formal concepts.

In Section 2, we elaborate on the lessons of the bus hypothetical by contextualizing it within some relevant issues in contemporary epistemology—an area we both, in different ways, have argued provides a better theoretical foundation for the law of evidence than those based on formal models (Pardo 2005; Allen and Leiter 2001).[5] Section 3 then applies the lessons of Section 2 to aspects of juridical proof.

---

5. This does not mean that the two are mutually exclusive. Indeed, mathematical modeling is a subset of epistemology. Any modeling or theoretical discussion that allows

2. EPISTEMOLOGY, EVIDENCE, AND REFERENCE CLASSES: FAKE BARNS AND BLUE BUSES

Evidence law has epistemic aims: to promote true conclusions arrived at via reliable evidence and rational reasoning methods and to prevent false, arbitrary, or irrational conclusions.[6] These aims are, of course, subject to all sorts of competing considerations and goals such as time, money, protecting privacy, promoting relationships, and so on. Nevertheless, evidence law's core epistemic focus suggests that contemporary epistemological theory can illuminate epistemological issues in the law of evidence (Pardo 2005). In this section, we show how one such epistemological issue regarding the concept of reliability can provide conceptual insight into the probative value of evidence—insight that, in Section 3, will aid in showing the limits of attempts to model the value of evidence mathematically.

One primary epistemological project is to explain under what conditions true beliefs qualify as knowledge. One such attempt is through the concept of causation; namely, a true belief or conclusion qualifies as knowledge if some aspect of its truth causes an agent to hold the belief or accept the conclusion. For example, suppose someone drives past a barn under good observation conditions and utters to a passenger, "There's a barn." The utterance is both true and may qualify as knowledge because a real barn, in good observation conditions, caused the agent to utter the statement. The philosopher Alvin Goldman destroyed this simple causal account of knowledge (Goldman 1976; see also Pardo 2005, pp. 347–51).

Take the above example, Goldman (1976) explains, and suppose the agent was in Fake Barn Town. Although the agent observed a real barn, it is one of the few real barns in a town filled with hundreds of barn facades, which, although they look like barns from the front, are just fake barn fronts and not real barns. Even though the agent's conclusion was true and its truth (seeing a real barn) caused the conclusion and even though it was formed by a reliable process[7] (perception under good

---

for epistemically arbitrary decisions is one that should be jettisoned, assuming the goal is accurate decision making. For a development of this argument see Pardo (2005, pp. 359–92).

6. In addition to these core epistemic aims, evidence law also has subsidiary epistemic interests such as rationally persuading parties and the public that correct results have been reached.

7. Reliability may be fleshed out in various ways. But it leads to a generality problem (Pardo 2005, pp. 348–49).

conditions), the conclusion does not qualify as knowledge because while true, it is only accidentally so. The agent does not know he saw a real barn. The agent would have formed the same belief even if he had observed one of the hundreds of fake barns in the town. To qualify as knowledge, the reporter would need to be able to distinguish between relevant counterfactual situations. Because the agent's capacities are not sensitive enough under these circumstances, he is an unreliable reporter of barns in this town.[8]

Now suppose that Fake Barn Town sits within Barn County, in which real barns vastly outnumber the barn facades (Brandom 2000, pp. 112–16; Pardo 2005, pp. 351–59). As if by magic, the agent, who looked unreliable in Fake Barn Town, now appears reliable when we attend to the fact that he is in Barn County. But now suppose further that in Fake Barn State (in which the county is located) barn facades vastly outnumber real ones. The switch flips, and the agent is now an unreliable reporter in the state, and so on. Or let us go in the reverse direction: suppose that the observation took place on Real Barn Street, in which all the barns are real. He is a reliable reporter on that street, and so forth.

Here is the critical point. The event under consideration (the observation of the barn) is a member of an infinite number of reference classes, the boundary conditions of which can be gerrymandered in countless ways, some of which lead to the inference that the agent is reliable and some to the inference that he is unreliable, given that particular class. And—outside of the reference class consisting only of the event itself— nothing in the natural world privileges or picks out one of the classes as the right one; rather, our interests in the various inferences they generate pick out certain classes as more or less relevant.[9] To see the bite

8. The conclusion of this artificial example has significant real-world consequences. As persuasive psychological research suggests, testing procedures and conditions at lineups, show ups, and photographic arrays may so affect a witness's choices that even accurate identifications (as confirmed after the fact) should be discarded if the witness would have made the same choices regardless of accuracy. For an example of the psychological literature, see Wells, Olson, and Charman (2003).

9. The problem cannot be solved by picking the smallest reference class either (Pardo 2005, p. 376; Colyvan, Regan, and Ferson 2001, p. 172). What matters is homogeneity within the class. How to specify the appropriate reference class for determining hypothesis confirmation has been a prominent issue in the philosophy of science. Hans Reichenbach (1949, p. 374) suggested we choose the smallest class for which reliable statistics were available; Wesley Salmon, by contrast, advocated that for single cases we ought to select the broadest homogeneous class. For a discussion of these positions see Salmon (1967, pp. 91, 124).

of this point, and in particular its bite for juridical purposes, suppose an empirical test were being run as to the ability of our agent (a witness at trial, for example) to identify barns accurately. What is the "proper" baseline (base rate) for running such a test? Is it the proportion of true barns on Real Barn Street, Fake Barn Town, Barn County, or Fake Barn State (or maybe the United Barn States of America)? There is no a priori correct answer; it depends on the interests at stake.

The probative value of juridical evidence is structurally similar to reliability in the above example (Pardo 2005, pp. 374–83). Instead of being natural facts consigned to predetermined reference classes with labels attached to designate the proper class (see Allen 1994), the evidence and the events on which it is based are members of an infinite number of reference classes, which lead to various inferences of various strength depending on how the boundary conditions of those classes are specified.

The blue bus hypothetical with which we began this paper exemplifies the general implications of reference classes, and those implications would hold for practically any attempt to quantify a priori the probative value of evidence.[10] Consider another, and more realistic, example—that of an eyewitness identification made at a lineup. Any attempt to quantify the likelihood ratio of this evidence (the probability of picking the defendant given that he or she did it divided by the probability of picking him or her given that somebody else did it) quickly runs into the reference-class problem. Do we take the ratios of all identifications ever made? Those made (or not made, depending on the circumstance) across racial differences? Those made by this witness? Those made by this witness under similar lighting conditions? Those made on the same day of the week, or month, or year, and so on? In each case, the reference class will likely change, and hence the quantified value will as well. But the evidence, the identification, has not changed. Thus it has no fixed, privileged, quantified value—save the event itself, which has a value of one or zero.

The demonstration above reveals several points. First, the value of evidence is not its likelihood ratio given a certain specified reference class. Evidence has countless likelihood ratios corresponding to its various reference classes. An explanation or justification for choosing any particular one must be provided, and there will invariably be reasonable

10. The reference-class problem may be universal, but we do not need to establish that.

alternatives. Second, for the same reason, the value of evidence is not, alternatively, its information gain in a given context, namely, the increase in probability of a hypothesis (for example, the defendant did it) from the prior probability without the evidence. This view still requires a likelihood-ratio calculation based on a chosen reference class; it just combines that likelihood with the prior probability.[11] Third, instead of capturing the probative value of evidence, the various statistics or likelihood ratios flowing from various reference classes are just more evidence and, as such, must themselves be interpreted and explained.[12] In Section 3 we apply these lessons to various aspects of juridical proof.

3. EPISTEMOLOGICAL LIMITATIONS OF MATHEMATICAL MODELS OF EVIDENCE

Questions at trial often focus on what happened specifically at a certain moment of time. Rarely is the ultimate issue a relative frequency about a class of events (disparate-impact issues in discrimination cases being a possible exception). The reference-class problem demonstrates that objective probabilities based on a particular class of which an item of evidence is a member cannot typically (and maybe never) capture the probative value of that evidence for establishing facts relating to a specific event. The only class that would accurately capture the "objective" value would be the event itself, which would have a probability of one or zero, respectively.

