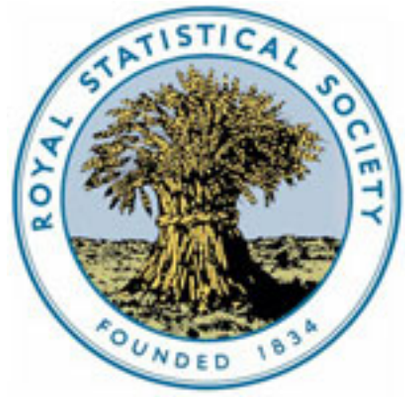

J. R. Statist. Soc. A (1979),
142, *Part* 2, *pp.* 146–180

On the Reconciliation of Probability Assessments

D. V. LINDLEY, A. TVERSKY and R. V. BROWN

University College London *Stanford University* *Decision Science Consortium, Inc.*

[Read before the ROYAL STATISTICAL SOCIETY on Wednesday, January 17th, 1979, Professor J. F. C. KINGMAN in the Chair]

SUMMARY

This paper investigates the question of how to reconcile incoherent probability assessments, i.e. assessments that are inconsistent with the laws of probability. A general model for the analysis of probability assessments is introduced, and two approaches to the reconciliation problem are developed. In the internal approach, one estimates the subject's "true" probabilities on the basis of his assessments. In the external approach, an external observer updates his own coherent probabilities in the light of the assessments made by the subject. The two approaches are illustrated and discussed. Least-squares procedures for reconciliation are developed within the internal approach.

Keywords: PROBABILITY ASSESSMENTS; CALIBRATION; COHERENCE; BAYES THEOREM; LEAST SQUARES; PRECISION; MIXTURES OF DISTRIBUTIONS; LOG-ODDS

1. INTRODUCTION

DECISIONS often depend on the probability of uncertain events such as the result of an election, the state of the economy, the outbreak of war, the guilt of a defendant or the outcome of a medical operation. Because such events are essentially unique, the assessment of their probability must be based on personal judgement. Thus, the human mind is used as an instrument for the assessment of uncertainty, much as the ruler, the pan balance and the pendulum are used as instruments for the measurement of length, weight and time. In fact, the axiomatic analysis of subjective probability (Ramsey, 1931; Savage, 1954; de Finetti, 1974) is closely related to the axiomatic analysis of the measurement of physical attributes like length, weight and time. (See Krantz *et al.*, 1971.)

Despite the formal similarity between the analyses, the measurement of subjective probability is considerably more problematic and less satisfactory than the measurement of physical attributes such as length or mass and psychological attributes such as loudness or brightness. Indeed, the assessment of subjective probability is beset with severe problems of both theoretical and practical nature. First, subjective probability is a measure of degree of belief, which reflects one's state of information. It is not only subjective but also variable since it can change from one situation to another. Second, it is not possible in general to obtain repeated independent measurements of subjective probability from the same individual because he is likely to remember his previous thoughts and responses. Consequently, there are no procedures for the measurement of belief that permit the application of the law of large numbers to reduce measurement errors.

The difficulties involved in applying standard measurement criteria of reliability and validity to the measurement of belief give rise to the questions of how to evaluate and improve assessments of subjective probability. Three types of criteria that could be called pragmatic, semantic and syntactic have been employed. Pragmatic tests refer to comparisons of assessments with reality, and they are applicable whenever the assessed probability of an event (e.g. a royal flush in poker, or an accident on the highway) can be meaningfully compared to a value that is computed in accord with the probability calculus, or estimated from empirical data. Unfortunately, such tests cannot be applied in most cases of interest because of the difficulties involved in estimating so-called objective probabilities.

If pragmatic tests are not applicable, however, it is still possible to evaluate probability assessments in terms of a semantic criterion that pertains to the meaning of the probability scale. Clearly, there is no way of validating, for example, a meteorologist's single judgement that the probability of rain is 2/3. If the meteorologist is using the scale properly, however, we would expect that rain would occur on about two-thirds of the days to which he assigns a rain probability of 2/3. This criterion is called calibration. Formally, a person is calibrated if the proportion of correct statements, among those that were assigned the same probability, matches the stated probability, i.e. if his hit-rate matches his confidence. If only half of the days to which the meteorologist assigned a rain probability of 2/3 were rainy, then he is not calibrated. This does not mean that his assessments are worthless or non-informative. The relationship between subjective probability assessments and calibration will be discussed below (Section 2.3).

Besides the pragmatic and the semantic criteria, subjective probabilities should also obey syntactic rules; that is, the relations between assessments should be governed by the laws of probability. For example, if A and B are disjoint events, then the assessed probability of the event, A or B, should be equal to the sum of the assessed probabilities for A and for B. A set of probability assessments is (internally) coherent only if it is compatible with the probability axioms. Coherence is clearly essential if we are to treat assessments as probabilities and manipulate them according to probabilistic laws.

Common observations and experimental studies (see Tversky and Kahneman, 1974; Slovic *et al.*, 1977) show that the pragmatic, semantic and syntactic criteria are not always satisfied. Thus, assessments of probability, produced by laymen and experts alike, are often inaccurate, uncalibrated and incoherent. But since subjective judgements constitute the major data base for the measurement of uncertainty, the question is not whether to accept subjective judgements at face value or reject them altogether, but rather how to debias and improve them. Procedures for the elicitation and debiasing of subjective probabilities have been discussed by several authors, among them Winkler (1969), Spetzler and Stael von Holstein (1975) and Kahneman and Tversky (1978). These procedures were designed to obtain probability assessments that are more accurate and better calibrated. The present paper is concerned with the problem of coherence, namely, how to reconcile probability assessments that are incoherent or mutually inconsistent. Before we formalize the problem, it is instructive to examine some examples.

Example 1

Consider the possible causes of death. Suppose H denotes heart failure, C denotes cancer, D denotes any other form of natural death and N denotes all causes of natural death. Let $\bar{N}$ denote the complement of N, etc. The following probability assessments, denoted by q, were actually made by one of the authors concerning the possible causes of death of another author:

$$q(H) = 0{\cdot}33, \quad q(C) = 0{\cdot}27, \quad q(D) = 0{\cdot}23, \quad q(\bar{N}) = 0{\cdot}12.$$

$$q(H|N) = 0{\cdot}41, \quad q(C|N) = 0{\cdot}31, \quad q(D|N) = 0{\cdot}28.$$

Note that these assessments are incoherent. First, $q(H)+q(C)+q(D)+q(\bar{N}) = 0{\cdot}95$ instead of unity. Such failures of additivity are quite common when the number of events exceeds 2 or 3. Second, the conditional probability ratios do not coincide with the ratios of the corresponding unconditional probabilities. For example, $q(H|N)/q(C|N) = 1{\cdot}32$, while $q(H)/q(C) = 1{\cdot}22$. Clearly, the two ratios should be equal because both H and C are subsets of N. The assessor is now faced with the task of reconciling his assessments so as to achieve coherence. How should he do it? What additional information, if any, is required to reconcile the inconsistent assessments?

Example 2

Let C denote the occurrence of a major energy crisis in the United States during the next decade, and let q_1 and q_2 be two different assessments of the probability of C. These values may emerge from two different ways of thinking about the problem, say, one in terms of specific circumstances that could lead to an energy crisis, and one in terms of a particular economic model. Alternatively, q_1 and q_2 may represent the judgements of two experts about the probability of C. In general, of course, q_1 and q_2 do not coincide. Thus, we have to amalgamate the two estimates, that is, assess the probability of C in the light of q_1 and q_2, denoted $p(C|q_1, q_2)$. Note that from a purely formal viewpoint, it is immaterial whether the two estimates were produced by one person using two different methods, or by two different individuals. In both cases we need to reconcile the difference and produce a single estimate.

The problem of reconciling inconsistent observations is not unique to the measurement of belief. For example, a surveyor who uses a theodolite to measure distance faces a similar problem. Because of measurement errors, the assessments of angle and distance are generally incompatible with the laws of plane geometry. Hence, the surveyor must reconcile the inconsistent measurements to obtain a coherent set of estimates. His problem, however, is simpler because he can readily obtain repeated observations and thereby reduce errors of measurement. Although it is generally not possible to obtain independent repeated measurements of subjective probability, the analogy between the measurement of distance and the assessment of belief is instructive. In particular, it suggests the possibility of exploiting the constraints imposed by the probability laws to obtain improved estimates of subjective probability, much as the surveyor exploits the constraints imposed by plane geometry.

