yet be the seed from which grows a useful method for comparing and evaluating forecasters. One step in this direction has been taken by Rubin (1984), but more work is needed.

Acknowledgment. The author would like to thank Teddy Seidenfeld, Phil Dawid, Rob Kass, Morris DeGroot, and Joseph Verducci for serious discussions on this subject.

REFERENCES

DAWID, A. P. (1982). The well-calibrated Bayesian. *J. Amer. Statist. Assoc.* **77** 605–613.
DE FINETTI, B. (1974). *Theory of Probability.* Wiley, New York.
OAKES, D. (1985). Self calibrating priors do not exist. *J. Amer. Statist. Assoc.* **80** 339.
RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172.
SCHERVISH, M. J. (1983). A general method for comparing probability assessors. Technical Report 275, Department of Statistics, Carnegie-Mellon University.
SCHERVISH, M. J. (1985). Comment on "Self calibrating priors do not exist" by David Oakes. *J. Amer. Statist. Assoc.* **80** 341–342.
WALKER, A. M. (1969). On the asymptotic behaviour of posterior distributions. *J. Roy. Statist. Soc. Ser. B* **31** 80–88.

DEPARTMENT OF STATISTICS
CARNEGIE-MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213

REJOINDER

A. P. DAWID

University College London

Mark Schervish musters some convincing arguments and examples to back up his position, outlined in my final paragraph, that the mathematics I have developed cannot be regarded as establishing the concept of empirical probability on a firm footing. All in all, I am in agreement with him. The essentially asymptotic nature of any criteria for empirical validity of probability assignments must mean, quite simply, that these can never be applied to finite experience in anything other than a nonrigorous and suggestive way. (The half-baked suggestions of my Section 13.4 clearly attest to this.)

This consideration applies just as much to traditional frequency-based interpretations of probability as to my attempted extension. Indeed, I have considered elsewhere (Dawid, 1985c) some of the logical difficulties that dog attempts to understand the probability assignments of the Bernoulli model in terms of limiting relative frequencies, and reached conclusions similar to Schervish's, arguing that an entirely subjective approach to the relationship between prob-

abilities and empirical frequencies, based on de Finetti's exchangability concept, is the only logically satisfying one. That paper considered the construction and interpretation of models for data still to be observed, rather than, as here, the empirical validation of proposed models in the light of data. This raises new problems of applied inductive inference.

Thus consider an observed finite sequence of outcomes of coin tosses that we are happy to explain as Bernoulli trials with probability $\frac{1}{2}$. We might have looked at calibration, or applied various familiar statistical tests (but not too many!) to satisfy ourselves of this. But, however long that sequence may be, we cannot legislate to Nature that the penny must continue to yield Bernoulli trials, or that heads and tails should be equally frequent. Based on much accumulated experience of similar processes we may indeed expect this, but there is no "law of Nature" which ensures it (any such law could only be tentative, and subject to revision if Nature decided—why not?—to behave differently). Thus any projection into the future of the empirical adequacy of probability assignments must remain speculative.

The subjective and tentative nature of such projection is clearly brought out in Schervish's Example 4.3: While it is logically possible that the bum's coin does forecast the weather perfectly, our background experience would suggest a strong belief in the contrary (this is my interpretation of the phrase "Nature assures us that the bum is just lucky"). Consequently, even if he has forecast correctly, for a very long period, we might still have doubts about projecting this ability into the future. However, our doubts would probably not be so strong if the same forecasts had been produced, instead, by an experienced weather forecaster.

If now the bum's success continued for an exceedingly long time, we might eventually feel obliged to take it seriously. But, having done so (and so having, in effect, postulated a new "law of Nature" connecting the bum's coin with the weather), how can we be sure that we might not then be rewarded by a complete breakdown of the connection at some later point? The same essential nonprojectability holds if the bum is spinning a pointer to provide well-calibrated probability forecasts, rather than categorical ones, and continues to hold if we consider instead the performance of an experienced forecaster. All this is simply to agree with Schervish that (as I pointed out in Dawid, 1985b), it is impossible to guarantee good forecasting performance, even though past data indicate it.