Any attempt to mathematically model the value of evidence, however, must somehow try to isolate an item of evidence's probability for establishing a particular conclusion. Generating these probabilities will, in turn, involve isolating characteristics about the evidence, the event, and the relationship among those characteristics. This relationship may be established either by objectively known base rates or through subjective assessments. In either case, the modeled values arise through abstracting from the specific evidence and event under discussion and placing various aspects of each within particular classes, with varying frequencies, propensities, or subjective probabilities instantiated by the various characteristics on which one has chosen to attend (for example,

11. Or, again, it simply informs the formation of a prior probability.

12. The value of evidence, whether the original propositions or their likelihood ratios, is the strength it provides a particular inference in a particular context (Pardo 2005, pp. 374–83), and this strength will be determined by the plausibility of alternative, contrary inferences (Allen 1994).

the frequency with which defendants who exhibit characteristic X com-
mit crime Y). An important lesson of the fake barn and blue bus ex-
amples in Section 2, however, was that adjustments in the boundary
conditions of the relevant classes may alter the strength of the inference
from the evidence to the conclusion that the event instantiates the char-
acteristic for which the evidence is offered (for example, whether this
defendant committed crime Y).[13]

The reference-class problem, in other words, is an epistemological
limitation on attempts to establish the probative value of particular items
of legal evidence (Pardo 2005, pp. 374–83). It is an epistemological
limitation because different classes may point in opposite directions and
nothing, other than the event itself, necessarily privileges one over an-
other. To be sure, some will be better or worse than others because some
will provide better or worse information about what we are trying to
infer regarding the underlying event. But the question of which is which
will, like any other evidence, be the subject of argument and, ultimately,
judgment. These conceptual points place significant limitations on at-
tempts to mathematically model the value of legal evidence.

We first list these limitations generally and then illustrate them with
specific examples. First, and most important, the probative value of legal
evidence cannot be equated with the probabilities flowing from any given
reference class for which base-rate data are available. Related to this
point, probative value likewise cannot be equated with the difference
between prior and posterior probabilities on the basis of such data, nor
is it sensible simply to translate directly an available statistic into a prior
probability. Second, the above problem regarding establishing probative
value cannot be solved by merely specifying the relevant classes with
more detailed, complex, or "realistic" characteristics. Third, while
switching from objective to subjective probability assessments better ac-
commodates unstable probative values of evidence, it nevertheless still
illustrates the pervasiveness of the reference-class problem because of its
presence even when evaluating such subjective assessments. Finally, the

---

13. Even when the material proposition for which evidence is offered itself involves a
frequency, these reference-class issues still arise. For example, consider evidence establishing
a racially disparate hiring practice or epidemiological studies establishing increased disease
among those who took a particular drug. In the typical case, this evidence is being used
to establish that a particular plaintiff was discriminated against or injured, and the issues
discussed in this paper will arise as to what larger class is appropriate to compare with
the evidence in the case. In addition, the ratios in the evidence itself will raise issues
regarding, for example, error rates and fraud, which makes the value of this evidence for
the proposition for which it is offered, not the overt statistic.

reference-class problem is so pervasive that it arises whenever one assesses the probative value of evidence, even when one is not trying to fix a specific numeric value to particular items of evidence—for example, when assessing whether evidence satisfies a standard of proof.

We next illustrate these general lessons by discussing several examples of attempts to model the probative value of evidence. Our point in critiquing these models is not to criticize these models in particular. To the contrary, we think they are quite useful in helping to understand the nature of legal evidence. Our point is to show that the epistemological limitations discussed above adhere in any such attempts. Thus, our criticisms concern primarily the conclusions that may—and may not—be inferred from such models. The limitations that are discussed, we contend, undermine the strong conclusions that are drawn from such models.

3.1. Modeling Objective Probative Value

We illustrate the first two lessons above by discussing three attempts to mathematically model the objective probative value of legal evidence. All three equate probative value either with likelihood ratios from base-rate data or with the difference between prior and posterior probabilities based on such data. The conclusions drawn from all three are undermined by the reference-class problem.

3.1.1. Carpet Fibers. Finkelstein and Levin (2003, p. 266) attempt to model the probative value of a found carpet fiber. They present and analyze two variations on the following scenario: "A crime has been committed and an unusual carpet fiber is found at the scene. Based on manufacturing records, an expert testifies that the frequency of such fibers in carpets is less than 1 in 500."

A match is found among carpet fibers taken from a suspect, a neighbor named Jones. The authors analyze the probative value of the match assuming the police tested 20 samples from Jones's various carpets and assuming the police tested one sample from Jones and one sample from 19 other suspects.

With regard to the first scenario, Finkelstein and Levin (2003, p. 266) first argue that it would be inaccurate for a prosecutor to argue "that there is only one chance in five hundred of such a match if the crime-scene fiber had come from some place other than Jones's carpets." The reason the prosecutor's argument would be inaccurate is that, because 20 samples were tested and each could have been a match, the probability

of one of them matching is much higher than if only one sample fiber had been tested (pp. 266–67). The authors invoke Bonferroni's inequality theorem—when there are multiple samples tested, the probability of at least one matching is less than or equal to the sum of the individual probabilities of each sample matching—and conclude, "So our prosecutor could only say that the probability of seeing a match if the crime fiber came from another source was less than 20 × 1/500, or 1 in 25, not 1 in 500," and thus "the search among the fibers of the suspect's carpets significantly reduced the probative value of what was found" (pp. 266–67).[14]

With regard to the second scenario, Finkelstein and Levin (2003, p. 267) conclude that the prosecutor could indeed argue "that there is only one chance in five hundred of finding such a match if the fiber did not come from Jones's carpet" because this number captures the "probative value" or "probative force" or "probative effect" of the fiber evidence. (They use all three terms, apparently interchangeably; see pp. 267, 268, and 269 n.8.) They use this value to calculate the likelihood ratio for the evidence in order to combine it (via Bayes's theorem) with the prior odds of guilt given the other evidence: "The likelihood ratio is thus $1/(1/500) = 500$. Given the match, the odds that the fiber came from Jones's home (versus a different home) are 500 times greater than the odds would have been ignoring the match evidence. This is undoubtedly powerful evidence" (p. 269).

It may very well be powerful evidence, but, as with the fake barn example, there is a reference-class issue, and that issue vitiates the showing that Finkelstein and Levin are trying to make. The probative value (or force or effect) of the evidence (the fiber found at the scene and its match to Jones) is not the above-quoted likelihood ratio in the second hypothetical, and it has no obvious application to the first hypothetical either. The problem with both conclusions arises from an ambiguity in the sentence, "Based on manufacturing records, an expert testifies that the frequency of such fibers in carpets is less than 1 in 500." What does

14. This is not, however, necessarily true. If the fibers are identical, or are from the same source (a uniform rug throughout the house), multiple testing will not change the probability of a match. It is thus possible that 20 tests are equivalent to one; whether this is true or false is obviously an empirical question and cannot be resolved analytically. For that matter, why should samples of the suspect's carpeting be viewed as random samples from a population? Is there some reason to think he bought 20 different carpets randomly across the United States (or is the world the right reference class, or the local Carpet Town USA)?

this mean? Whose records? Which records? Does the statistic refer to those who make a particular kind of carpet, or all U.S. manufacturers, or all manufacturers in the world? Or all carpets ever made in the history of the world to date? And once we know the class to which it applies, why is this the appropriate class in which to place Jones and his carpet sample? Is the fiber more or less prevalent in his part of the world, country, state, region, age group, gender, profession, socioeconomic class, and so on? Each of the different classes suggested by these questions would reveal different probabilities and likelihood ratios, but the evidence under consideration has not changed. Indeed the evidence would likely have widely varying likelihood ratios. The probative value of the evidence cannot be simply the ratio derived from any arbitrarily chosen reference class. Therefore, to argue, as Finkelstein and Levin do, that the probative value is the likelihood ratio of 500, they would first need an argument that the appropriate class has been employed, one that is not obvious given the paucity of information in the example.