To illustrate this idea, suppose you have to assess, as in Example 2, the probability of a major energy crisis in the United States during the next decade, denoted C. There are several different approaches to the problem. You could adopt an intuitive, wholistic approach where you contemplate the energy situation, the United States economy, the international scene, etc., and make an intuitive estimate on the basis of these considerations. Alternatively, you might wish to develop an explicit model for the supply and demand of energy, in which $p(C)$ can be expressed as a function of some parameters that either are known, or can be estimated. A third possibility, which lies somewhere between wholistic assessment and explicit modelling, is to decompose $p(C)$ and assess the components separately. For example, let E denote an oil embargo on the United States during the relevant time period. Following the decomposition approach, you could assess the probability of an energy crisis given an oil embargo $p(C|E)$, the probability of an energy crisis in the absence of an oil embargo $p(C|\bar{E})$ and the probability of an oil embargo $p(E)$. The overall probability of an energy crisis is given by

$$p(C) = p(C|E)p(E) + p(C|\bar{E})\{1 - p(E)\}.$$

Thus, the laws of probability allow us to compute $p(C)$ from $p(E)$ and the conditional probabilities $p(C|E)$ and $p(C|\bar{E})$, just as the laws of geometry allow us to compute the distance between A and B, say, from the distance between A and X, the distance between B and X, and the angle AXB. Just as the distance between A and B can be measured using a different auxiliary point from X, the probability of C can be computed using a different conditioning event from E. Thus, one may compute $p(C)$ from $p(C|D)$, $p(C|\bar{D})$ and $p(D)$, where D denotes the development of new effective methods for using solar energy.

The use of different conditioning events such as D and E to compute $p(C)$ can be viewed as two different ways of thinking about the target event C. A direct wholistic assessment of C represents a third way of looking at the problem. Generally, the different procedures yield different estimates of $p(C)$ that have to be reconciled. If each of the estimates, however, conveys some valid information that is not included in the others, then the precision associated with the reconciled value will be greater than that of the separate estimates, in the same way

that the precision in the location of a point increases by determining several different (inconsistent) bearings.

The present paper is concerned with the development of models for the reconciliation of incoherence which extend earlier ideas (Brown and Lindley, 1978; Good, 1952, 1971). We first outline a general framework for the analysis of probability assessment; we then investigate how it can be used to reconcile inconsistent judgements. We develop two approaches to the reconciliation problem, which we label internal and external. In the internal approach, the observed probability assessments are related to internal coherent probabilities in a manner analogous to the relation between the observed score and the true score in test theory. Thus, it is assumed that the subject has, in some sense, a set of coherent probabilities that are distorted in the elicitation process. The internal approach is concerned, then, with the attempt to estimate the underlying "true" probabilities using the observed assessments. This approach also permits the calculation of the precisions associated with the reconciled values, given the precisions associated with the observed assessments. The external approach does not explicitly address any "true" probabilities, but concerns itself only with deriving coherent probabilities based on the original set of incoherent assessments.

In both approaches we introduce an investigator who assesses his own coherent probabilities on the basis of the judgements produced by the subject. Here, the investigator plays a role that is similar to that of a surveyor who uses fallible measurements to produce a coherent set of distances.

In the next section we develop a general framework for the analysis of fallible probability assessments, and introduce the internal and the external approaches to the reconciliation problem. Section 3 illustrates the application of the basic model in a simple case. Least-squares procedures for the internal approach are discussed and illustrated in Section 4. The philosophical and practical problems associated with the present development are discussed in Section 5, along with directions for future research.

2. THE BASIC MODEL

2.1. *Description of the Model*

We are concerned with an individual or a subject, denoted S, who considers a sequence $A = (A_1, A_2, \ldots, A_m)$ of events about which he is uncertain. For example, consider a meteorologist contemplating the rain pattern for the next m days where A_i denotes rain on the ith day.

We suppose that S wishes to describe his uncertainty about A through a coherent probability specification for A. S therefore has a probability distribution $\pi(A)$ for A. Real S is not necessarily coherent and his assessments of probabilities do not always obey the rules of the calculus; even if they do, they may be defective because of S's weakness as a probability appraiser. Thus, in the meteorological example S may assess the probability of rain on day 1 as 0·4, on day 2 given that day 1 is wet as 0·8, and yet say that the probability of rain on both days is 0·2, and not 0·32 as implied by the above assessments and the demands of coherence. The assessed values will be described by a vector $q(A)$.

Our model therefore contains, in addition to S, three elements: A, $\pi(A)$ and $q(A)$. The first describes the world external to S, the second describes a coherent S and the third gives S's stated view of the world. In terms of the analogy with the measurement of length, q corresponds to the observed measurements of distances and angles, and π could be regarded as the true distances between the points. Like their physical counterparts, q is directly observable but π is not.

In addition to the subject, S, we consider an investigator, N. Unlike S, N is coherent and his task is to reconcile S's stated values q and to provide an assessment of π. Alternatively, N could assess his own probabilities for A in the light of the information q provided by S.

N can be thought of as the surveyor who uses Euclidean geometry to provide estimates of the true positions except that he uses the probability calculus instead of geometry.

From N's viewpoint, all the elements of the model, S, A, π and q are part of the external world about which he is uncertain and which are described by a probability distribution $p(A, \pi, q)$ of the uncertain quantities. (We prefer Schlaifer's term "uncertain quantity" to the more usual "random variable" because the expressions are fixed and not variable; but they are unknown and hence uncertain.) The notation $p(\cdot)$ will be reserved for N's probabilities: the Greek equivalent $\pi(\cdot)$ similarly refers to S's coherent probabilities; a different letter, q, is used for S's assessments which need not be coherent.

The joint distribution may conveniently be described in three stages. First, there is the distribution of A, $p(A)$. Second, there is the conditional distribution of π given A, $p(\pi|A)$; third, there is the conditional distribution of q given the other two elements, $p(q|\pi, A)$. These three distributions completely specify the joint distribution of the uncertain quantities and summarize the situation as far as N is concerned. Notice that each of the three distributions corresponds to a different aspect of N's contemplation of the problem. His view of the world external to both N and S is described by $p(A)$; whereas S's, when coherent, is $\pi(A)$. His view of coherent-S's knowledge of A is included in $p(\pi|A)$. Finally, $p(q|\pi, A)$ gives N's opinion of S as a probability appraiser. In the meteorological example, $p(A_1)$ is N's probability for rain on the first day; $p\{\pi(A_1)|A_1\}$ is N's probability that the meteorologist's true probability of rain on the first day is π when it truly will rain then. The final conditional distribution describes what N thinks the meteorologist will actually state when his true value is $\pi(A_1)$ and it will rain.

Since q differs only from π because of S's difficulties in articulating his probabilities and therefore describes S as a measuring instrument on π, it seems reasonable to suppose that the conditional distribution of q, given π, does not depend on the state of the world external to S described by A. That is, given π, q and A are independent, or

$$p(q|\pi, A) = p(q|\pi). \tag{1}$$

Indeed, the major function of the unobservable "true" probability π is to stand between S and the external world so that the two are related only through π. This assumption is in the spirit of the ttandard measurement model, where different measurements of the same quantity are treated as independent—given the "true" value of the quantity.

In summary, our model involves

I. $p(A)$: N's appreciation of the world

II. $p(\pi|A)$: N's opinion of S's knowledge of the world.

III. $p(q|\pi)$: N's opinion of S as a probability assessor.

These core distributions are N's, and all the calculations with them are performed by N according to the rules of the probability calculus, that is, coherently. We now describe two approaches to the reconciliation problem, called the internal and the external approaches. In the internal approach N is concerned with π, in the external approach with A; both in the light of S's stated q, π being internal, and A external, to S.

2.2. *The Internal Approach*

Here N is concerned only with π and q. We have†

$$p(\pi) = \sum_A p(\pi|A)\, p(A) \tag{2}$$

† The notation Σ_A means a summation over a partition of events in A. Thus if $A = (A_1, A_2)$ then the summation is over A_1A_2, $A_1\bar{A}_2$, $\bar{A}_1A_2$ and $\bar{A}_1\bar{A}_2$.

and $p(q|\pi)$ directly, providing a complete probabilistic description of q and π. By Bayes' Theorem we have

$$p(\pi|q) \propto p(q|\pi)p(\pi) \tag{3}$$

as N's appraisal of coherent S after S has reported his assessments q.

This method can be used when S has made several probability assessments $q = (q_1, q_2, \ldots, q_m)$, finds them to be incoherent—as in Example 1 or the meteorological problem above—and wishes to reconcile them to coherent values. Let $(\pi_1, \pi_2, \ldots, \pi_m)$ be the true values that correspond to $(q_1, q_2, \ldots, q_m)$, respectively. There will typically be constraints among the π_i's corresponding to coherence requirements. Thus, in the meteorological problem with $q_1 = q(A_1) = 0{\cdot}4$, $q_2 = q(A_2|A_1) = 0{\cdot}8$ and $q_3 = q(A_1A_2) = 0{\cdot}2$ we will have $\pi_3 = \pi_1\pi_2$. A possible set of reconciled values is given by $\hat{\pi}_i = E(\pi_i|q)$, the means of the distribution $p(\pi|q)$. In statistical language, the $\hat{\pi}$'s are estimates of the π's. The precision† of the reconciled value may be described by the inverse of the variance of π_i given q. Notice that, in general, the covariances will not be zero and the reconciled values will not be independent.

A special case arises when one or more of the events in A are of particular interest—we call them target events—and the other events are introduced in order to increase the precisions. It is then only necessary to calculate the marginal distribution of these π's which refer to the target events. We call this procedure "extension of the conversation"—from the target events to other events—enlarging its usual use. The increase in precision essentially results from the increased exposure S has to the constraints of coherence when he contemplates many events. Example 2 above provides an illustration of his procedure. We believe that this method could be of considerable value in improving probability assessments. The calculations are described in detail in Section 4.2.

