WALD, A. (1937). Die Widerspruchsfreiheit des Kollektivbegriffes der Wahrscheinlichkeitsrechnung. *Ergebnisse eines mathematischen Kolloquiums* **8** 38–72.

DEPARTMENT OF STATISTICAL SCIENCE
UNIVERSITY COLLEGE LONDON
LONDON WC1E 6BT
ENGLAND

DISCUSSION

MARK J. SCHERVISH

Carnegie-Mellon University

I wish to thank Professor Dawid for providing such a thought-provoking paper to discuss. He has raised an interesting question in his paper, namely, "Do objective probabilities for events exist, relative to a given information base?" Professor Dawid suggests that the answer is yes, while this discussant believes that the answer is no.

1. Existence. Professor Dawid's main Theorems 7.1 and 9.1 prove the asymptotic closeness of computably calibrated computable forecasts. Their existence for any given forecasting problem is an open question. The purpose of this section is to cast doubt on their existence.

Whether or not there exists a single sequence of computably calibrated computable forecasts depends on exactly which sequence a actually occurs. Schervish (1985) has shown that there are uncountably many sequences a such that not a single computably calibrated computable forecasting system exists. That is, there are as many noncalibrable sequences as there are calibrable ones. The claim, which Professor Dawid makes, that the noncalibrable sequences are sparse in an intuitive sense, is an understandable outgrowth of the fact that, as statisticians, we view the world through the rose-colored glasses of computable forecasting systems. Hence, we see only calibrable sequences (with probability 1). But Nature is not (to my knowledge) hampered by the same computability restrictions as statisticians are. It follows, then, from the cardinality argument above that the most positive answer we can give to the question of the existence of objective probabilities is "Maybe they exist, maybe not." In Section 2 we will show that even such a weak positive answer is unwarranted.

Even if the sequence a is noncalibrable, there is no cause for alarm in the forecasting community. It may very well be the case that, for many forecasters, the majority of forecasts in any finite initial segment are still quite good. That is, most forecasts may still be close to the indicators of the forecast events.

2. Probabilities of events. Suppose that the sequence a which will occur will be calibrable. (Please, do not ask how we might know this.) What then are

the objective probabilities for the events being forecast? The answer ought to be "Whatever forecasts are given by a calibrated forecaster." Unfortunately, there will be infinitely many such forecasters and they will disagree on every forecast! At first, this would seem to conflict with Professor Dawid's Theorems, but the following example will clarify matters.

EXAMPLE 2.1. Suppose Nature is providing a collective $\{a_j\}$ with proportion of 1s equal to 0.2. Consider the following set of computable forecast sequences denoted C_i. (Note that i indexes the forecasting system and j indexes the sequence of forecasts for each system.) C_{ij} equals 0.8 for $j \leq i$ and equals 0.2 for $j > i$. It is clear that according to Professor Dawid's criteria, all of the forecasting systems C_i are calibrated. However, for every N, no matter how large, infinitely many of these calibrated forecasters will be providing the "wrong" forecast on trial N.

The reason for the apparent conflict between Example 2.1 and Theorems 7.1 and 9.1, is that the theorems only show that *every given finite set* of calibrated forecasters will agree asymptotically. The problem is that for any finite set of calibrated forecasters, there are infinitely many other calibrated forecasters that do not yet agree with the given finite set. Which set of forecasters is providing the "asymptotically unique" forecasts?

Of course, Professor Dawid is aware, as he states, that "No such theory can ever justify assigning particular probabilities to particular events". But Example 2.1 shows that the theory cannot even justify "asymptotically" assigning probabilities to sequences of events. Since every one of the C_i systems is calibrated, 0.8 is as objectively valid a forecast as 0.2 for every trial but not in the limit. Here we have another example of a result which is true for every N but false in the limit. Infinity can play dirty tricks on us.

The situation illustrated in Example 2.1 is not peculiar. In fact, it will always be the case (when a is calibrable) that for every event, infinitely many calibrated forecasters will disagree. Hence, we see that even if we consider the question of the "asymptotic" existence and/or uniqueness of objective probabilities for events, the answer is that even if they exist, they are not unique.