A second problem with their conclusions concerns how the fiber evidence connects with other evidence. They contend, "The presence of other evidence does not change the analysis given above because the increase in probability that Jones's house was the source of the fiber associated with finding a match is not affected by the other evidence; it is only the degree of probability based on the other evidence and the fiber matching that is so affected" (p. 268).[15] Regardless of whether the probability—that Jones's house was the source of the fiber, given the match—is not affected by other evidence, the probative value of the evidence (a matching fiber) is so affected by other evidence. For example,

15. Likewise, Kaye and Koehler (2003, p. 651) assert that a likelihood ratio "is a constant—it does not change according to one's prior belief." While there is a sense in which it may be true that a likelihood ratio does not vary on the basis of prior beliefs (although we are not sure how one could calculate such a ratio without relying on prior beliefs), it is false that likelihood ratios are constants. Constants may exist in some scientific areas, but in most juridical contexts they can be formed only by exercises of judgment, including picking appropriate reference classes based on ill-quantified data. Thus, instead of being constants, they will typically be contestable propositions. Perhaps Kaye and Koehler meant to limit their remark to certain narrow aspects of scientific evidence, but even then there will be a reference-class issue, as our discussion of DNA evidence demonstrates. Alternatively, perhaps they meant that the processes of forming prior beliefs and forming likelihood ratios are hermetically sealed off from one another, but our discussion shows this to be in error. One's investigations of the scene of the crime, coupled with increasing knowledge of the carpet industry as it bears on the crime, plainly could affect one's construction of a likelihood ratio. Obviously, forming a likelihood ratio depends on some set of prior beliefs, and there is no reason why that set and whatever forms a prior probability must be completely distinct from each other.

conclusive evidence that the crime scene fiber had been planted after the fact to frame Jones would reduce the value of the fiber evidence to zero. Even if we have no evidence about this possibility, how do we know that it was brought from the suspect's home? Even if it was, how do we know that it was from carpeting in his home rather than, say, from having been picked up on the shoes of the actual perpetrator when he was at a party at the home of the person wrongly accused of the crime? These possibilities further show the disjunct between the value of evidence, on one hand, and the likelihood ratio calculated on the basis of a specified reference class, on the other. And the divergence between the two shows the mismatch that can occur when the former is modeled mathematically on the basis of an arbitrarily chosen reference class.

The references classes employed by Finkelstein and Levin are by no means more plausible than many one can imagine. The problem, however, is that there may be no data for other plausible reference classes, which means that the mathematics can be done only by picking these or some variant. Thus, instead of showing something true about the events under consideration, they merely show one of an enormous number of calculations that might be done if one had different data. Using the data one has does not make the proffered analysis correct or true in some sense; instead, it is reminiscent of relying on the lamppost more for support than illumination. Instead of being an objective datum that captures the value of evidence, the numbers that Finkelstein and Levin discuss are just more evidence, which itself must be interpreted to assess its value in the particular case.

3.1.2. Infidelity. Davis and Follette (2002, p. 156) also purport to demonstrate that "the probative value of evidence can be mathematically/ empirically established" if one knows the "pertinent base rates or contingent probabilities or both in the appropriate population." To illustrate this they consider the value of a defendant's infidelity in the murder of his wife: "The fact of infidelity is not probative of whether a man murdered or will murder his wife. In fact, the relative increase in likelihood that an unfaithful man will murder his wife, over the likelihood that a faithful man will murder his wife is so infinitesimal (.0923%) as to be totally insignificant" (p. 139). They arrive at this calculation by defining probative value as "the difference between the probability of murder given the infidelity and the probability of murder given no infidelity" (p. 137). In order to determine this, something like a likelihood ratio (in that it compares two relative frequencies) was obtained by dividing

the rate of murdered wives per million men by the rate of infidelity. This figure is used to determine what they call a "maximum probative value" of infidelity on the assumptions that faithful husbands never murder their spouses and that at most it is only .0923 percent more likely that an unfaithful husband will murder his wife during their marriage than a faithful husband will (p. 137). Davis and Follette then assert that those who assign greater value to the evidence are engaged in inaccurate "intuitive profiling," which simply means giving too much credence to the belief that an unfaithful husband more likely fits the profile of one who would murder his wife than a faithful one and thus is more likely to have done so (pp. 152–54). The correct probative value, under their analysis, is the value arrived at above via the chosen base rates.

As with the carpet fiber example, the calculations work only once a particular event is placed within a particular reference class, and there is no reason to privilege the particular base rates Davis and Follette employ. With any given suspect, different rates exist that would vary from being highly probative to being virtually irrelevant, depending on such variables as age, geographic location, types or amount of infidelities, types of murder, and so on. The evidence (infidelity) would remain the same as we vary the reference class, but the value of the evidence would change.

Davis and Follette recognize this general issue. They discuss the simplified example in order to discuss a real case (p. 135; for a discussion of the case, see Friedman and Park 2003, pp. 640–43). According to their description of the facts, a woman and her husband were riding on a snowmobile (the woman driving and her husband on the back) when the woman drowned as a result of a crash (pp. 135–36). The prosecution theorized that "the defendant had deliberately drowned his wife once they had fallen into a ditch, and that he may have somehow caused the crash, thereafter faking his own unconsciousness/inability to breathe [when paramedics arrived]. Physical evidence of each of these assertions was extraordinarily weak, particularly evidence of whether the wife's death was the result of murder or accident" (pp. 135–36). The prosecution relied heavily on motive evidence: the defendant had purchased a large life insurance policy on his wife within the past year and had had several extramarital affairs. Recognizing the reference-class issue, the authors contend that even more favorable base rates were available to support the defendant given his characteristics (white, in his 30s, middle class) than those mentioned above (about four in 1 million) (pp.

149–50).[16] Nevertheless, they relied on the base rates for all married men in order to pick a class most favorable to the prosecution. The court refused to allow the base-rate evidence, concluding it would be misleading or prejudicial. The authors, however, contend that such testimony should be allowed "to help the jury understand that the intuitive profile or stereotype telling them the evidence is probative of guilt is misleading" (p. 153).[17]

As with the simplified example above, this particular defendant is part of a large number of classes, each with its own base rate, not just the general ones such as married, unfaithful, white, 30s, middle class, and so on. Indeed, if one starts specifying more and more details, one will arrive at the event itself, which will have a base rate of one or zero. Short of that, there is no class that uniquely captures the probative value of the evidence (that is, infidelity). This fact undermines attempts to equate probative value with probabilities based on base rates.

In response to the Davis and Follette (2002), Friedman and Park (2003) criticize the conclusions and choice of base rates. They point out that the defendant's reference class will likely vary when the boundary conditions of the class are altered to account for more realistic characteristics: "not all histories of infidelity or of spousal abuse have the same probative value" (p. 639). Second, they contend that the infidelity evidence may combine with other evidence in ways that are too complex to quantify: "[t]he insurance evidence combines with the evidence of infidelity in a way that cannot be captured by schematic quantitative analysis" (p. 642). We agree with their criticisms but think they do not go far enough.

The first criticism is illuminating because Friedman and Park effectively show that on the facts of the case there are very likely reference classes that suggest a much higher probability of guilt than the "most liberal base rate[s]" employed by Davis and Follette (2002, p. 150). This, of course, is a partial demonstration of the central point of this paper that there is a large class of reference classes. The reason Friedman and Park's critique does not go far enough is their suggestion that the Davis and Follette reference class is too simplistic ("Real life is far more

---

16. The base rates were higher for nonwhite, younger married men from lower economic categories (Davis and Follette 2002, p. 150).

17. Likewise, in a subsequent response, Davis and Follette (2003, p. 672) argue that when jurors and judges use such evidence to construct stories, they may give it excessive weight in relation to its "true utility." The true utility, for them, would be the result of their analysis.

complicated than [Davis and Follette's] scenarios" [p. 639]), which suggests that the more complex examples they give are more realistic. This can be true only if there is a "realistic" reference class (other than the event itself), but there virtually never is. As we have tried to demonstrate, generally if not always there is a practically unbounded set of reference classes with probabilities within those reference classes ranging from zero to one, and nothing privileges any particular class. Because there is no unique base rate that would capture the value of the evidence, other than the event itself,[18] all Friedman and Park can do is articulate why they think some other reference class is more pertinent, but of course Davis and Follette could offer yet another competing class that lowers the probability, and so on. The effect of specifying more details to make the rate more realistic will depend on the relevance of those details and may or may not take one closer to the actual value, which will be zero or one.