2.3. *The External Approach*

Here N is concerned only with q and A, so that S plays no role except as the provider of data q for N to update his probabilities of A. We have‡

$$p(q|A) = \sum_{\pi} p(q|\pi)\,p(\pi|A), \tag{4}$$

using the independence condition (1), and $p(A)$ is available directly providing a complete probabilistic description of q and A. By Bayes' Theorem we have

$$p(A|q) \propto p(q|A)p(A) \tag{5}$$

as N's appraisal of the uncertain events after S has reported his assessments q.

The external approach could be useful where N is a decision maker who requires a probability for A in order to take action. He may consult an expert, or a group of experts, each of whom reports his probability for A and the decision maker has to reach an overall judgement (see Example 2).

Notice that although both approaches depend on the common core of distributions I–III above, $p(A)$, $p(\pi|A)$ and $p(q|\pi)$, they can be used without N determining all three. Thus, in the internal approach, $p(\pi)$ may be assessed directly, rather than through (2), when combination with $p(q|\pi)$ in (3) gives the required result. Similarly, in the external approach, $p(q|A)$ may be assessed directly, rather than through (4), and combined with $p(A)$ in (5) to produce the result. If these ideas are adopted, the internal approach is seen to be simpler than the external one because it avoids the second of the core distributions, $p(\pi|A)$, or the derived $p(q|A)$.

† Precision, so defined, is attractive because, in normal situations, it is additive: see the definition of w in Section 3.

‡ We use Σ_π to denote a summation over π. In application below it will be an integration over real vector space.

Both these distributions are relatively unfamiliar, even to a Bayesian, in comparison with I and III, which are respectively a "prior" and a likelihood. However, II requires a judgement by N of what S believes about A (in the case of π) or will say he believes (for q) for each constituent event in A.

In the weather example, N has to consider what the meteorologist might say about rain tomorrow both on those occasions when it will rain, and on those when it will not. Presumably, for a reputable weather forecaster, the former probability will be the higher. The distributions $p(\pi|A)$ and $p(\pi|\bar{A})$ are measures of the quality of the expert, S. Note that $p(q|A)$ is related to the idea of calibration discussed in Section 1. To see the relation, consider a sequence of occasions on which S asserted a value q for the probability of an event and let $f(A|q)$ denote the relative frequency with which the events in the sequence occur. In the frequentist view of probability $f(A|q)$ is closely related to $p(A|q)$ which, in turn, is related to $p(q|A)$ by the one-dimensional form of (5). Recall that a person is calibrated if $f(A|q) = q$; that is, if a proportion q of the events to which he has assigned a probability q actually occur. Indeed, if $p(A|q) = q$, N will take S's announced value, q, to be his probability for A.

We now add a remark about the internal approach in its direct form using $p(\pi)$ and $p(q|\pi)$ to obtain $p(\pi|q)$. In this form, π plays the role of a set of parameters, q is data, and the "prior" $p(\pi)$ is updated by the likelihood $p(q|\pi)$ to give a posterior $p(\pi|q)$. The use of reconciled values $\hat{\pi} = E(\pi|q)$ with their associated precisions is closely related to the method of least squares. Suppose the data $q_1 \ldots q_m$ are, given the π's, independent and normally distributed about the π's with constant variance; suppose further that the prior for the π's is, relative to this likelihood, rather smooth. It then follows by standard theory that the π's given the q's are also normal. The means are the least-squares estimates obtained by minimizing $\Sigma_i(q_i - \pi_i)^2$ over the π's, subject to the coherence constraints. (This result is exact if the constraints are linear and will be approximate for non-linear constraints.) The matrix of second derivatives of the sum of squares at the minimum when inverted gives approximate variances and covariances. If the q's are not independent but have a general normal distribution, then a weighted sum of squares and products replaces the direct sum above. Because the q's are bounded, they cannot be normally distributed so that it may be preferable to convert them to log-odds, $\ln\{q/(1-q)\}$. The whole argument then goes through with log-odds instead of probabilities, both for the q's and the π's.

This technique is particularly simple and is the most usable method we have for reconciliation of probabilities. It is discussed in some detail in Section 4. Notice that all it requires, in addition to S's provision of the data, q, are their variances and covariances since these completely describe the normal likelihood, and the prior is assumed to be "flat". In our formulation, the likelihood $p(q|\pi)$ has been thought of as N's but S may provide his own second moments. If these are used directly by N, our argument is unaffected.

3. TECHNICAL DEVELOPMENTS

In this section, we take the general model of the previous section, insert specific forms for the core distributions and calculate other distributions of interest. To avoid excessive technicalities, we confine ourselves to the simplest case to demonstrate the feasibility of the model. In Section 4, by specializing and using least-squares ideas, we come nearer to results of practical use.

In the case considered there is a single event, A, for which S has true probability $\pi(A)$, or simply π, and for which he reports the single value $q(A)$, or q. Hence all quantities are one-dimensional. With one value reported, there is no opportunity to use coherence; nevertheless, reconciliation, in the sense of calculating $\hat{\pi}$, might be appropriate depending on N's judgement about S expressed through the core distributions I–III. In any case, N may need to calculate his probability of A in the light of S's reported value q, as is the case when the weather forecaster reports the probability of rain tomorrow to be q. Another example arises in medical

diagnosis: a physician, S, reports the probability q that N has appendicitis. What is N's probability for appendicitis?

To specify the model completely, we need to describe $p(A)$, $p(\pi|A)$, $p(\pi|\bar{A})$ and $p(q|\pi)$. It is convenient and sensible to work with log-odds (see above) rather than with probabilities. To economize on notation, we use q and π to denote log-odds rather than probabilities. The log-odds for A is written $lo(A)$. Clearly, the general theory so far discussed is unaffected by this change.

There are two reasons for changing to log-odds. First, it is necessary to handle bivariate distributions, and the normal is computationally the most attractive; so variables, like log-odds, with infinite ranges are to be preferred. Second, it is more reasonable to suppose that the measurement error has constant variance when expressed in log-odds rather than in probabilities, since values of the latter are, in absolute terms, more precisely assessed when near 0 or 1 than when around $\frac{1}{2}$.

We therefore suppose that the last of our core distributions $p(q|\pi)$ is, in log-odds, normal with mean π and constant variance, σ^2, abbreviated to $N(\pi, \sigma^2)$. That is, N views S's measurement of log-odds as unbiased with constant variance. With $p(A) = \alpha$, say, we have only to specify $p(\pi|A)$ and $p(\pi|\bar{A})$. Suppose the former is $N(\mu_1, \tau^2)$. That is, if it really is going to rain tomorrow, then N expects S to have log-odds μ_1, with standard deviation τ. Similarly, suppose the latter (applying to the case of no rain) is $N(\mu_2, \tau^2)$. Presumably, for a good forecaster, $\mu_2 < 0 < \mu_1$. (Log-odds of zero correspond to a probability of $\frac{1}{2}$.)

A special case arises when S is thought to be just as good when A is true as when it is false: then $\mu_2 = -\mu_1$, for when A is true, his probability for A would then be expected to be the same as that for $\bar{A}$ when $\bar{A}$ is true, but $p(\bar{A}) = 1 - p(A)$ and hence $lo(\bar{A}) = -lo(A)$. This may be appropriate in the meteorological case but not in the appendicitis example, where it is easier to diagnose a real case of appendicitis than it is other sources of abdominal pain, so that perhaps $|\mu_2| < \mu_1$. The variance has been supposed the same for A and for $\bar{A}$: it is possible to handle the more general case, but the algebra is untidy and little extra insight is gained. Notice how $p(\pi|A)$ and $p(\pi|\bar{A})$ together describe N's opinion of S as a probability appraiser. The values are related to the errors of the two kinds studied in statistics, or to the false positives and false negatives considered in medicine.

In summary, the model is as follows:

I. $p(A) = \alpha$.

II. $p(\pi|A) \sim N(\mu_1, \tau^2), p(\pi|\bar{A}) \sim N(\mu_2, \tau^2)$.

III. $p(q|\pi) \sim N(\pi, \sigma^2)$.

Consider first the internal approach. We have, equation (2),

$$p(\pi) = p(\pi|A)p(A) + p(\pi|\bar{A})p(\bar{A}) = \alpha N(\mu_1, \tau^2) + (1-\alpha) N(\mu_2, \tau^2), \tag{6}$$

a weighted average of two normals. For sufficiently small values of τ^2 the distribution is bimodal. Thus, good meteorologists (with small standard deviations) typically have probabilities for rain which are either high (when it is going to rain) or low (when not), and only rarely are they so undecided as to give values around $\frac{1}{2}$, log-odds of zero.