If I am inclined to agree with Schervish's criticisms of empirical probability, what can I rescue from my work? The mathematics still stands, of course, but Schervish points to the lack of any logical implications for the practice of Statistics. While this is so, there are, I think, some important lessons that, although not logically mandatory, should guide us in that practice.

First, it does make sense to reject (even though only tentatively) a probability model in the light of data: One should not forever doggedly hold onto an empirically invalid model. This implies, in particular, that a Bayesian might have to reconsider an initially proposed subjective distribution for all the data, in the light of partial data. Likewise a classical statistician may need to reject his statistical model, without having an alternative.

Second, this informal but very real testability of forecasts against outcomes strongly supports the meaningfulness, and hence the practical value, of formulating statistical inferences in terms of such sequential forecasts [the "prequential approach" of Dawid (1984)], rather than in terms of forever unobservable parameters. We should not be afraid of making our inferences testable.

Third, it surely makes sense to consider, ahead of getting data, the likely extent of empirical validity of our future forecasts. What will happen if Nature will generate outcomes from a distribution Q [i.e., so that forecasts (q_i) from Q will be empirically valid], but we use forecasts (p_i) constructed from P? Can, perhaps, a single P be expected to do well for each $Q \in \mathcal{Q}$, some postulated statistical family of possible distributions for the data? What we would want here is $p_i - q_i \to 0$ with probability 1 for each $Q \in \mathcal{Q}$, where the (p_i) are fixed but the (q_i) depend on Q. This is the property of *prequential consistency*, which can indeed be achieved for parametric, and many nonparametric, families $\mathcal{Q}$. (Refinements relating to the speed of convergence of $p_i - q_i$ to 0 are also available, and form the subject of *prequential efficiency*.) To the extent that much of traditional statistical theory is content to work within the ambit of a family $\mathcal{Q}$ of possible distributions, it should be granted that considerations of expected empirical validity within this restricted ambit could be of interest and value.

Another advantage of restricting in advance the ways in which Nature is considered able to behave, by assuming she will use some $Q \in \mathcal{Q}$, is that many of the general problems of projectability of past forecasting performance, to which Schervish draws attention, disappear. In such a restricted (but still very broad) context, where we are imposing on Nature our (subjective) beliefs that she will behave herself by using a common model for both past and future, model-based forecasts that have performed well for a long time can be expected to continue to do so. Thus one of Schervish's "tricks of infinity" can be brought under control by working within the standard statistical framework.

This restriction of scope does not solve the problems of deciding when we are close enough to infinity for asymptotic results to apply, but these are not essentially new. For example, Schervish's Example 2.1 is no different in principle from the asymptotic behaviour of Bayesian posterior distributions in a regular parametric problem (Walker, 1969). Given enough data, any two Bayesians with mutually absolutely continuous prior distributions will come into close agreement over the posterior distribution (a normal distribution centered on the maximum likelihood estimate). However, for any fixed quantity of data, we can always find two Bayesians whose priors were so far apart that their posteriors are still in wide disagreement. (Similar considerations apply to the asymptotic sampling distribution of the maximum likelihood estimator conditional on various possible choices of an asymptotically ancillary statistic.) This nonuniformity of convergence does not, however, mean that considerations of asymptotic posterior distributions, free of the prior (or of sampling distributions, free of ancillaries) are completely useless. In just the same way "asymptotically unique" valid probability forecasts are not completely useless, and I dispute Schervish's conclusions from his Example 2.1. I do, however, agree with him that more careful study of the use to be made of such asymptotic animals must involve subjective considera-

tions, broadly comparable to the analysis by Edwards et al. (1963) of the effect on the posterior distribution of prior inputs.

In summary, I am extremely grateful to Mark Schervish for his penetrating commentary on my work, which greatly helps in clarifying its practical implications and limitations.

Additional Reference

DAWID, A. P. (1985c). Probability, symmetry and frequency, *Brit. J. Phil. Sci.* **36**, 107–128.

DEPT. OF STATISTICAL SCIENCE
UNIVERSITY COLLEGE LONDON
LONDON WC1E 6BT
ENGLAND