convey some sense of the power of the laboratory's evidence.

Second, I worry that forcing jurors to articulate prior odds conditioned on the non-DNA evidence and to multiply this prior by a likelihood ratio may omit (and possibly divert attention from) major uncertainties in the experimental evidence. As indicated in Sections 1 and 2, the likelihood ratio $R$ does not account for the risks of missing bands, extra bands, population misspecification and substructure.

One can respond, as I suspect Berry might, that all conceivable sources of error need not be reflected in a single figure for the posterior odds. One might treat Berry's analysis as conditioned on the absence of experimental embarassments, at least where the laboratory has observed rigorous protocols (compare OTA, 1990). Where it is not clear whether a suspect is homozygous or a fragment has gone undetected, one can compute distinct values of $R$ under each assumption—as in Berry's discussion of *Castro*. Similarly, one can perform multiple computations of $R$ and hence $P(G \mid X)$ for different racial categories.

The final result, however, is no longer a simple posterior probability for guilt or even a single table of posteriors and priors. It is a set of competing numbers or tables—accompanied, quite possibly, by some nagging doubts that must be left out of the equations for want of adequate data or analytic tools. If the residual uncertainty is substantial, then the jury must attend to it in some intuitive fashion anyway. It cannot take $P(G \mid X)$ at face value if the defendant (or the prosecution in a case in which the defendant offers an exculpatory DNA profile) raises serious questions about population structure or other uncertainties not included in sensitivity analysis of $R$ and $P(G)$. And if this situation does materialize, one is left too wonder once again whether the expected payoff from the Bayesian format is worth the demands it places on the experts, the parties and the court.

## 4. CONCLUSION

As a lawyer, I see in Berry's article a cogent and powerful indictment of the matching and binning reasoning now used in single-locus DNA profiling. Berry builds an impressive case for using likelihoods that (a) make better use of the information in the test results and the population data and that (b) handle more of the uncertainties now present in DNA evidence.

I am less enamored of the strong Bayesian demand that jurors should quantify their prior probabilities and combine them with likelihood ratios based on certain simplifying assumptions to return a verdict of guilt or innocence. Like the courts, however, I am not prepared to say that there is no room for some form of a Bayesian presentation in a criminal trial. Considering the difficulties that many courts, attorneys and jurors face in assessing quantitative evidence, the efforts of Berry and other statisticians (e.g., Kadane, 1990; Fienberg and Kadane, 1983) to develop suitable Bayesian analyses for forensic applications are a most welcome development.

# Comment

**Ian Evett**

Professor Berry's analysis of DNA profiling is elegant and penetrating. I will not discuss the detail of his treatment but will concentrate on issues touched on by the other discussants that are relevant to the work of the forensic scientist.

First, should the forensic scientist adopt a Bayesian view of evidence evaluation? It has been the convention, from the first glimmerings of the science, to view evidence from a frequentist perspective. Consider a simple case where the evidence consists solely of a blood stain at a crime scene and there is a single suspect who gives a sample of blood. Assuming a system of discrete alleles with no measurement error then, if the suspect's blood and the scene blood are the same type—say X1—the scientist will refer to a data collection of some sort and, as well as reporting a

*Ian Evett is on the staff at Central Research and Support Establishment, Home Office Forensic Science Service, Aldermaston, Reading, Berkshire, RG7 4PN, England.*

match, will make a statement of the form "this blood type occurs in about 1 person in 100." The implied inferential theory is that the smaller the frequency, the stronger the evidence of association. This seems intuitively reasonable, but what happens in more complicated cases? Assume that the crime had been committed by two men, both of whom left blood stains at the crime scene one being type X1, the other type X2. A single suspect is apprehended and he is found to be type X1. That type, as before, has a relative frequency of 1 in 100, whereas type X2 has a frequency of 1 in 10. The frequentist solution is to take the sum of the two frequencies but this, stemming from intuition rather than logic, leads to illogical corollaries. The Bayesian approach (as described by Evett, 1987) provides a logically consistent solution. In the first case the likelihood ratio is 100, in the second it is 50.

Now consider a more complex case. A person has been stabbed to death; an examination of the suspect's clothing reveals blood staining that is type X1, the same type as that of the victim, the suspect being of some other type. Clearly, the statistic of 1 in 100 has some relevance here but what of the other aspects of the evidence? Is this the sort of blood staining that the expert would expect if the suspect were indeed the person who had stabbed the victim? (What if, for example, the blood staining is pinhead size on the inside of the back of the suspect's jacket?) Is this the sort of blood staining that the expert would expect to find on the clothing of an innocent member of the population? In this sort of situation, the statistic may actually be one of the less important aspects of the case. The scientist needs some sort of framework for coherently fusing his (or her) expert judgement—the "soft" probabilities—with whatever "hard" statistics may be available. The Bayesian paradigm provides the framework.

In the Forensic Science Service of the Home Office we have, for a few years, been exploring such inferential issues through workshops and it has been useful to encapsulate the Bayesian approach in three precepts.

**1. To evaluate evidence it is necessary to consider two explanations for its occurrence.** Strictly speaking, of course, this should read "two or more" but the principles emerge more clearly from simple examples, and it is noticeable that in most cases it is possible to reduce the alternatives to two: one prosecution, the other defense.