It is convenient to write $p(\pi|A) = p_1(\pi)$ and $p(\pi|\bar{A}) = p_2(\pi)$. Then, by Bayes' theorem,

$$p(\pi|q) = p(q|\pi)p(\pi)/\sum_{\pi} p(q|\pi)p(\pi) = \frac{p(q|\pi)\{\alpha p_1(\pi) + (1-\alpha)p_2(\pi)\}}{\alpha p_1(q) + (1-\alpha)p_2(q)}$$

$$= p_1(\pi|q)\alpha' + p_2(\pi|q)(1-\alpha'), \tag{7}$$

where

$$p_i(q) = \sum_{\pi} p(q|\pi)p_i(\pi), \quad p_i(\pi|q) = p(q|\pi)p_i(\pi)/p_i(q) \quad \text{and} \quad \alpha' = \alpha p_1(q)/\{\alpha p_1(q) + (1-\alpha)p_2(q)\}.$$

From standard normal theory, $p_i(\pi|q)$ is normal with mean $wq+(1-w)\mu_i$ and variance $w\sigma^2$, where $w=\tau^2/(\sigma^2+\tau^2)$; also $p_i(q)\sim N(\mu_i, \sigma^2+\tau^2)$, so that $p(\pi|q)$ is also a mixture of normals. The mean is

$$\hat{\pi} = E(\pi|q) = wq+(1-w)\{\alpha'\mu_1+(1-\alpha')\mu_2\}. \tag{8}$$

This is the reconciled value of π on the basis of the stated value q. It is near q if w is near one, that is if τ^2 is much greater than σ^2. A large value of τ means that $p(\pi)$ [equation (6)] has large spread and that the likelihood $p(q|\pi)$ with smaller spread σ is dominant. This is the situation discussed in connection with least-squares methods in Section 2.3. Consequently, if S has small measurement error compared with appraisal error, the reconciled value will be essentially the stated value, and no change will occur. The variance, var $(\pi|q)$, of the reconciled value can be found as the familiar variance of a mixture. It is an untidy expression; but in the case where τ^2 is much larger than σ^2, it is approximately equal to σ^2.

A numerical example when τ^2 is of the same order as σ^2, so that reconciliation away from q may take place, is instructive. Suppose $\alpha=\frac{1}{2}$, $\mu_1=-\mu_2=1{\cdot}0$ (so that S is equally competent whether A is true or false), $\tau=1{\cdot}0$ and $\sigma=0{\cdot}5$. At two standard deviations, when A is true, S is anticipated to have true log-odds between $-1{\cdot}0$ and $3{\cdot}0$ (probabilities between 0·27 and 0·95) and when false between $-3{\cdot}0$ and $1{\cdot}0$ (0·05 and 0·73 for probabilities). But the reported odds can differ by as much as $1{\cdot}0$ $(=2\sigma)$ from the true values. (A probability of 0·5 can be reported anywhere in the range 0·27 to 0·73.) Then $w=0{\cdot}8$ so that in $\hat{\pi}$, equation (8), 80 per cent of the weight goes on q and 20 per cent depends on the anticipated performance of S. If $q=0{\cdot}7$ (stated probability of 0·67), the weight $\alpha'=0{\cdot}75$ and $\hat{\pi}$ is 0·66 (a reconciled probability of 0·66). If $q=1{\cdot}5$ (probability 0·82), the weight $\alpha'=0{\cdot}92$ and $\hat{\pi}$ is $1{\cdot}37$ (probability 0·80). Both probabilities are lowered slightly, the mean of the prior distribution of π being zero. But the changes are small, and even here, where τ is only twice σ, the approximation that assumes τ is large is not unreasonable. In this case, reconciliation is merely allowing for the effect of measurement error.

As an intermediary between the internal and external approaches, we can consider $p(A|\pi)$, which is a special case of the external approach as $\sigma^2\to 0$. We have

$$\frac{p(A|\pi)}{p(\bar{A}|\pi)} = \frac{p_1(\pi)}{p_2(\pi)}\frac{\alpha}{1-\alpha}$$

and hence

$$lo(A|\pi) = lo(A)-(\mu_1^2-\mu_2^2)/2\tau^2+(\mu_1-\mu_2)\pi/\tau^2. \tag{9}$$

For this to be π, two conditions must be fulfilled, namely

$$(\mu_1-\mu_2)/\tau^2=1 \quad \text{and} \quad \ln\{\alpha/(1-\alpha)\} = (\mu_1^2-\mu_2^2)/2\tau^2.$$

These are equivalently

$$(\mu_1-\mu_2)/\tau^2=1 \quad \text{and} \quad \tfrac{1}{2}(\mu_1+\mu_2) = \ln\{\alpha/(1-\alpha)\}. \tag{10}$$

The second condition says that the prior log-odds have to equal the average performance. This is satisfied in the special case $\mu_2=-\mu_1$ when $\alpha=\frac{1}{2}$. The first condition says that the difference between the expected performances under the two conditions, A and $\bar{A}$, must equal the appraisal variance τ^2. Thus, a complete matching of N and coherent-S when considering A depends on a rather complex combination of prior views by N and S's appraisals. A change in the former would affect this agreement. If N judges A to be as likely as not, $\alpha=\frac{1}{2}$, the conditions for agreement are that $\mu_2=-\mu_1$, so that S performs equally well under A as $\bar{A}$, and $2\mu_1/\tau^2=\lambda$, say is one. No simple interpretation of this final requirement is known to us.

Consider next the external approach. It is immediate from the core distributions II and III that

$$p(q|A)\sim N(\mu_1, \sigma^2+\tau^2) \quad \text{and} \quad p(q|\bar{A})\sim N(\mu_2, \sigma^2+\tau^2), \tag{11}$$

so that the replacement of π (in II) by q here merely replaces τ^2 by $\sigma^2+\tau^2$. Consequently, $p(A|q)$ is given by the same result as (9) but with τ^2 everywhere replaced by $\sigma^2+\tau^2$ and π by q.

In the numerical illustration with $\alpha = \frac{1}{2}$, $\mu_1 = -\mu_2 = 1{\cdot}0$, $\tau = 1{\cdot}0$ and $\sigma = 0{\cdot}5$, so that $\lambda = 2$, the log-odds given q are, from (9) with $\sigma^2+\tau^2$ replacing τ^2 and q for π, equal to $(\mu_1-\mu_2)q/(\sigma^2+\tau^2)$, or 1·6 times q. Hence q of 0·7 (probability 0·67) is raised to q of 1·12 (probability 0·75) and a q of 1·5 (probability 0·82) to one of 2·4 (probability 0·92). These changes are much larger than those caused by reconciliation. The intuitive explanation for this is most easily seen by considering the higher value of q, 1·5. Such a value is much more likely to have arisen from the distribution when A is true, $p_1(q) \sim N(1, 1{\cdot}25)$ rather than from the distribution when A is false, $p_2(q) \sim N(-1, 1{\cdot}25)$. This is described by α' which is here 0·92, so that there is a 92 per cent chance that q came from A. Consequently, N can substantially increase S's stated value. Notice how this change depends heavily on S as an expert, whereas reconciliation does not, $p(\pi|A)$ and $p(\pi|\bar{A})$ entering only through $p(\pi)$. If τ increases, the effect on N will become less. In an extreme case where τ and σ tend to zero, N will shift S's stated value to either 1 or 0. A meteorologist who always says 0·6 chance of rain on rainy days and 0·4 when it is to be dry entitles N to be sure of rain when 0·6 is announced.

This example shows that even the simplest case involves quite complicated calculations. Much of the complexity arises through the need to consider the second of the core distributions $p(\pi|A)$ and $p(\pi|\bar{A})$, reflecting S's expertise. While regretting the awkward nature of the results, it does seem that these distributions reflect an essential ingredient of the situation. In the next section we consider the internal approach where π is estimated from q and where $p(\pi)$ is "flat", $\tau^2 \to \infty$, and least-squares procedures can be used. Here the calculations are simpler and more readily related to practical requirements.

4. Least-squares Procedures with the Internal Approach

We begin with a general description of the procedure. We are concerned only with q and π, S's stated and coherent values for some events. N's opinion about π is diffuse so that $p(\pi)$ is approximately constant. In the simpler situations, each element q_i of q is approximately normally distributed about π_i, the corresponding element of π, with constant variance σ^2 say, these being independent. Under these circumstances, the reconciled values $\hat{\pi}_i = E(\pi_i|q)$ are given approximately by the values of the π's that minimize

$$\sum_i (q_i - \pi_i)^2$$

subject to any constraints on the π's that coherence imposes. In other situations, the q's will be correlated and have different variances, when a quadratic form

$$\sum w_{ij}(q_i - \pi_i)(q_j - \pi_j)$$

will be minimized subject to the constraints, the w's being weights, the elements in the inverse of the dispersion matrix of the q's. Minimization is performed by equating the first derivatives to zero. The matrix of second derivatives at the minimum, when inverted, provides the approximate variances, $\mathrm{var}(\pi_i|q)$, and covariances of the reconciled values. Throughout this argument, the π's and q's may be probabilities or some suitable transform of them, such as log-odds.

The whole procedure is straightforward except for one difficulty: the constraints imposed by coherence are typically non-linear, and the resulting equations are therefore non-linear. There is no simple resolution of this difficulty. The power of the probability calculus lies in the ability of probabilities to combine both additively and multiplicatively. If linearity of the latter is imposed by taking logarithms, the linearity of the former is destroyed. It is, therefore, a fundamental difficulty but one that can sometimes be alleviated by suitable approximations.