3. Interpretation of Theorems 7.1 and 9.1. Since we now realize that unique objective probabilities (in the sense Professor Dawid intends them) do not exist, how are we to interpret the main results (Theorems 7.1 and 9.1)? The usual way to interpret limit theorems is as approximations for large sample problems. Professor Dawid suggests this in his section on data analysis, and with the comment that all "empirically valid" forecasts "must be in essential agreement, given sufficiently extensive experience." To see that even such a statement is not correct, recall Example 2.1. There we saw that no matter how extensive our experience was, calibrated forecasters were *never* in essential agreement. That is,

the "sufficiently extensive experience" needed to ensure that all "empirically valid" forecasts are in essential agreement is nothing short of infinite.

The problems with interpreting Theorems 7.1 and 9.1 for finite sample sizes are even more serious than Example 2.1 indicates. First, consider the following embellishment of Example 2.1.

EXAMPLE 3.1. Under the same conditions as Example 2.1, let $D_{ij} = 1 - C_{ij}$. It is clear that none of the D_i is calibrated. However, for every N, no matter how large, after N trials, the results will suggest that all but finitely many of the D_i are calibrated and that only finitely many of the C_i are calibrated. The evidence based on *every* arbitrarily large initial set of trials will suggest just the opposite of what will actually happen.

Next, consider the following even more problematic situation, which arises in every forecasting problem, no matter what a is.

EXAMPLE 3.2. Consider the sequence of forecasting systems F_i, where the jth forecast made by system F_i is denoted by F_{ij}. To determine what the forecasts are, let m be the integer part of $(i + 2^j - 1)/2^j$, and let F_{ij} equal 0 if m is odd and 1 if m is even. No matter what a is, and no matter how big n is, so long as the information base always contains n, *infinitely many* of the F_i systems will not only appear calibrated by time n, but will have predicted perfectly so far.

What Examples 3.1 and 3.2 tell us is that, no matter how long an initial string of events and forecasts we have seen, we *cannot* yet tell which forecasters are calibrated and which are not. In fact, we will be led to believe that infinitely many calibrated forecasters are not calibrated and that infinitely many non-calibrated forecasters are calibrated. Such a result should not be surprising; it is a simple consequence of the fact that no finite initial segment of a sequence, no matter how long, sheds any light whatsoever on the question of the convergence of the sequence. Convergence is always a function of the tail of the sequence. Similarly, calibration is always a function of the future events, not of the initial segments.

Professor Dawid's Theorems 7.1 and 9.1 are not like other asymptotic results, which do have finite sample interpretations. Take, for example, the asymptotic normality of a posterior distribution (c.f. Walker, 1969). For a given sample of size n (with a given likelihood and a given prior), one can calculate the actual posterior density as well as the normal approximation. If the normal approximation is close enough to the actual posterior (a subjective judgment), it does not matter whether or not there actually exists an infinite sample; one can use the approximation in the finite sample. On the other hand, Theorems 7.1 and 9.1 require not only the existence of the entire infinite sequence a, but also properties of the sequence which can never be checked, even approximately, with finite initial segments. No finite sample analog of the definition of calibration (based solely on the outcomes and forecasts known up to time n) could possibly imply that future forecasts of two arbitrary systems would have to be close together.

That is, if all we knew about two forecasting systems was that they had each produced forecasts which satisfied some finite sample analog of calibration at time n, there would be no way we could guarantee that any future forecasts of the two systems would be close together, since there are some forecasting systems for which the initial segments do not control the future forecasts. For example, two F_is from Example 3.2 can agree on all of the first N forecasts, but they will disagree infinitely often thereafter.

4. Philosophical implications. Professor Dawid suggests that his uniqueness result "raises difficulties for the forecaster who cannot guarantee that he will produce objective forecasts." Surely, it must be evident that *no* forecaster can guarantee that he will produce objective forecasts. Even if a is calibrable (which nobody can guarantee), Oakes (1985) proves that there is no universal algorithm to guarantee calibration. Hence, the scope for subjective disagreement between forecasters remains intact, because each forecaster believes that he/she is the one who will be calibrated in the end (Dawid, 1982). Nobody can prove otherwise before time infinity. It is true that some forecasters will look worse than others before time infinity, but, as Examples 2.1, 3.1, and 3.2 show, looks can be deceptive.