With regard to combining evidence, Friedman and Park's (2003) critique seems to assume the epistemological limitation is one of computational complexity only (p. 639). If we knew the likelihoods for, say, combinations of insurance and infidelity, then these values may indeed reflect the probative value of the evidence. We disagree; we think the problem is much deeper. Even if we knew those base rates, the reference-class problem, as we have presented it, would arise once again with the various combinations or any other combination.

Replying to Friedman and Park, Davis and Follette (2003) adjust their analysis in two ways. They focus more specifically on the base-rate/reference-class issue (pp. 669–70), and they emphasize that their theory of probative value is not the likelihood ratio but instead the "information gain" that evidence provides (the difference between prior and posterior probabilities) (pp. 668–69, 673–78). Neither emendation removes their analysis from our critique. With regard to base rates, they assert, "Although selection of relevant base rates that will be generally accepted will be 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

---

18. Of course, some could be more or less realistic and thus tell us more about the situation, but the key point is that none of these values would uniquely capture the probative value of the evidence. That is a different question from the relative persuasiveness of arguments about reference classes. Friedman and Park present, in our opinion, a strong argument that Davis and Follette (2002) substantially underestimated the strength of the evidence of guilt.

evidentiary rulings" (p. 669). But, to belabor a point, there is only one empirically "objective" reference class—the event itself. Among the various other reference classes, there is no other unique class that will capture the probative value of the evidence. Moreover, emphasizing information gain as the measure of probative value does not respond to the reference-class problem because information gain depends on base rates and thus on reference classes.

Like the carpet fiber example, the probative value of the evidence of infidelity cannot be objectively captured mathematically.[19] Rather, the varying statistics are just themselves more evidence that must be interpreted. Like all other evidence, their value will depend on what can be inferred from them, which in turn will depend on how well they explain or are explained by the various hypotheses in issue.

3.1.3. DNA Evidence.

Nance and Morris (2002) employ models similar to the ones discussed above to empirically test and evaluate how jurors value scientific evidence. They present the evidence—expert testimony about a DNA profile match between a defendant and a semen sample recovered at a crime scene—in nonquantified form and then in a variety of quantified forms (pp. 411–15).[20] They conclude that the mock jurors "significantly undervalue[d] the DNA match evidence" (p. 435), and they suggest that use of the statistical methods they recommend "can indeed assist the jury in reaching more accurate verdicts" (p. 445).

The authors divided a large pool of jurors[21] into five groups; each group was asked to elicit a mathematical probability of guilt about a hypothetical rape case:

1. The first group (the control group) was told that the prosecution's evidence was the victim's identification of the defendant before and at

19. Kaye and Koehler (2003) criticize Davis and Follette on the ground that the likelihood ratio is a better measure of probative value than the change in posterior probability is. But as we have previously noted, this is premised on a correct and unchanging likelihood ratio. We doubt that there is such a thing anywhere in life of any relevance to the legal system, and the DNA example they employ certainly is not. DNA is very good evidence, but not because there can be no disputes about the reference class into which pieces of DNA evidence fall.

20. As Nance and Morris (2002) explain, nothing in the problem depends on the fact that it is DNA evidence as opposed to some other kind of quantified evidence based on expert testimony. Thus, their experiment and conclusions—as well as our criticisms—generalize readily.

21. The subjects in the study were jurors called for service in criminal courts in Kane County, Illinois (Nance and Morris 2002, p. 407).

trial. The defense's evidence consisted of an alibi witness. This group's mean estimate of guilt was .33.

In addition to the above evidence,

2. the second group was told a DNA test matched portions of the defendant's DNA to portions tested from a sample found at the crime scene so that "one could not rule out the defendant as the source of the semen" (p. 412). No statistics were given. This group's estimate of guilt was .60.

3. The third group was also told about the DNA match and that "DNA profiles found in this case occur in only 4% of the male population," that "we would expect to see the DNA profiles found in both the recovered semen sample and the defendant in approximately 1 person out of every 25 men in the population," and that "[i]n terms of lab error, I can tell you that in proficiency tests of forensic labs in the United States, matches are mistakenly declared in about 1 out of every 1,000 cases for which the samples are not from a common origin" (p. 412).[22] This group's estimate of guilt was .52.

4. In addition to the information given to group 3, the fourth group was given a likelihood ratio for the random match. They were told that "it is 25 times more likely that one would have a match if the defendant were the source of the sample taken from the victim than if he were not" (p. 413). This group's estimate was .58.

5. In addition to the information given to group 4, for the fifth group "the expert illustrated how a likelihood ratio relates the posterior probability to the prior probability. The expert presented a chart that shows the effect of a likelihood ratio of 25 on prior probabilities and allowed the respondent to select his or her own prior probability and read the chart to discern the corresponding posterior probability" (p. 413–14). This group's estimate was .65.

Everything presented thus far we find to be both helpful and fascinating in understanding juror decision making, particularly the variances among groups 3, 4, and 5. These results are also no doubt quite helpful to practitioners who must present such evidence. We part ways with Nance and Morris, however, when they construct a "Bayesian norm," which purports to establish the true or accurate probative value of the evidence, to conclude that the jurors' assessments are false and need to be corrected to the extent they deviate from this norm (pp. 420–26).

22. According to the authors, this format most closely resembles the most common form of presenting such evidence in criminal trials.

Likewise, the authors suggest ways to improve juror evaluations, where improvement means moving assessments closer to the norm (pp. 437–45).

Nance and Morris first calculated the likelihood ratio of the random-match evidence. The ratio is the probability of a match given the defendant is the source divided by the probability of a match given that he is not. The authors took the numerator to be approximately 1.[23] They arrived at the second number by adding three figures: the probability of lab error as specified in the problem (.001), the probability of a match due to random coincidence as specified in the problem (.04), and the probability as assessed by the subjects of other false positives attributable to causes other than lab error or random match, such as police mishandling (this number was .022). This generated a likelihood ratio of 15.87.[24]

Second, they employed the likelihood ratio to update via Bayes's theorem the prior probability of guilt without the evidence (as assessed by the individuals in group 1 [the control]) to arrive at a posterior probability of .72.[25] Under their analysis, the extent to which the subjects' estimates deviate from this posterior probability (namely, the Bayesian norm) is the extent to which the subjects over- or undervalued the true or accurate probative value of the random-match evidence. In the experiment, all of the groups undervalued the evidence because their final estimates of guilt were lower than the Bayesian norm (.72): group 2 (.60), group 3 (.52), group 4 (.58), and group 5 (.65). Not surprisingly, the one group that was given a chart showing them how to update their prior probabilities consistent with the Bayesian norm came closest to it. And the group that was confronted with a scenario that most resembled the way such evidence is actually presented in criminal trials was the most inaccurate. Nance and Morris suggest that employing such Bayesian techniques may make juror assessments more accurate.

Although this example is more complex, at root it is structurally similar to the carpet fiber and infidelity examples above, and the attempt

23. More specifically, they explain that although there is some possibility of false negatives, it will not have much effect on the calculations (Nance and Morris 2002, p. 421).

24. $= 1/(.001 + .04 + .022)$.

25. The authors arrive at this "Bayesian normative posterior probability" by applying the above likelihood ratio to the prior odds as assessed by each individual in the control group—that is, those who were not presented with the DNA evidence (Nance and Morris 2002, p. 423).

to mathematically model the probative value of the evidence has similar limitations. Like the above models, Nance and Morris's calculation of probative value depends on likelihood ratios based on a reference class. Therefore, the reference-class problem again arises.