We now pass to the consideration of several special cases. In any application of the least-squares ideas, we have to specify whether the calculation is in terms of probabilities directly, or in some function thereof, such as log-odds. We refer to this as the choice of metric: probability metric, log-odds metric, etc. The reason for preferring one metric to another is, as before, that the variance of the q's may be judged to be more reasonably constant in one metric than in another.

4.1. *Partition*: *General Metric*

Here $(A_1, A_2, \ldots, A_n)$ is a partition, and S provides values $q_i = q(A_i)$, $1 \leqslant i \leqslant n$ for $\pi_i = \pi(A_i)$ with the single† coherence constraint $\Sigma\pi_i = 1$. For a general metric $F(\cdot)$—for log-odds, $F(t) = \ln\{t/(1-t)\}$—the expression to be minimized is $\Sigma\{F(q_i)-F(\pi_i)\}^2$ subject to the constraint $\Sigma\pi_i = 1$. The constraint can either be incorporated with $\pi_n = 1-\Sigma_{i=1}^{n-1}\pi_i$, or by using a Lagrangian. For the probability metric, the equations are linear and have the exact solution

$$\hat{\pi}_i - q_i = n^{-1}(1-\Sigma q_j) \tag{12}$$

with

$$\text{var}(\hat{\pi}_i) = (1-n^{-1})\,\sigma^2, \tag{13}$$

σ^2 being the variance of each value stated by S. The improvement in precision due to coherence is only appreciable when n is small; for $n = 2$ it doubles. The form of (12) is worth a comment since it generalizes. The adjustment, $\hat{\pi}_i - q_i$, of the stated value to a reconciled value is a multiple, here n^{-1}, of the degree of incoherence, that is, how far Σq_j differs from its coherent value, 1. Here each q_i is altered by the same amount.

A subject was asked to assess the probabilities that his ultimate death would fall under one of the five categories: cancer, heart disease, stroke, other natural, unnatural (accident, suicide, etc.). He gave the values given in the second column of Table 1.‡ The total is 0·83, exhibiting a fair degree of incoherence perhaps caused by the unpleasant nature of the events. The reconciled values are each increased by $0{\cdot}17/5 = 0{\cdot}034$ and are given in column 6.

In other metrics the minimization equations

$$\{F(q_i)-F(\hat{\pi}_i)\}\,F'(\hat{\pi}_i) = \lambda$$

in Lagrange form are non-linear. A first guess for π_i might be q_i with the approximate result that

$$\hat{\pi}_i = q_i - \lambda/F'(q_i)^2. \tag{14}$$

With the log-odds metric $F'(x) = x^{-1}(1-x)^{-1}$ so that

$$\hat{\pi}_i = q_i - \lambda q_i^2(1-q_i)^2 \tag{15}$$

providing a larger correction when q_i is near $\frac{1}{2}$ than elsewhere. In the medical example, the changes upwards are 0·06, 0·04, 0·01, 0·04 and 0·01, instead of a constant 0·034 in the probability metric. The reconciled values are given in column 7 of Table 1.

The reason for using other than the probability metric is the fact that not all stated probabilities may have the same errors associated with them. An alternative to changing the metric is for either S or N to provide the precision associated with each value. The precision of an assessment is closely related to its stability, i.e. the degree to which it varies upon further reflection. The assessment of the probability that a coin will come up heads, for example, is very precise: it is not likely to depart from 0·5.

† It is assumed throughout that S always gives values for the q's that lie in the unit interval.

‡ Column 8 gives the frequencies (unknown to S) for US residents. These values are not to be interpreted as the correct probabilities. For example, a person with a known history of heart trouble would have a higher probability than the frequency value.

TABLE 1

Subject's probabilities for various causes of death

Column: 1 *Cause of death*	2 *Stated values* q_i	3 *Ranges for* q_i	4 *Weights*	5 *Reconciled values* $\hat{\pi}_i$	6 $\hat{\pi}_i$ *Unweighted*	7 $\hat{\pi}_i$ *Log-odds*	8 *US frequencies*
Cancer	0·25	0·15–0·35	25	0·28	0·28	0·31	0·18
Heart disease	0·20	0·05–0·40	8	0·25	0·23	0·24	0·40
Stroke	0·10	0·05–0·25	25	0·12	0·13	0·11	0·11
Other natural	0·20	0·05–0·45	6	0·27	0·23	0·24	0·25
Unnatural	0·08	0·01–0·15	51	0·09	0·11	0·09	0·06
	0·83						

To elicit the precision associated with each of the q's, we asked the subject to quote a range of values for each assessed probability (column 3 of Table 1). The quoted ranges were interpreted as multiples of standard deviations and weights used inversely proportional to the variances (column 4). We now have to minimize $\Sigma w_i(q_i-\pi_i)^2$ with the result that

$$\hat{\pi}_i - q_i = (1-\Sigma q_j)/w_i \Sigma w_j^{-1}. \tag{16}$$

The results for the medical example are given in column 5 of Table 1.

4.2. *External Conditioning*

In Section 2.2 we mentioned the possibility of having some target whose probabilities are required and that, in order to assess these values, the conversation is extended to other events. Here we consider the simplest case of a single target event A and the extension to another event, denoted X to distinguish it clearly from a target event. Suppose S is sufficiently conscious of coherence so that his assessments for any pair of complementary events add to unity. Suppose further that S assesses $\pi(A)$, $\pi(A\,|\,X)$, $\pi(A\,|\,\bar{X})$ and $\pi(X)$. Here, the opportunity for incoherence arises since $\pi(A) = \pi(A\,|\,X)\pi(X)+\pi(A\,|\,\bar{X})\pi(\bar{X})$ may not hold for the corresponding q's. Our hope is that the reconciled value $\hat{\pi}(A)$ will be an improvement over the raw assessment $q(A)$, and we therefore wish to evaluate the precision of the reconciled value. The coherence constraint above is non-linear; to avoid this let us suppose $\pi(X)$ is known† so that $q(X) = \pi(X)$. We later generalize to the case where this is also in error. The process will be called *external conditioning*, since the probability of the target event is considered conditional on some external event X.

Since no new difficulties arise by generalizing, we consider a partition into $X_1, X_2, \ldots, X_n$ and not just X and $\bar{X}$, $n = 2$. Write $q(A\,|\,X_i) = q_i$, $q(A) = q'$, the π's analogously and $q(X_i) = \kappa_i = \pi(X_i)$, with $\Sigma\kappa_i = 1$. The coherence constraint is that $\pi(A) = \Sigma\pi(A\,|\,X_i)\pi(X_i)$ or $\pi' = \Sigma\pi_i\kappa_i$, and in the probability metric we have to minimize

$$\Sigma(q_i-\pi_i)^2+(q'-\Sigma\pi_i\kappa_i)^2, \tag{17}$$

on assuming equal weights and zero correlations for the q's. The result, for the target event A, is that

$$\hat{\pi}' - q' = \frac{\Sigma q_i\kappa_i - q'}{1+\Sigma\kappa_i^2}. \tag{18}$$

Stated in words, the correction to q' for incoherence is the departure from coherence, $\Sigma q_i\kappa_i - q'$,

† An example is where X is the event that a team wins the toss at the beginning of a contest. It would then generally be agreed that $q(X) = \pi(X) = p(X) = \frac{1}{2}$. The event A could then be that the team wins the contest.

divided by $(1+\Sigma\kappa_i^2)$. This divisor is the variance of the departure, assuming all q's have unit variance. The precision of the reconciled value is therefore easily found to be

$$1+(\Sigma\,\kappa_i^2)^{-1} \tag{19}$$

times the precision of each stated value.

There is another simple way of looking at $\hat{\pi}$ that generalizes. In effect, S has provided two values for $\pi(A)$; q' directly and $\Sigma q_i\,\kappa_i$ indirectly. These have variances σ^2 and $\Sigma\kappa_i^2\sigma^2$, where σ^2 is the variance of each statement by S. Taking a weighted average of these two values with weights inversely proportional to the variances, we obtain $\hat{\pi}'$ as given by (18).

Returning to a single event X, the case $n=2$, the precision of $\hat{\pi}\{A\,|\,q(A), q(A\,|\,X), q(A\,|\,\bar{X})\}$ is $1+\{\pi(X)^2+\pi(\bar{X})^2\}^{-1}$ relative to that for $q(A)$. The improvement is remarkable. If $\pi(X)=\frac{1}{2}$, as in the example of a toss, the reconciled value has three times the precision of the original value. The extension of the conversation to include X has tripled the precision, and it would appear that the extension is important in assessing probabilities. As we shall see below, the increase in precision is considerably smaller when the assessments are positively correlated, as is likely to be the case in most applications.

It is easy to see in the general case that the precision (19) is maximized when all the κ's are equal, that is, $\kappa_i = n^{-1}$: it is best to choose a partition in which the constituent events are all equally likely. The precision for the reconciled value is then $(n+1)$ times the precision for the original value. (This was the case with the example of winning the toss.) Expressed differently, the increase in precision due to refining a partition from n to $(n+1)$ elements is equivalent to that which could be had from an independent assessment for the probability of the target event. Since such independence judgements are not available in practice, the principle of extension of the conversation can be used to play a similar role.