But Professor Dawid claims that his objective forecasts express a "quasi-logical relationship between the information utilized and the outcomes. In effect they provide a measure of 'partial implication', i.e., the strength with which it is reasonable to assert that the forecast events will occur, on the (generally inconclusive) evidence of the data gathered." Another example will illustrate the flaw in this argument.

EXAMPLE 4.1. Consider two different forecasting systems A and B, perhaps based on mutually singular probability models for Nature. Suppose Nature decides to ensure that either A or B is calibrated, but not both. But she postpones the decision of which one to calibrate until after some arbitrarily large N. In fact, she may alternate between them for an arbitrarily long time, never knowing for sure whether she will change her mind back to the other one. That is, no matter how large N is, not even Nature knows which system A or B will be calibrated, although one of them will be.

How can we possibly say that the forecasts of either A or B (but not both) are the "strength with which it is reasonable to assert that the forecast events will occur, on the evidence of the data gathered"? The evidence of the data gathered is completely irrelevant to the issue of which of them provides the objective forecasts. Nature has not yet even determined which of them is calibrated! It is the *evidence of the data not yet gathered* that provides the strength with which it is reasonable to assert that the forecast events will occur. In fact, long before the necessary evidence is gathered, we already know which events occurred and the whole issue is moot.

The device which Professor Dawid employs, of using the future to determine what is "valid" now, is typical of frequentist attempts to discredit the theory of

subjective probability. In fact, Professor Dawid himself states that from the subjectivist viewpoint espoused by de Finetti (1974) it would "seem that any set of forecasts is as good as any other." de Finetti, however, was only trying to set up minimal criteria that forecasts should satisfy to be coherent based on data currently available. One forecast is as valid as any other *now* if they all satisfy the requirements for coherence based on what knowledge is available *now*. This is not to say that all coherent forecasts should be equally well received by all users. Each individual user of forecasts will have subjective criteria upon which to judge the goodness of forecasts. Some of these criteria may be measures of past performance. Others might be measures of confidence in the systems and information bases being used. (See Example 4.3 below for a case in which such confidence might be lacking.) But all of the criteria must be available at the time the judgment is to be made. Professor Dawid is suggesting that we judge forecasts based on data *not yet* available. Of course, certain forecasts will look better after it has been discovered what the outcomes of the events are. But when we have to compare forecasters before learning the outcomes of the events, we must base the comparisons on information available before learning the outcomes of the events. Perhaps another example will help to clarify this point.

EXAMPLE 4.2. An urn is filled with 10 red and 10 blue balls. Balls will be drawn with replacement and forecasters must provide probabilities of the events $E_n = \{\text{draw a blue ball on attempt } n\}$. If a forecaster B consistently assigns probability 0.6 to E_n, are his forecasts invalid? Professor Dawid would have us believe that the answer depends on what the actual sequence of draws looks like. Maybe B believes that the balls are not being mixed well or that blue balls are more pleasant to the touch of the person drawing, etc. Even if we sample many times and obtain blue balls approximately half of the time, how can we claim, based on the evidence gathered so far, that B's forecasts are invalid? After all, whenever a blue ball was drawn, 0.6 was a better forecast than 0.5. We may not agree with B's forecasts and we may choose to model the draws from the urn differently than B does, but such a choice is a subjective judgment on our part and not "objective" or "empirically valid." Whether we agree with B's forecasts will depend on how seriously we take his reasons behind them.

On the other hand, assume that 20 draws will be made without replacement. Then it is clearly incoherent, invalid, and simply wrong to assign probability 0.6 consistently to each of the events E_n (assuming B knows the composition of the urn and what "without replacement" means). But this is because we *know* the "limiting" proportion of blue balls is 0.5. Unless we *know* (rather than simply believe) something similar in the example with replacement, we have no right to invalidate the assignment of 0.6. The same thing occurs in all forecasting problems. Unless we know something about the future, we cannot *now* say that a forecast is invalid. We do not have to accept it as our own subjective probability, but it is still a valid forecast.