Moreover, while the presence of two ratios (randomness and error rate) may give the problem an initial appearance of greater precision, it actually creates two rather than one reference-class problem. First, consider randomness. The subjects were told that 4 percent, or one in 25, of males in the population share these DNA characteristics. But the jury is never told what "the population" means. Does it mean in the whole world, state, city, neighborhood, block, house next door? If it means among all human males, why is this the appropriate class? Suppose, for example, the victim shared some of the same characteristics as the perpetrator—which would suggest that a relative may be her attacker or someone from the same racial, ethnic, or national group. In this case, while the likelihood of the characteristics among all human males would remain 4 percent, the probability among the set of possible suspects (say, a relative) may be much higher, which would significantly reduce the value of the evidence.

The lab error rate evidence is equally problematic. The jury was told only that matches were mistakenly declared in approximately one out of 1,000 cases "in proficiency tests of forensic labs in the United States" (Nance and Morris 2002, p. 412). (Presumably this means proficiency tests of DNA tests similar to the one at issue here, but note the ambiguity/reference-class problem.) But what about the proficiency rates for the lab that conducted this test or the technician or technicians who conducted the test? If these rates differed significantly from the likelihood ratio above, so would the value of the evidence and so would the Bayesian norm as constructed by the authors. But the match evidence itself has not changed.

For all of these reasons then it seems quite inaccurate to us to assert that juror assessments that do not conform to the authors' norm are inaccurate or false or to suggest that jurors may not be fully rational in their decision making. Jurors may quite rationally be responding to these reference-class concerns (consciously or not, intuitively or not), and in any event quite plausible explanations of the empirical results exist that do not depend on the charge of error or irrationality. With the lab error rate statistics in particular, perhaps the jurors intuited the limits of the data and wanted better and more appropriate information about this lab. Given the ambiguity regarding the statistics, that they may flow

from less appropriate reference classes, and the potentially missing information regarding them, perhaps the jurors did just what one might expect—hold these problems against the party that presented the evidence (in this case, the prosecution). Rather than reach inaccurate conclusions because of their distance from the authors' Bayesian norm,[26] the jurors may have been perfectly justified (from an epistemic standpoint) in discounting the evidence to a value below what the authors believed to be appropriate. These statistics, like those regarding the carpet fiber and infidelity, were just more evidence that needed to be interpreted.

In a subsequent study, Nance and Morris (2005) focus more specifically on presenting statistics regarding error rates. Again they measure "divergence of the respondents' assigned guilt probabilities from the guilt probabilities that the respondents ought to have assigned," with the latter figure calculated by using "Bayes's rule to generate normative measures of the probability of guilt" (p. 400). As with the first study (which they refer to as phase 1), this phase 2 study uses the same error-rate information; that is, the jurors were again told that the nationwide error rate was one in 1,000. The study tested various formats for presenting this information such as providing the statistic in frequency form, in a likelihood ratio, and with a chart showing how to aggregate the error statistic with the random-match statistic (pp. 401–404). In addition, this study focused on large lab error rates as compared with the match statistics, whereas in the first study lab error rates were low in comparison with the match statistics.

The results were consistent with the first study, with Nance and Morris concluding again that jurors undervalued the statistical evidence in general but came closest to the correct value when a similar chart method was employed (pp. 433–34). The authors speculate that perhaps this is because the jurors discounted the evidence when they did not know how to appropriately combine the two statistics—a problem alleviated by the chart (p. 434). But, as we suggest above, they may have discounted the

26. To be clear, we, of course, take no issue with Bayes's theorem as a formal matter, nor with its epistemological usefulness in maintaining coherence (Talbott 2001). Rather, we question its usefulness in the juridical proof context, where there are few overt statistics that do not suffer from the issues we are discussing. Indeed, the manipulability of such numbers to accord with subjective assessments would cause some Bayesians to accept conclusions that are completely irrational from an epistemic point of view (Hájek 2003). And it is the epistemic viewpoint, we contend, that matters most in the juridical proof context.

value of the evidence below what the authors take to be the objective value (the Bayesian norm) because they were reacting (even if implicitly) to this reference-class issue. To repeat, why should national statistics be the appropriate class? Why not lab- or analyst-specific results? Indeed, in a section entitled "Limitations of the Study" (p. 428), Nance and Morris now note that "[u]se of lab-specific or even analyst-specific proficiency test data, when such are available, might affect the results, especially if such error rates are much higher or lower than prevailing juror expectations."[27] Not surprisingly, jurors would be less inclined to respond to their intuitions regarding such reference-class problems when they are given a chart telling them that the objectively correct value of the aggregate of the statistics is a certain value.[28] This might explain why the chart method yielded results closest to the norm. But, as with the first study, the experimenters do not establish that the chart values capture the correct probative value, while the discounted values from the subjects—who may very well be intuitively responding to the reference-class problem—are wrong.[29] Thus, the authors are right to flag this limitation, but they are too quick to dismiss its significance. The significance is that it shows that probative value cannot be equated with likelihoods from reference classes.

27. They also later recognize this issue when they assert that "prosecutors should not be too concerned with preventing the introduction of the results of lab proficiency studies, unless of course what is offered is the performance of the specific lab in question and its lab-specific error rate is significantly higher than national averages and prevailing juror expectations" (p. 434). Of course, prosecutors should be concerned in the latter case because the more specific rates most likely will be typically viewed by fact finders as having a higher probative value than the national ones. Note also that this reference-class problem would still arise with lab or analyst statistics. Class issues could arise with other variables as well, such as time. For example, rates for a lab over the past 10 years may reveal wildly different rates than for that same lab over the past year, and so on. Recall also, however, that trying to specify more and more details will eventually collapse the class into the event itself.

28. The studies may therefore demonstrate an unintended phenomenon. They both show that presenting evidence in different ways can move juror decision making in the direction of the bias of the experimenter. Signposts about what the experimenters wanted may be correlated with deference by the jurors to those desires, which makes this another confirmation of the Rosenthal effect (Rosenthal 1966). Perhaps strong evidence of juror rationality in these studies is that even with the charts staring at them, the jurors still insisted on discounting the evidence.

29. Indeed, juror decisions that discount in this manner would have been consistent with the National Research Council's (1996, pp. 85–87) report on DNA, which asserted that general lab proficiency statistics are not appropriate measures of possible error in a particular case.

3.2. Modeling Subjective Probative Value

The more important limitations due to the reference-class problem concerned the objective models discussed here. But such limitations also arise with attempts to model subjective assessments of probative value. To illustrate this point and also to show how such models may be put to illuminating and beneficial use despite such limitations, consider Friedman's (1991) model of character evidence used to impeach a witness.

According to his model, the probative value of any given witness's testimony is captured by a specific type of likelihood ratio: the probability of the witness testifying this way given the testimony is true—divided by the probability of the witness offering this testimony if it is false (p. 656). Character impeachment evidence is relevant to the issue of the probability of whether a witness's testimony is false; its probative value is the extent to which it affects this probability and hence the value of the testimony and the testimony's effect on the prior probability.

He presents the following example: "Suppose that Wendy Whitney was a passenger in a car driven by her wealthy neighbor Dollar when Dollar got into an accident with Poor. Poor is now suing Dollar. Whitney has testified for Dollar that Poor was contributorily negligent. Poor now wants to impeach with proof of Whitney's prior conviction for petty larceny" (p. 679).[30] The evidence at issue is Whitney's prior conviction. Friedman assumes that juror assessments of the value of the impeachment evidence will be greater than it would be if Whitney were a party because jurors will assume parties are more likely to lie even without seeing such evidence (p. 685).[31] These subjective probabilities will arise from various background assumptions about the relevant event, each of which places the event in a reference class, and the subjective assessments of the impeachment evidence may vary widely when equally plausible background assumptions are made. For example, jurors may alternatively assume that Whitney takes the threat of a perjury conviction seriously (perhaps because of fear of a harsher sentence due to her previous conviction); they may also assume that a perjury conviction is a serious

---

30. Were this model based on objective probabilities, the reference-class problem would arise quite readily. Probabilities for such categories as neighbors with convictions, neighbors convicted of petty larceny, different kinds of petty larceny, neighbors who have known each other for longer or shorter times, and so on, can vary quite dramatically.