The result that among partitions of size n the precision is maximized when all probabilities in the partition are equal is a very special solution to what we call the *design problem*, that is, the problem of designing questions to be answered probabilistically by S in such a way that the expected precision of the reconciled value for the target event is maximized. It is analogous to the design problem in weighing objects, where the purpose is to choose the weighings in such a way as to maximize the precision of the final determination of weight. We do not discuss this problem in the present paper except, as here, when solutions arise as a by-product of our investigation of the principle of reconciliation. One cannot choose the best design until the reconciliation problem for that design has been solved.

We have also considered external conditioning with the log-odds metric and with correlation present between the various assessments. Space prevents details being given but a numerical example in the three cases may prove illuminating. Consider the assessment of the probability of an energy crisis in the United States during the next decade, denoted A. Let X denote the development of new, effective methods for the use of solar energy during the same period. Suppose $q' = q(A) = 0{\cdot}7$, $q_1 = q(A\,|\,X) = 0{\cdot}5$, $q_2 = q(A\,|\,\bar{X}) = 0{\cdot}8$ and $\kappa = q(X) = \pi(X) = 0{\cdot}5$. With the probability metric, the reconciled value is 0·67 with precision three times that of each of the q's, as we have just seen. With the log-odds metric, the reconciled value is still 0·67, and the precision is slightly more than three times that of the original value, 0·7. This suggests that the procedure is fairly robust to changes in the metric.

In the case where correlations are present, suppose they all have the value $\frac{1}{2}$ and all the precisions, in the probability metric, are equal. Then the reconciled value is 0·67 exactly as before, but the variance is decreased from 1·0, say, to 0·67; that is, the precision is only increased by 50 per cent rather than tripled as before. This result reflects a general tendency for correlations to have little effect on the probabilities but a substantial effect on the precisions.

4.3. *External Conditioning*: $\pi(X)$ *Assessed*

To help in our understanding of the situation, external conditioning has been considered only when the probability of the conditioning event is known, since this makes the analysis

linear. We now pass to the general case supposing that all four probabilities are assessed with the same errors and with no correlations. As before, we write $q(A) = q'$, $q(A \mid X) = q_1$, $q(A \mid \bar{X}) = q_2$ and introduce $q(X) = q_3$. The π's correspond and the coherence constraint is that $\pi' = \pi_1\pi_3 + \pi_2(1-\pi_3)$. The function to be minimized is

$$(q_1-\pi_1)^2+(q_2-\pi_2)^2+(q_3-\pi_3)^2+(q'-\pi')^2. \tag{20}$$

An approximation is available using the alternative way of interpreting $\hat{\pi}'$ in (18) as a weighted average. Here the two values for $\pi(A)$ cited by S are q' directly and $q_1q_3+q_2(1-q_3)$ indirectly. The variance of the former is σ^2, say. The latter is non-linear, but its differential is $q_3\,\delta q_1+(1-q_3)\,\delta q_2+(q_1-q_2)\,\delta q_3$ so that its variance is approximately $q_3^2+(1-q_3)^2+(q_1-q_2)^2$, times σ^2. The evaluations are independent. Taking the weighted average and rearranging to obtain the same form as before, we have

$$\hat{\pi}'-q' = \frac{q_1q_3+q_2(1-q_3)-q'}{1+q_3^2+(1-q_3)^2+(q_1-q_2)^2}. \tag{21}$$

The precision is

$$1+\{q_3^2+(1-q_3)^2+(q_1-q_2)^2\}^{-1}. \tag{22}$$

Both these results are approximate but probably adequate for most applications. The result for the precision is especially interesting since it shows, in comparison with the result for known $\pi(X)$, the reduction in precision due to uncertainty about $\pi(X)$. A straight use of the precision result for known $\pi(X)$, equation (19), would give, in the present notation $1+\{q_3^2+(1-q_3)^2\}^{-1}$, differing from (22) in not having the term $(q_1-q_2)^2$ in braces. It is this term that produces reduction in precision. Hence, lack of knowledge of $\pi(X)$ is most critical when $q_1 \neq q_2$; that is, $q(A \mid X) \neq q(A \mid \bar{X})$. If $q_1 = q_2$ then it does not matter. This is intuitively sensible.

In the energy problem, we have as before: $q' = 0{\cdot}7$, $q_1 = 0{\cdot}5$, $q_2 = 0{\cdot}8$ and $q_3 = 0{\cdot}5$, except that q_3 is now uncertain. The reconciled value of π is 0·67 as before, illustrating the robustness of the reconciliation procedure. The precision is increased by the external conditioning from 1 to 2·70, instead of 3·0 as before. By using the full equations (20), it is possible to calculate the other reconciled values, which are $\hat{\pi}_1 = 0{\cdot}52$, $\hat{\pi}_2 = 0{\cdot}82$ and $\hat{\pi}_3 = 0{\cdot}49$.

The extension to a general partition is straightforward.

4.4. *External Conditioning*: *Causation*

A special case of external conditioning arises when the conditioning event is a necessary antecedent to the target event. For example, let A be the event that S will die from cancer, and let X be the event that S will develop cancer at some time during his life. A formal way of expressing the special feature of this situation is to say $\pi(A \mid \bar{X}) = q(A \mid \bar{X}) = 0$. S will then provide $q' = q(A)$, $q_1 = q(A \mid X)$ and $q_3 = q(X)$, and coherence imposes the constraint $\pi(A) = \pi(A \mid X)\,\pi(X)$, or $\pi' = \pi_1\pi_3$ which is non-linear. As before, we can think of S as providing two estimates of π': q' directly, and q_1q_3 indirectly. If the former has variance σ^2, the latter, assuming all values stated by S are equally precise and uncorrelated, has approximate variance $(q_1^2+q_3^2)\,\sigma^2$. Hence, weighting inversely proportional to the variance, we have for a reconciled value, $\hat{\pi}'$, the result

$$\hat{\pi}'-q' = \frac{q_1q_3-q'}{1+q_1^2+q_3^2} \tag{23}$$

with precision $1+(q_1^2+q_3^2)^{-1}$ times the original precision.

Suppose S assesses the probability of his death from cancer as $q(A) = 0{\cdot}35$, the probability of his acquiring a cancer as $q(X) = 0{\cdot}40$, and finally the probability of dying from an acquired cancer as $q(A \mid X) = 0{\cdot}70$. With $q' = 0{\cdot}35$, $q_1 = 0{\cdot}70$ and $q_3 = 0{\cdot}40$, the reconciled value is 0·31; and the precision is increased to 2·54 times its original value.

Notice that this method is especially valuable when q_1 and q_3, and hence q', are small, since then the increase in precision will be greatest. However, the probability metric may not be too appropriate for small probabilities and a log-odds metric preferred.

5. DISCUSSION

In his treatise on the foundations of statistics, Savage writes:

> "According to the personalistic view, the role of the mathematical theory of probability is to enable the person using it to detect inconsistencies in his own real or envisaged behavior. It is also understood that, having detected an inconsistency, he will remove it. An inconsistency is typically removable in many different ways, among which the theory gives no guidance for choosing. Silence on this point does not seem altogether appropriate, so there may be room to improve the theory here . . .
>
> "To approach the matter in a somewhat different way, there seem to be some probability relations about which we feel relatively 'sure' as compared with others. When our opinions, as reflected in real or envisaged action, are inconsistent, we sacrifice the unsure opinion to the sure ones. The notion of 'sure' and 'unsure' introduced here is vague, and my complaint is precisely that neither the theory of personal probability, as it is developed in this book, nor any other device known to me renders the notion less vague." Savage (1954, pp. 57–58).

The present paper represents an attempt to deal with the issues raised by Savage, namely, the resolution of inconsistencies and the weighting of opinions. Our approach is based on a division of labour between a fallible subject S and a coherent investigator N who uses S's assessments to estimate S's "true" probabilities in the internal method, or to update his own beliefs about the world in the external method. Naturally, such an approach encounters both philosophical and practical difficulties.

Perhaps the most obvious philosophical objection pertains to the coherence of N. Why permit first-order incoherence of q but exclude second-order incoherence of p? If people are inevitably fallible, is it reasonable to postulate a coherent N? Alternatively, one could argue, if S has an access to a coherent N, why permit inconsistency in the first place?

The coherence of N is needed to ensure the coherence of the reconciled values. If the core distributions, for example, are also allowed to be incoherent, they must be reconciled before they can be used to reconcile the basic assessments. This leads to an infinite regress that can be avoided only by assuming coherence somewhere in the process. Indeed, it does not appear unreasonable to assume a fallible assessor who is capable—in a more reflective mood and perhaps with the help of paper and pencil—of detecting and reconciling his own inconsistencies.

Notice that once N's coherent assessments, the core distributions, are determined, they are available for the resolution of any problem posed by S. In other words, the role of N is to link together the possible situations that S might face. This is in the true spirit of coherence in which a problem is not considered in isolation but viewed in conjunction with other problems, both real and conjectural. If the reader considers a simple example, he can easily conclude that our methods are unnecessarily involved—easier, he might say to do a naive reconciliation rather than determine the core distributions and then use these to effect the reconciliation. It is only when a set of examples is considered that the power of the methods is revealed, for then, a single determination of core elements serves for the whole set.