Professor Dawid himself notes that "no finite collection of probability forecasts can be declared invalid." And yet he still suggests that we use significance

tests to detect the "acceptable" finite sequences. Are the finite sequences which fail the test unacceptable but not invalid?

A final philosophical problem with Professor Dawid's program arises out of his metacriteria. M4 seems to have a built-in bias in favor of Theorems 7.1 and 9.1. In fact, in Section 5 below, we point out a case in which one may not desire such a criterion. Criterion M3 seems innoccuous, while M1 is biased toward frequentist criteria. It is metacriterion M2 which is the enigma. On the one hand, it intuitively makes no sense to evaluate a forecasting system on the basis of what might have happened, but it also makes no sense to ignore any belief one might have about what is likely to happen. Consider the following example.

EXAMPLE 4.3. A bum sits on the streetcorner flipping a coin once each day. Eventually, he begins obtaining heads before each rainy day and tails before each dry day. (By "eventually" I mean the same thing as Professor Dawid does when he says "asymptotically".) This sequence of coin flips can be considered as a forecasting system (always forecasting probability 0 or 1 for rain). Suppose that Nature assures us that the bum is just lucky. Nevertheless, assuming that the information base contains the bum's coin flip, all computable forecasters will (eventually) have to agree with the bum if they hope to be calibrated.

We are compelled by metacriterion M2 to treat the bum's forecasts on the same footing as every scientifically based system. It is true that the bum's forecasts were the "best" in a technical sense, but to call them "objective" seems to be overdoing it. Just because the bum gave the best forecasts, do we want to call him the best forecaster? It is fine to use metacriterion M2 for awarding prizes to forecasters after the fact, but if one wishes to describe a relationship between the forecasts and the outcomes, such as "partial implication," then one must be careful to separate the sublime from the ridiculous.

5. Practical implications. Of more interest than the philosophical implications of Professor Dawid's results are the practical implications. That is, what use can a forecaster make of Theorems 7.1 and 9.1 if he/she must deal with finite information bases? The simplest and most correct answer is, of course, "None." But Professor Dawid suggests that his results might be "suitably interpreted for finite outcome sequences." For example, he suggests that "we might choose some collection of computable selection rules, ordered in some reasonable way," while noting that this choice will necessarily leave out some computable selection rules. In fact, some of the selection rules left out will be of the form needed to prove Theorems 7.1 and 9.1. Hence these theorems may not even be true if we are required to restrict attention to a computable sequence of computable selection rules. That is, there may actually exist two computable forecasting systems, each of which is calibrated for a with respect to every selection rule in our computable sequence, but which do not asymptotically produce identical forecasts. This would then violate metacriterion M4. If, on the other hand, one were willing to allow a noncomputable ordering of the selection rules for checking calibration sequentially over time, then there would be noncomputable forecasters who make

use of the ordering of selection rules (sequentially over time), and Theorems 7.1 and 9.1 would not apply to these forecasters.

Another problem which arises in finite samples is that of the prognosticating bum in Example 4.3. How many correct forecasts in a row will we need to see before we start paying more attention to the bum than to the local weatherman? A significance test would be particularly useless here because it takes no account of the presumably low prior probability that the bum's coin is actually related in any meaningful way to the weather. In fact, a problem more serious than that of the bum occurs in finite sample problems. This was illustrated in Example 3.2. In that case we saw that, even if there is no bum on the streetcorner, there will always be infinitely many computable forecasters who have predicted perfectly so far. It may be that none of these will be calibrated, however.

Since Professor Dawid's results depend too heavily on the infinitary aspects of the theory and cannot be "suitably interpreted for finite outcome sequences," this discussant chooses an alternative approach to comparing forecasters. This approach is described in more detail by Schervish (1983). Briefly, the approach is to assume that each forecast is to be used as the probability of the event being forecast in a simple two-decision problem, with a fixed loss function. After each event occurs, one calculates the loss one would have incurred had one made the optimal decision based on the forecast, and then accumulates the incurred losses over time. Next, one repeats this process for every possible loss function (essentially a one-parameter family) and obtains the accumulated loss as a function of one variable (the loss-function parameter). The smaller this function is, the "better" the forecaster has performed. There is no need to consider an infinite sequence of possible forecasts. And the determination of who has performed better can be made at any (and every) time one wishes.