31. Friedman (1991) also argues that jurors will assume that criminal defendants already have sufficient incentive to lie and, therefore, that character impeachment evidence regarding them should be categorically excluded when defendants testify (pp. 655–69). For criticisms of this conclusion, see Uviller (1993); see also Friedman's reply (1994a).

enough threat that parties in an ordinary civil case would not lie, unless they have shown a previous willingness to flout the law. These alternative assumptions would thus flip the conclusions.

Rather than try to model a correct or true probative value, Friedman's model is thus an excellent example of the way such models can be used for less problematic purposes. His argument does not depend on any assertions of the true or correct probative value of evidence; it depends on whether his intuitions map on to those of his readers. Nor is this a trivial exercise. He has shown that, if one accepts his intuitions about how people will behave and how they should behave, then the law plausibly should be changed.[32]

3.3. *United States v. Shonubi*

The reference-class problem arises even with more general assessments of evidence, for example, when assessing whether evidence meets a standard of proof. To illustrate this point, our final example focuses on the saga and debates regarding the *Shonubi* case. (The saga generated five opinions: 962 F. Supp. 370 [E.D.N.Y. 1997]; 103 F.3d 1085 [2d Cir. 1997]; 895 F. Supp. 460 [E.D.N.Y. 1995]; 998 F.2d 84 [2d Cir. 1993]; and 802 F. Supp. 859 [E.D.N.Y. 1992]).[33] The debate among scholars has explicitly focused on the reference-class problem (Colyvan, Regan, and Ferson 2001; Tillers 2005). We first discuss the case and the commentary and then offer an interpretation and partial defense of the Second Circuit's analysis.

In 1991 Charles Shonubi was arrested at John F. Kennedy International Airport after arriving from Nigeria, and, over a 2-day period, he passed 103 previously swallowed balloons containing in total 427.4 grams of heroin. A jury convicted him of importing and possessing heroin (103 F.3d at 1087). At trial, the government also established that Shonubi, a Nigerian citizen and U.S. resident, had made seven previous smuggling trips from Nigeria to the United States (895 F. Supp. at 488). At sentencing, Judge Jack Weinstein had to determine by a preponderance of evidence the amount Shonubi had smuggled during the previous seven

32. And we largely are persuaded that his intuitions as to how jurors will and should behave are plausible, although they have been criticized (see Uviller 1993).

33. The fifth opinion contains Judge Weinstein's assessment that the Second Circuit's analysis was of "dubious validity" and that its requirement of "specific evidence" regarding the defendant's conduct "represents a retrogressive step towards the practice relied upon from the Middle Ages to the late Nineteenth Century, which often limited the use and weight of evidence by category of evidence and type of case" (962 F. Supp. at 375).

trips. This amount constituted "relevant conduct," and Shonubi was to be sentenced on the basis of the total amount smuggled, not the amount for which he was convicted.

After a remand from the Second Circuit, Judge Weinstein held a hearing to determine the previous amount smuggled. The most significant evidence proffered during sentencing was data provided by a government expert concerning the "quantities seized from 177 Nigerian heroin swallowers arrested at JFK Airport during the same time period that spanned Shonubi's eight trips" (103 F.3d at 1088). Four experts offered their independent analyses of how to interpret these data: one prosecution expert, one defense expert, and two court-appointed experts. On the basis of the analyses of these experts (895 F. Supp. at 499–511), Judge Weinstein's own assessment of the statistical evidence, and other evidence such as Shonubi's demeanor and the judge's knowledge of the drug trade, Judge Weinstein concluded that Shonubi had smuggled between 1,000 and 3,000 grams (the relevant guideline category) over his eight trips and sentenced him accordingly (895 F. Supp. at 530).

In a short opinion, the appellate court again reversed, remanding to sentence Shonubi on the basis of the amount for which he was convicted (103 F.3d at 1087). The court concluded that the statistical evidence offered regarding other smugglers did not meet the requirement of "specific evidence" that they had previously required (103 F.3d at 1090–93). Specific evidence, the court explained, consists of evidence such as reports, admissions, and testimony that "points specifically to drug quantity for which the defendant is responsible" (103 F.3d at 1089–90). The court further concluded that none of the evidence—previous trial transcripts, demeanor, and knowledge of drug trafficking generally—provided specific evidence concerning the quantities carried on the previous seven trips (103 F.3d at 1091).

Although not explicit in the Second Circuit's opinion, the reference-class problem provides the basis for a recent defense of that opinion by Colyvan, Regan, and Ferson (2001). The authors, a philosopher and two mathematicians, conclude that in rejecting the evidence the court was implicitly responding to the issue of reference classes: "[W]hile the Second Circuit's reasons for rejecting the statistical evidence [in *Shonubi*] might not have been expressed as clearly as one would like, their decision was correct. The statistical evidence presented [in *Shonubi*] ignored the well-known reference-class problem—statistical data were presented that were based on a particular reference class as though doing so was un-

controversial. One cannot simply assume that the reference class used is privileged" (p. 176).

According to the authors, "the real issue is whether Shonubi should have been sentenced on the basis of evidence gathered from *other people*" (p. 171). Specifically, Shonubi's sentence was based on placing him in the class of Nigerian heroin smugglers arrested at Kennedy Airport between September 1, 1990, and December 10, 1991.[34] Moreover, the effect of such assumptions is to treat Shonubi as though he were actually convicted beyond reasonable doubt of the factual conclusion generated by the assumption. Colyvan, Regan, and Ferson (2001, p. 175) offer their interpretation of the Second's Circuit specific-evidence requirement to mean evidence of "Shonubi's previous behavior" such as "previous convictions, financial dealings and so on."

In response, Tillers (2005) also explicitly focuses on the reference-class issue, but he criticizes both the Second Circuit and Colyvan, Regan, and Ferson (2001) for challenging the epistemic legitimacy of group-to-individual inferences. Tillers notes that evidence law frequently endorses such inferences, for example, in allowing evidence of routine practices of an organization or behavior of a street gang to show action in conformity with the practices or behavior by individuals. He speculates that the real (unexpressed) concerns of the Second Circuit were their beliefs that such group-to-individual inferences were morally, politically, or socially problematic—rather than being epistemically illegitimate (pp. 42, 46–47). Tillers thus concludes, "No one can altogether avoid signs, signposts, and evidentiary hints that the operations of the world and other people have created. It is the failure to appreciate this basic point that makes the positions of Judge Newman [the author of the Second Circuit's opinion], Colyvan, Regan and Ferson unsatisfactory, and, ultimately, untrue" (p. 49).

We agree with these authors that *Shonubi* raises important reference-class issues. We disagree, however, with different aspects of their assessments of this importance. Colyvan, Regan, and Ferson correctly focused on the problematic inattention given to the choice of reference classes—it does indeed beg the question whether Shonubi should be seen as a typical member of the class for which the government offered base-

34. This category would be appropriate, incontestably we think, for assessing Shonubi's conduct if, for example, there were complete homogeneity among the class members or there were an immutable causal relationship between class membership and the amount smuggled; but in the absence of evidence establishing either, to assume so is to beg the critical question.

rate evidence. Their response—focusing on Shonubi-specific evidence—is, however, at least unclear and perhaps unconvincing, precisely for the reasons given by Tillers. It would be impossible to draw inferences about a defendant's conduct without placing him within classes that include conduct by other people. To interpret any statements made by a defendant (at trial or previous statements), the fact finder will assume that those words mean what they mean to other speakers in the community; to infer any kind of mental state (such as intent or knowledge), the jury will evaluate the defendant's behavior for whether it is consistent with the behavior of people generally when acting in that mental state. It is, of course, possible that the defendant was speaking or acting idiosyncratically, but even such idiosyncrasy will make sense only against a background of shared linguistic and behavioral practices (see Davidson 2005).