**2. It is necessary to establish the probability of the evidence given each of the two explanations.** This is the toughest one to assimilate but it is, of course, the most important. Familiarity does not come easily which is why we favor the workshop approach to interpretation training.

**3. The ratio of the two probabilities measures the strength of evidence in relation to the two explanations.** Pure statisticians may recoil from such an imprecise framing of Bayes' theorem but, in the real operational world I believe that broad concepts are more important than rigor.

I find that, in general, scientists react well to this kind of approach but eventually we arrive at the crunch question: "How do I explain this to a jury?" Kaye distinguishes between "weak" and "strong" Bayesian formats and these are useful distinctions. Frankly, I don't consider the strong format practical. Remember that the forensic scientist is no statistician and the strong format requires, in my opinion, a statistician of authority with highly persuasive powers of communication to pull it off. Although I have never tried it in the witness box, I have done so in case conferences with, I estimate, 50% success and in lectures to lawyers with a similar degree of success. The main problem, I find, is that, contrary to what many statisticians take a given, laymen are not comfortable with thinking in terms of odds. There seems to be a tendency among statisticians to regard the majority of the population as betting people, but this is not the case and the actual proportion who can reason coherently in this way is, I suspect, very small. Why should we believe that lawyers, judges and juries are special in this context?

But just leaving a court with a likelihood ratio does not seem enough. The scientist, taking great care, may say something like "the evidence is 100 times more likely if the first explanation is the true one" but it can be almost be guaranteed that the lawyer will respond with "do you mean that the first explanation is 100 times more likely than the second?" The response "no, I don't mean that" will be met with a somewhat impatient "well what *do* you mean?" Whereas we can discourse at length from the comfort of our armchairs about this, we must recognize the difficulties posed by being in one of the loneliest places on earth—the witness box!

Accordingly, I have come to a compromise stance that orthodox Bayesians will consider rank heresy. I favor a verbal convention, which maps from ranges of the likelihood ratio to selected phrases. Thus, for example, a likelihood ratio of 450 may be described as "strong evidence"; one of, say, 10, would be described as "weak evidence." I fully realize that there are problems with this approach, but, first, I do not believe that numerical precision is necessary at presentation stage and, second, no statistician has yet to suggest to me a more practical method of solving the problem. I eagerly await the response of readers of this journal.

Two of the other discussants raise the issue of band independence and multiplying probabilities. This has attracted enormous interest in the United States. A lot of it, apparently, from eminent people who hitherto had little or no experience in the forensic field. It is good that such people should take an interest, but it is important that they should realize that they have entered a field that may have requirements, strictures and problems that are different from those that they are accustomed to working with. Frankly, I find the discussions of Hardy-Weinberg equilibrium, linkage disequilibrium and population substructuring confusing and often of no more than tangential interest. The important question for me is not "can I prove independence?" but "is there any evidence of dependence effects that would have any practical impact on operational casework?" The robustness studies that I have carried out with colleagues have served to increase the confidence with which we carry out our casework procedures. Even when we have constructed artificially stratified simulated populations (Evett and Gill, 1991), we have failed to produce effects which would cause operational disquiet. We have carried out all $N(N-1)/2$ comparisons in a file of $N$ Caucasians using a file of Afro-Caribbeans for estimating frequencies (Evett and Pinchin, 1991) and have shown that even under these conditions our operational procedures are robust. Our most recent work (Evett, Scranage, Pinchin and Buffery, 1991) has shown that, if there are any between-probe dependence effects in U.K. Caucasians, they are too small to have any practical effects in case work. We do not take these results as a source of complacency, nor do we claim that they have universal implications for all countries and racial groups. However, we do suggest that the fears that have been expressed may sometimes grow out of reasonable proportion.

# Rejoinder

**Donald A. Berry**

I thank the dicussants for their clear and insightful comments. All discussants have important concerns and identify important problems for future research. I am pleased that all four seem to favor the approach I describe in preference to match/binning. The editors tried to find discussants who use match/binning and who would argue its merits, but unfortunately they were unsuccessful.

While I have no major disagreements with the discussants, I will respond to some of the points they raise.

## RESPONSE TO LANGE

Lange correctly points out that I did not dwell on departures from the independence assumptions. In Berry, Evett and Pinchin (1991), we extend the results of the current paper to the bivariate setting of pairs of bands on a single-locus probe. The approach does not assume independence of the two bands. The second "key independence assumption" is more difficult to relax since going to higher dimensions has calculational and sample size implications. Unpublished results of Evett and his colleagues (see Evett's discussion) indicate that independence is not a concern for the probes used in the UK Home Office Forensic Science Service. However, *measurement errors* across probes are highly correlated; research to account for such "band shifting" across probes is ongoing.

Lange likes the name "identity index" for R. I like it too. Actually, while I will continue to use both, neither "Bayes factor" nor "likelihood ratio" is ideal. The former carries a bit more philosophical baggage than R deserves. The latter is somewhat of a misnomer because R involves Bayesian averaging in both numerator and denominator.

Lange worries about the ability of judges and juries to adjust priors to posteriors. This worry is shared by Kaye, Evett and many others—including me! Judges and juries should be given (1) information they can understand, and (2) information that is correct. There can be no compromise regarding (2). If something we provide is correct but we know