Another advantage of the alternative approach described above is that it allows comparisons of forecasters who use different information bases. And it does not penalize a forecaster for using a larger information base even if he/she turns out to be not calibrated with respect to the larger information base. Consider the following example.

EXAMPLE 5.1. On 100 days, it rained 50 times and was dry 50 times. One forecaster always assigned probability 0.5, because he was using an empty information base. He is clearly calibrated (in any finite-sample sense) with respect to the only selection rule allowed by his information base. The second forecaster, however, used a larger information base and forecasted 0.2 on 40 days and 0.8 the other 60 days. Of the 40 days on which she forecasted 0.2, it rained 10 days and was dry 30. Of the 60 days on which she forecasted 0.8, it rained 40 days and was dry 20. This second forecaster is neither calibrated with respect to the empty information base nor with respect to her larger base. In fact, she is not even computable with respect to the empty base. And yet it is fairly clear to see that, so long as all 100 days were equally important, the second forecaster was at least as good as the first in any sense one cares to name. In particular, she was better in the sense of Schervish (1983) and in the sense of every strictly proper scoring rule. It is true that she would have been even better had her 0.2 forecasts been

0.25 and her 0.8 forecast been 0.667, but how was she to know this ahead of time? And how does she know if the proportions will continue that way in the future?

What this example illustrates is that the information base can be more important to making good forecasts than calibration. The goodness criteria should allow comparisons across information bases as well as within bases. A forecaster can be bad for using a small information base even if he/she is unable to be calibrated with respect to a larger base.

The approach described above is not designed to say how well a forecasting system is likely to perform in the future, but rather only attempts to say how well it has performed in the past. Nor is any claim made that the best forecaster is in any way "objective," "correct," or "valid." In fact, the bum of Example 4.3 will perform well both by the above method and by any method (such as scoring rules) which considers only the actual forecasts and outcomes. One area where further research is needed is in formulating models and criteria on how to decide which forecasts are likely to be better in the future given data available from the past. The problems of comparing forecasters based on past data and assessing their potential to forecast future data are clearly quite distinct as Examples 3.2 and 4.3 illustrate.

6. Conclusions. Professor Dawid has given us many interesting propositions to think over. However, the thought we should leave with is that *there are no objective or empirically valid probability forecasts.* The existence of calibrated forecasters is not guaranteed in general, and even when they exist, not a single event in the entire sequence a has an objective probability associated with it. Theorems 7.1 and 9.1 do not even imply that each computable subsequence of a has a unique limiting average probability. All they show is that if a calibrated forecaster exists, every other calibrated forecaster eventually agrees with him/her, but it always takes forever for all calibrated forecasters to agree. In the meantime, it is unknown (even unknowable) which of them are giving the calibrated forecasts. Hence, there is no time, after which we can be sure that we are receiving calibrated forecasts, even if we know who the calibrated forecasters are. This realization invalidates Professor Dawid's claim that "An attempt to make inferences about these objective probabilities is therefore justified to the extent that it is a hunt for something which does, at least, have a unique existence, at any rate asymptotically." It appears that Professor Dawid is attempting to resurrect objectivist/frequentist statistics within the Bayesian paradigm. Of course, this attempt fails for the same reasons that all other attempts to lay a foundation for objectivist/frequentist statistics have failed, namely, there are no objective or frequency-based probabilities on which to found such a theory.

On the other hand, some of Professor Dawid's concepts may have a role to play in creating models and criteria for deciding which forecasters are likely to provide good forecasts in the future. So long as one keeps in mind that the most a probability forecast can be is a measure of how strongly one believes that an event will occur (based on evidence currently available, not based on evidence yet to be observed), some of the concepts presented in Professor Dawid's paper may

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-