On the other hand, we disagree with Tillers that the Second Circuit's specific-evidence requirement was necessarily a condemnation of group-to-individual inferences, and we think he errs in his strident criticism of Colyvan, Regan, and Ferson by accusing them of adopting such a position, although they are not entirely clear about the matter. Precisely for the reasons given by Colyvan, Regan, and Ferson, the court may have been concerned about whether the district court used an appropriate class from which to draw inferences about Shonubi's conduct. We therefore suggest that a more charitable reading of the Second Circuit's opinion would be to read the specific-evidence requirement as a requirement that the evidence reasonably personalize the matter or that there be evidence or argument showing that this is an appropriate class in which to place the defendant.

The evidence presented failed to do either. It was quite general and problematic. Apparently, the most solid evidence of eight trips was absences from work; Shonubi's own passports directly indicated five trips (895 F. Supp. at 467). Whether the first of these eight trips was truly his first was unknown, and thus whether he was a novice or a seasoned veteran was unestablished. Whether any of the trips were for other purposes or failed in their illegal importation objectives is nowhere addressed. The data that the government's experts relied on had various curiosities (such as a number of the data points indicating greater net than gross weight, no data on learning curves, and so on), and there were no data on the kind of questions raised above. There were no data, for example, on the frequency with which illegal importers return from

Nigeria empty-handed or go for other reasons (such as organizing their affairs).

To be sure, it is reasonable to believe that the evidence showed a reasonable probability that Shonubi had made eight trips importing drugs illegally and that he probably imported at least the 1,000 grams critical to the sentence. Nonetheless, we also would say we are not terribly confident in those conclusions. There is, in other words, a second-order proof problem here, as Colyvan, Regan, and Ferson point out. Now suppose either of two counterfactuals. First, suppose that on the third of Shonubi's trips he had also been caught in possession of 400 grams of drugs. The second-order proof problem would recede dramatically. And if there were similar evidence about another of the trips, it would become vanishingly small. Next, suppose that the government adduced very good data about the careers and activities of Nigerian drug smugglers and that it was perfectly plain that they were a very homogenous group with a very short learning curve and that they never returned from Nigeria empty stomached, as it were. Again, the second-order proof problem about Shonubi's actual activity would recede dramatically.

If either supposition were accurate, we predict that the district judge's sentence would have been upheld. There is a strongly held belief that personalizing data—evidence that includes observational data about the particular individual—reduces ambiguity, and better data generally are better data. Neither is necessarily true—it is theoretically possible to have more personalizing and better data and to simultaneously increase the probability of an error[35]—but generally things do not seem to work that way. The Second Circuit's opinions, we suggest, are merely operationalizing this belief in the context of a sentencing hearing that had attributes of a trial on the merits.

Seen this way, the Second Circuit's opinions do not constitute a rejection of probabilistic reasoning, the importance of generalizations, or the relevance of background knowledge and experience. They are merely assertions that, given the interests at stake, the government's evidence was insufficient to justify the sentence imposed. Something more specific, that is, more reliable, was needed.[36] Exactly what something more re-

35. As the district court pointed out, the specific-evidence requirement could lead to less accuracy in certain cases (896 F. Supp. at 479).

36. The appellate court explained, "A guideline system that prescribes punishment for unconvicted conduct at the same level of severity as convicted conduct obviously obliges courts to proceed carefully in determining the standards for establishing whether the relevant conduct has been proven" (103 F.3d at 1089).

liable may be is unclear, a point on which we agree with Tillers, but we think this ambiguity affects virtually all inference problems at trial—which is the central point of this paper in a sense. The inferential problems at trial appear to defy formal treatment; it is as simple as that. This can be seen by flipping the *Shonubi* question and asking the district judge or his academic assistants to explain why the evidence actually provided at trial and sentencing established by a preponderance of the evidence that Shonubi imported a certain amount of drugs. No formal answer will be forthcoming to that question, and the answer ultimately will be in the form of the exercise of judgment. Precisely so, as always will be the case at trial, we suspect. Thus, the dispute between the trial and appellate courts was not a profound disagreement about probability theory or statistics but instead about the exercise of judgment over the facts of this case.

The *Shonubi* example connects with our analysis throughout this paper. As in the previous examples, the possible inferences drawn from statistical evidence about particular events depended on reference classes. The appropriateness of any such class, however, cannot simply be taken for granted or assumed to be appropriate just because it is the class for which data are available. It will be the subject of argument as to why one class is more appropriate, particularly when others are plausible. It is the failure to provide such evidence and argument that may best explain the Second Circuit's reaction in *Shonubi*. For this reason, the *Shonubi* saga provides a vivid example of the importance of the reference-class issue and the need for argument about appropriate classes. This importance arises not only for those attempting to model and evaluate such evidence, but for advocates as well. After all, it was the failure of the prosecution to provide such evidence regarding Shonubi that ultimately doomed its position.[37]

4. CONCLUSION

Because of the epistemological limitations flowing from the reference-class issue, mathematical models do not very well capture the probative value of evidence. The statistics that are relied on and provided by such models are just more evidence that must be interpreted. How will this

---

37. In remanding for the second time, the appellate court explained that "since the Government has now had two opportunities to present such 'specific evidence' to the sentencing court, no further opportunity is warranted" (103 F.3d at 1092).

proceed? We suggest this occurs by comparing the various hypotheses that may explain the evidence. In other words, at the microlevel of determining the probative value of evidence, fact finders are engaged in a process of inference to the best explanation (Pardo 2005, p. 382)[38] in which the contest is largely over the relative plausibility of the competing hypotheses advanced by the parties.[39] Thus, the probative value of any given item of evidence is the strength it provides a particular inference at trial. Evidence will be stronger when it supports an inference to one hypothesis (for example, this person committed the crime) over a competing one (for example, somebody else committed the crime) and weaker when it does not exclude plausible alternative hypotheses that rely on alternative assumptions. This conclusion shows an additional problem with the models discussed above: they do not account for this explanatory phenomenon. They fail to greater or lesser degrees to exclude the plausible alternative explanations that may better explain the evidence, given plausible assumptions flowing from alternative reference classes.

This process of inference to the best explanation does not much depend on the quantification of the value of individual items of evidence, although this in no way suggests that overtly statistical evidence cannot

38. Lipton (2004) is one of several leading accounts of scientific inquiry (see also Leiter 2001). The account is consistent with the general Bayesian approaches employed in the models in Section 3. Lipton (2004, p. 120) explains, "Bayesianism poses no particular threat to Inference to the Best Explanation. Bayes's theorem provides a constraint on the rational distribution of degrees of belief, but this is compatible with the view that explanatory considerations play a crucial role in the evolution of those beliefs, and indeed a crucial role in the mechanism by which we attempt, with considerable but not complete success, to meet that constraint." For empirical support for the proposition that jurors are engaged in inference to the best explanation, see Pennington and Hastie (1991).

39. This is also the best explanation of the macrostructure of trials (Allen 1994). Interestingly, there are a number of cases that, with varying disagrees of explicitness, recognize that the structure of trials involves the relative plausibility of the various explanations advanced rather than a cardinal appraisal of the probability of the plaintiff's case being true (with reciprocal appraisals of defenses) (see, for example, *Anderson v. Griffin,* 397 F.3d 515 [7th Cir. 2005]; *United States v. Beard,* 354 F.3d 691 [7th Cir. 2004]). In criminal cases, the issue is whether there is a plausible story of guilt and no plausible story of innocence (see, for example, *United States v. Newell,* 293 F.3d 917 [7th Cir. 2001]). The Supreme Court acknowledged the relative nature of proof in *Los Angeles v. Alameda Books, Inc.* (535 U.S. 425, 437–38 [2002]); in finding that the city justified its statute restricting adult businesses, the Court said, "Neither the Court of Appeals, nor respondents, nor the dissent provides any reason to question the city's theory. In particular, they do not offer a competing theory."

be quite probative.[40] Consider, for example, statistical evidence that shows a high percentage of racially disparate hiring practices or epidemiological studies showing a significant increase in disease among those exposed to a particular product. The best explanation for this evidence, absent better competing explanations, is that the employer employed racially motivated hiring practices and the product caused the disease. Now, consider the blue bus example: the fact that the Blue Company bus caused the accident may explain why the witness saw a bus, but it does not explain why the company owns most of the buses in the town (Pardo 2005).