The present approach can be viewed as a compromise between two extreme positions on the nature of probability assessments: the rationalistic position that assumes coherence, and the empiricistic position that denies it. Neither position deals with the reconciliation problem; the former effectively ignores the issue, while the latter cannot solve it. By modelling a person as composed of a fallible S and a coherent N, we attempt to provide a more realistic idealization which could, nevertheless, be used to achieve rational reconciliation.

A central feature of the approach developed in this paper is the reliance on the core distributions and Bayes' rule to reconcile incoherence. Alternatively, one could start by considering the set of all conceivable reconciled values and then introduce criteria or axioms that restrict the choice of an admissible reconciliation. For example, if $q(A) = 0{\cdot}62$ and $q(\bar{A}) = 0{\cdot}34$, one may wish to restrict π so that $0{\cdot}62 \leqslant \pi(A) \leqslant 1 - 0{\cdot}34$. Additional constraints could further restrict the set of admissible values. This approach represents a viable alternative to the presens procedure. It remains to be seen, however, whether one could develop a compelling set of criteria for reconciliation that would lead to a unique, or at least a highly constrained, solution.

From a practical standpoint, the major obstacle to the application of the proposed procedure is the difficulty in assessing the core distributions. All three distributions are readily interpretable: $p(A)$ is N's prior, $p(\pi|A)$ describes the relation between S's true beliefs and the external world and $p(q|\pi)$ describes the relation between S's assessments and his true beliefs. Although these expressions are psychologically meaningful, their assessment may prove very difficult in many cases. We are keenly aware of this problem, and much of the specific assumptions discussed in Sections 3 and 4 were introduced to simplify the assessment of the core distributions. It remains to be demonstrated that this information can be elicited from laymen and/or experts for non-trivial problems.

Reflection suggests to us that the introduction of these core distributions and, in particular, the occurrence of precisions and correlations, is a reasonable, and perhaps necessary, requirement in the real-world situation before reconciliation is possible.

In the light of these difficulties, it could be argued that instead of applying the formal procedures developed in this paper, we could simply instruct the subject to resolve his own inconsistencies in the way that he finds most appropriate. Although this approach could often be employed, we believe that an explicit model provides a useful tool for the analysis of coherence. It focuses attention on the data needed to resolve incoherence, and it provides a rational procedure for reconciliation.

Of the three elements in the core distributions, $p(A)$ and $p(q|\pi)$ are of familiar types, but the third, $p(\pi|A)$, is rather novel. It provides, together with $p(\pi|\bar{A})$, an expression of S's ability as a probability assessor; and, in particular, a statement of S's variability. In effect, it looks at S as a diagnostic instrument: in medical terms, if the patient has appendicitis, A, what probability is the doctor, S, going to assign to A; and similarly for a patient with abdominal pain not originating from a ruptured appendix. We have seen that it is related to the calibration concept, $f(A|\pi)$. It leads, for example in equation (9), to unexpected adaptations from π (or q) to revised probabilities for A, $p(A|\pi)$ or $p(A|q)$, which cannot always be π or q.

S. French, in a paper not yet published, has suggested to us that $p(\pi|A)$ might depend on $p(A)$. In words, N's assessment of S might depend on his own views about A, the event under discussion. He cites the example where past experience has shown that in similar situations, N and S have tended to agree: then $p(\pi|A)$ might peak around $\pi = p(A)$. If this is admitted, we are led to a curious probability, $p(\pi|A, p(A))$ and an unexpected form of Bayes' theorem that French has considered, namely

$$p(A|\pi, p) \propto p(\pi|A, p)\, p(A),$$

where $p = p(A)$. This is an interesting idea that introduces a reasonable correlation between N and S, but it should be noted that our model partly allows for such correlation. We suppose π is unaffected by q, given A and given $\bar{A}$, but this will not imply that π is unconditionally free of q; indeed, quite the contrary if S is a good appraiser. Since p and π are also connected, this could induce a connection between N and S's assessments.

There are at least two ways in which the model can be generalized. First, the world external to both N and S can contain uncertain quantities and not just uncertain events. This has been discussed by Morris (1974, 1977). An example might be in the case where the

meteorologist is forecasting the amount of rain tomorrow rather than whether or not it will rain.

A second generalization is to the case where the data-base for S changes. In our discussion it has been supposed fixed. An example arises when the meteorologist, forecasting the weather on day 2 learns about what the weather has been on day 1. The role of the data-base certainly needs more examination: it might resolve the dilemma proposed by French, in that the data common to N and S might explain the possible correlation suggested by him.

Acknowledgements

This work was supported in part by the Office of Naval Research, Engineering Psychology Programs Contract No. N00014–75–C–0426 through Decisions and Designs Inc. and by ONR Naval Analysis Programs through Decision Science Consortium Inc.

References

Brown, R. V. and Lindley, D. V. (1978). Reconciling incoherent judgments—toward principles of personal rationality. Technical Report No. 78, 8–72. McLean, VA: Decisions and Designs, Inc.

de Finetti, B. (1974). *Theory of Probability*. 2 volumes. New York: Wiley.

French, S. (1978). Updating of belief in the light of someone else's opinion. Submitted for publication.

Good, I. J. (1952). Rational decisions. *J. R. Statist. Soc.* B. **14**, 107–114.

—— (1971). Twenty seven principles of rationality. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds), pp. 124–127. New York: Holt.

Kahneman, D. and Tversky, A. (1978). Intuitive prediction: biases and corrective procedures. *Management Sci.*, in press.

Krantz, D. H., Luce, R. D., Suppes, P. and Tversky, A. (1971). *Foundations of Measurement*, vol. 1. New York: Academic Press.

Morris, P. A. (1974). Decision analysis expert use. *Management Sci.*, **20**, 1233–1241.

—— (1977). Combining expert judgments: a Bayesian approach. *Management Sci.*, **23**, 679–693.

Ramsey, F. P. (1931). Truth and probability. In F. P. Ramsey, *The Foundations of Mathematics and Other Logical Essays*, pp. 156–198. New York: Harcourt, Brace. Reprinted in H. E. Kyburg, Jr and H. E. Smokler (eds), *Studies in Subjective Probability*. New York: Wiley, 1954.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

Slovic, P., Foschhoff, B. and Lichtenstein, S. (1977). Behavioral decision theory. *Ann. Rev. Psychol.*, **28**, 1–39.

Spetzler, C. S. and Stael von Holstein, C. A. S. (1975). Probability encoding in decision analysis. *Management Sci.*, **22**, 340–358.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, **135**, 1124–1131.

Winkler, R. (1969). Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Ass.*, **64**, 1073–1078.

Discussion of the Paper by Professor Lindley, Professor Tversky and Dr Brown

Professor P. G. Moore (London Business School): This evening's paper is a rewarding one from a number of points of view. Primarily, its interest for a meeting such as this must be in the intrinsic merits of the work done that lies behind the paper. I believe that tonight's audience will be well satisfied on that count.

Secondly, the paper explores one facet of a topic that is of interest to many of us who are anxious to see probability notions applied to the widest possible range of practical problems.

Thirdly, the paper brings together in a fruitful manner the work of three men from two continents with very differing backgrounds and experience who have worked in this particular field. It is a pleasure to see the way in which so much diverse wisdom and experience has been distilled into the one paper.

It is perhaps customary in this Society at this point to pause—and then to say, *however*—followed by another pause, before proceeding to tear the edifice apart. I should like to take a rather different tack from that and reflect on the first three sentences of the paper. I believe that these sentences, innocent-sounding though they may be at a first reading—and I suspect that most of us did not read them thoroughly anyway—contain the nub of a substantial problem which leads to

a whole series of subsidiary problems, one of which is tackled with some energy in tonight's paper. Let me draw your attention in particular to the words that occur in the middle of those sentences:

> "Because such events are essentially unique, the assessment of their probability must be based on personal judgement."

The implication here seems to be that they must be based solely on personal judgement and I wonder whether things are quite so black and white in practice as this extract, perhaps unintentionally, seems to imply.

An insurance company quoting for a life assurance policy, or a house fire policy, may have a mass of statistical data available that enables it to make what appears to be an objective assessment of the risk involved. On the other hand, the branch bank manager, or some body such as FFI (Funds for Industry), faced with a proposition from an entrepreneur to set up a fresh organization to manufacture and market a new product, may seem to have little or no information and to have to rely on personal judgement alone—a judgement that may well be based on assessing the man or woman more than the project.

But this dichotomy is not as watertight as it may appear to be. For example, suppose that the occupation of the 25-year-old proposer for life assurance to that insurance company is that of a North Sea diver. Does his mortality follow that of the tables that have been so laboriously compiled by statisticians and actuaries over the years?

The costs of the insurance, moreover, depend also on the annual expenses of an office in maintaining a policy and on the interest it can earn on the money paid over as premiums. Will there be heavy inflation, or will current interest rates fall back?

Luckily, in practice some of these items tend to cancel each other out—for example, high inflation rates engender high interest rates, but, nevertheless, there remain some considerable areas where the statistics can be considered only as a guide, and where judgement is still needed in order to be able to quantify these various inputs.

Our entrepreneur likewise will be putting forward a proposition where some elements are capable of technical and financial evaluation of a direct and meaningful kind, but in addition there may well be substantial elements which are not so readily assessable, and hence the overall assessment is a mixture of hard fact and soft feel.