The explanatory connection may, of course, go in the other direction: the fact that most of the buses were from one company may explain why the one involved was likely to be from that company. May explain, however—depending on the strength of competing explanations, flowing from alternative classes, relying on alternative assumptions. This is what ultimately renders problematic the models in Section 3; the plausible alternative assumptions and classes make alternative explanations at least as, if not more, probable. The models thus fail to capture the value of the evidence. The models thereby fail to demonstrate the strong conclusions claimed on their behalf.[41]

This analysis also demonstrates the consanguinity of determining the probative value of evidence—the microlevel evidentiary question—and the macrostructure of trials. At both levels, the best explanation of what occurs at trial is inference to the best explanation.[42] Conventional probability approaches are difficult to reconcile with the evidence concerning the structure of trials; the macrostructure is better explained by the relative-plausibility theory (Allen 1994). We do not claim that the two levels are analytically related, however; it is theoretically possible that

40. Thus we disagree with the general skepticism toward statistical evidence expressed by Tribe (1971).

41. Interestingly, the law of evidence in many instances implicitly accommodates the reference-class problem. The most striking example is in the very definition of relevancy that looks to what a reasonable person might think rather than what a reasonable person must think.

42. A third alternative—in addition to Bayesian confirmation and best explanations—would test hypotheses by how well they survive "severe" tests (in other words, difficult challenges to their truth). This approach is associated with the views of Karl Popper and is a competing philosophy-of-science theory (Mayo 2005). At a general level, we take our view to be consistent with this approach. The structure of the trial—which attempts to narrow down the proof process to decisions about important, contested factual issues—would appear to create such severe tests; they would be the key pieces of evidence regarding one of the contested issues, which each side's theory must attempt to account for.

proof at the microlevel involves inference to the best explanation but at the macrolevel involves conventional probabilistic measures. In other words, one can reject relative plausibility as a general macrolevel account of the trial and still accept all of our conclusions in this paper. Nevertheless, we think that the best explanation of the evidence concerning trials is that proof at trial involves inference to the best explanation from beginning to end.

REFERENCES

Allen, Ronald J. 1991. The Nature of Juridical Proof. *Cardozo Law Review* 13: 373–422.

———. 1994. Factual Ambiguity and a Theory of Evidence. *Northwestern University Law Review* 88:604–40.

———. 2000. Clarifying the Burden of Persuasion and Bayesian Decision Rules: A Response to Professor Kaye. *International Journal of Evidence and Proof* 4:246–59.

Allen, Ronald J., and Brian Leiter. 2001. Naturalized Epistemology and the Law of Evidence. *Virginia Law Review* 87:1492–1550.

Brandom, Robert B. 2000. *Articulating Reasons.* Cambridge, Mass.: Harvard University Press.

Cohen, L. Jonathan. 1977. *The Probable and the Provable.* Oxford: Oxford University Press.

Colyvan, Mark, Helen M. Regan, and Scott Ferson. 2001. Is It a Crime to Belong to a Reference Class? *Journal of Political Philosophy* 9:168–81.

Davidson, Donald. 2005. James Joyce and Humpty Dumpty. Pp. 143–57 in *Language, Truth, and History.* Oxford: Oxford University Press.

Davis, Deborah, and William C. Follette. 2002. Rethinking the Probative Value of Evidence: Base Rates, Intuitive Profiling, and the "Postdiction" of Behavior. *Law and Human Behavior* 26:133–58.

———. 2003. Toward an Empirical Approach to Evidentiary Ruling. *Law and Human Behavior* 27:661–84.

Friedman, Richard D. 1987. Route Analysis of Credibility and Hearsay. *Yale Law Journal* 96:667–742.

———. 1991. Character Impeachment Evidence: Psycho-Bayesian [!?] Analysis and a Proposed Overhaul. *UCLA Law Review* 38:637–91.

———. 1994a. Character Impeachment Evidence: The Asymmetrical Interaction between Personality and Situation. *Duke Law Journal* 43:816–33.

———. 1994b. Conditional Probative Value: Neoclassicism without Myth. *Michigan Law Review* 93:439–77.

Friedman, Richard D., and Roger C. Park. 2003. Sometimes What Everybody Thinks They Know Is True. *Law and Human Behavior* 27:629–44.

Finkelstein, Michael, and William Fairley. 1970. A Bayesian Approach to Identification Evidence. *Harvard Law Review* 83:489–517.

Finkelstein, Michael O., and Bruce Levin. 2003. On the Probative Value of Evidence from a Screening Search. *Jurimetrics Journal* 43:265–90.

Goldman, Alvin I. 1976. Discrimination and Perceptual Knowledge. *Journal of Philosophy* 73:771–91.

Hájek, Alan. 2003. Interpretations of Probability. *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. http://plato.stanford.edu/entries/probability-interpret/.

Kaplan, John. 1968. Decision Theory and the Factfinding Process. *Stanford Law Review* 20:1065–92.

Kaye, D. H. 1995. The Relevance of "Matching" DNA: Is the Window Half Open or Half Shut? *Journal of Criminal Law and Criminology* 85:676–95.

Kaye, David H., and Jonathan Koehler. 2003. The Misquantification of Probative Value. *Law and Human Behavior* 27:645–59.

Koehler, Jonathan 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.

Leiter, Brian. 2001. Moral Facts and Best Explanations. *Social Philosophy and Policy* 18:79–101.

Lempert, Richard. 1977. Modeling Relevance. *Michigan Law Review* 75:1021–57.

Lipton, Peter. 2004. *Inference to the Best Explanation*. New York: Routledge.

Mayo, Deborah G. 2005. Evidence as Passing Severe Tests: Highly Probable versus Highly Probed Hypotheses. Pp. 95–128 in *Scientific Evidence: Philosophical Theories and Applications*, edited by Peter Achinstein. Baltimore: Johns Hopkins University Press.

Nance, Dale A., and Scott B. Morris. 2002. An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Large and Quantifiable Random Match Probability. *Jurimetrics Journal* 42:403–45.

———. 2005. Juror Understanding of DNA Evidence: An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Small Random-Match Probability. *Journal of Legal Studies* 34:395–443.

National Research Council. 1996. *The Evaluation of Forensic DNA Evidence*. Washington, D.C.: National Academies Press.

Pardo, Michael S. 2005. The Field of Evidence and the Field of Knowledge. *Law and Philosophy* 24:321–92.

Pennington, Nancy, and Reid Hastie. 1991. A Cognitive Model of Juror Decision Making: The Story Model. *Cardozo Law Review* 13:519–57.

Reichenbach, Hans. 1949. *The Theory of Probability*. Berkeley: University of California Press.

Rosenthal, Robert. 1966. *Experimenter Effects in Behavioral Research*. New York: Appleton-Century-Crofts.

Salmon, Wesley C. 1967. *The Foundations of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.

Talbott, William. 2001. Bayesian Epistemology. *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. http://plato.stanford.edu/entries/epistemology-bayesian/.

Tillers, Peter. 2005. If Wishes Were Horses: Discursive Comments on Attempts to Prevent Individuals from Being Unfairly Burdened by Their Reference Classes. *Law, Probability, and Risk* 4:33–49.

Tillers, Peter, and David A. Schum. 1992. Hearsay Logic. *Minnesota Law Review* 76:813–58.

Tribe, Laurence H. 1971. Trial by Mathematics: Precision and Ritual in the Legal Process. *Harvard Law Review* 84:1329–93.

Uviller, H. Richard. 1993. Credence, Character, and the Rules of Evidence: Seeing through the Liar's Tale. *Duke Law Journal* 42:776–832.

Wells, Gary L., Elizabeth A. Olson, and Steve D. Charman. 2003. Distorted Retrospective Eyewitness Reports as Functions of Feedback and Delay. *Journal of Experimental Psychology* 9:42–52.