Again, we could consider a problem that was touched on in the paper—that of medical diagnosis. A doctor facing a patient can make a spot diagnosis from the look of the patient, a feeling of his pulse, some knowledge of his past medical history, relative prevalence of different types of illness in the locality and so on. Alternatively, the doctor could go a stage further and take the patient's temperature and blood pressure, examine his urine and so on, before making a diagnosis. Or he could have the patient admitted to hospital and a battery of tests carried out before a diagnosis is made by a consultant. At each possible diagnosis point there is really a set of possible complaints, and the doctor is putting probabilities on each, deciding what to treat for on the relative magnitude of those probabilities. His assessment, therefore, can be said to be a mixture of information of varying degrees of quality to which he applies himself as an information processor to reach his diagnosis. The point I am trying to make, however—an important one, which I find to be frequently misunderstood—is that no probability can be proved to be correct in advance, but the quality of the assessments can be enhanced by the proper use of information, and coherence is one part of that proper use. In other words, we do not jump from a provable objective assessment to a non-provable, and hence—it is sometimes assumed—useless subjective assessment.

In real life no two situations are ever completely the same. The notion of repetition is a relative one, not an absolute distinction. No two matches at tennis are exactly the same, but a study of past matches of the opponents concerned can lead one to make a better forecast of the result than could be obtained by ignoring the information available and merely spinning a coin.

This point is discussed in Section 2.3, and it needs reiteration. The quality of the expert assessor concerned rests not only in the hit frequency rates but also in the conditional probability distributions, in that the expert who forecasts a 50 per cent chance of getting a head on each toss of a coin may be correct on his hit rate, but will not be greatly helping the English cricket captain.

Let me make one final point, which is in the nature of a challenge to the authors and other people. Many of the difficulties that arise in practical applications of probability judgements in the worlds of business and commerce, and in which we do look for things like coherence, are concerned with very small probabilities. For example, in lending money to business, the average annual loss

rate by American banks is under 1 per cent per annum. The figures for British banks are not published in exactly the same form, but I believe them to be rather similar. Banks cannot realistically afford this rate of loss to increase without substantially affecting their whole operations. This inevitably means that the question being asked in many instances is whether the chance of failure of some venture is or is not greater or less than 0·01. Commonly, it is suggested that this barrier could be raised, and higher compensating rates of interest charged. Again, is it possible, either objectively or subjectively, to feel confident that we can be correct to that degree of accuracy? The actuarial argument is that the number of such individual risks being handled is of such a magnitude that the benefit of averaging safeguards the overall financial situation of the organization concerned. But this is not the case for the individual manager in a large organization who is being judged on one or two particular projects, nor indeed for the branch bank manager who may desperately want to keep his total lending up, but his loss ratio down.

Incidentally, although it is not a subject that I can usefully develop this evening, the system in this country whereby so much risk lending is covered by loans and not by equity seems to me to elevate the importance of avoidance of loss on each particular project financed, and to downgrade the maximization of overall expected profits in the long run. Insurance companies would have had a bleak history if they had been forced in the past to operate on this principle!

Like many others, I have greatly enjoyed reading this paper. It gives me great pleasure to propose a cordial vote of thanks to the three authors.

Professor J. M. DICKEY (University College of Wales, Aberystwyth): This paper is important for suggesting a formalism for the problem that a person's elicited quantities may fail to be his subjective probabilities. This could be viewed simply in the spirit that verbal reports tend to contain reporting errors. But the problem goes deeper than that. One's actual beliefs may not cohere into joint probabilities.

In the philosophy of mind (e.g. Davidson, 1976) beliefs, opinions, uncertainties are viewed as what (together with values, preferences) determine human actions. Perhaps for convenience, one may say that beliefs actually *are* potential actions. Ramsey's normative theory of self-consistent behaviour tells us that technically non-foolish beliefs cohere into joint probabilities. I think, in practice, we should help each other to get the muddle out of our opinions, and this requires feedback on the ways we are incoherent during the process of probability assessment. Interactive computer terminals are ideal for this. So ultimately, I think the assessment process itself should remove any need for formal "reconciliation". There would remain, however, the need to reconcile different experts.

Philosophers call verbal reports of beliefs second-order beliefs. They are one's beliefs (verbal actions) concerning one's own beliefs (actions relevant to the uncertain event A in question). It is second-order beliefs that I wish to discuss here.

First, however, let me complain a bit about the authors' terminology. Subjective probability is sometimes viewed as a kind of language. (For example, the use of language has an effect on the speaker's thought, and the assessment of probabilities removes the expert's muddle, much as any tool has an effect on its user.) Now, "syntax" and "semantics" are words borrowed from the study of language. They are analogous to Frege's "sense" and "reference" in philosophical logic. Formal structure on the one hand, actual meaning on the other hand. But what can we say regarding probability? "Coherence" and what? Well, it depends on whether we want reference to the event A in question or to the person's first-order belief about A. The authors have neither. Their word "pragmatic" involves reference to specific symmetric probabilities. Their word "semantic" involves a mere calibration measure, a sample average. In my discussion to Dawid *et al.* (1973) I reveal this measure as one component of the Brier P score, which in turn measures reference to the event A itself.

May I complain also about the use of S or "subject" for "person". I should not, of course, blame these authors for merely borrowing an habitual usage of behaviourist psychology, wherein "subject" means "object". But it does seem backwards for "investigator" to mean the computer terminal or consulting statistician, instead of referring to the subject-matter expert who is interviewed. The humanistic American psychologist George Kelley (1955) should be mentioned here for his concept of a person who sequentially constructs a view of the world, a concept having much in common with de Finetti's (1970) open ended process in which the probability space increases in dimensionality.

In Aberystwyth we have a research programme underway in subjective probability modelling. For structures that are both interesting and useful, we work with distributions, instead of just

probabilities of simple events. We also model the dependence of opinion on a range of further information by conditioning the distributions on concomitant variables. For example, consider the t distribution as a location-scale family with the centre parameter taken as a linear form in the concomitant vector $\mathbf{x}$,

$$y \mid \mathbf{x} \sim \text{Student}_{\delta}(\mathbf{x}'\mathbf{b}, u).$$

The standard distribution in the family, $z \sim \text{Student}_{\delta}(0, 1)$, is the usual Student t distribution on δ degrees of freedom. Then $y = \mathbf{x}'\mathbf{b} + u^{\frac{1}{2}} z$. This can be generalized to a vector of dependent variables, a future sample $\mathbf{y}$, conditional on a matrix of concomitant vector values.

$$\mathbf{y} \mid X \sim \text{Student}_{\delta}(X\mathbf{b}, U).$$

Here the distribution is multivariate t, in location-scale form, $\mathbf{y} = X\mathbf{b} + C\mathbf{z}$ where $CC' = U$. Note that this provides a powerful theory of subjective response surfaces when $\mathbf{x}$ is taken as an arbitrary function of primary variables.

This model is useable as stated. For a special case, however, consider the familiar normal multiple regression sampling model, $\mathbf{y} \mid \boldsymbol{\beta}, \sigma, X \sim \text{Normal}(X\boldsymbol{\beta}, \sigma^2 I)$. If the parameters have a natural conjugate prior distribution, $\boldsymbol{\beta} \mid \sigma^2 \sim \text{Normal}(\mathbf{b}, \sigma^2 J)$, $\sigma^2 \sim s^2(\delta/\chi^2_\delta)$, then the preceding model is obtained as a Bayesian predictive distribution and

$$U = s^2(XJX' + I).$$

Again, I stress that this is a special case and, in general, subjective probability models with imbedded sampling models are special. Subjective probability models are useful to clarify belief, compare and possibly pool experts, update opinion to data, and plan experiments and make other decisions. In the presence of an underlying sampling model the updating is by Bayes' formula, so there is a bound on the dimensionality of the assessments needed.

Kadane *et al.* (1978) give methods for assessing such multivariate t models, that is, obtaining by interactive interview the values of the parameters, δ, $\mathbf{b}$, U, or in the special case, δ, $\mathbf{b}$, s^2, J. Dawid *et al.* (1979) extend the model and the methods to the matrix t. Of course, these models and parameters have to be "fitted" somehow to elicitations during the interview process. We elicit the expert's quantiles for univariate marginal and conditional distributions of $\mathbf{y}$, which are again Student t.

In the event that precise measurement considerations hold, there is no necessity for the interviewer N to infer the parameters as a formal Bayesian. However, in the contexts of the paper here today, the least squares methods proposed may fruitfully be extended to conjugate prior methods involving prior pseudo-observations. Fractional prior sample sizes would be needed.

But how are elicited quantiles distributed given the expert's actual subjective quantiles? If that were known, one could form a second-order-belief likelihood function and maximize it in δ, $\mathbf{b}$, U. We take a step in this direction by fitting the subjective-response-surface $X\mathbf{b}$ to the elicited marginal medians, since the person's actual medians under the model satisfy $\text{Med}(\mathbf{y}) = X\mathbf{b}$. Knowing the first two moments of the second-order-belief distribution of the elicited medians will permit the use of generalized least squares to obtain the fit. Dickey (1969) provides theory to the effect that the elicited medians might usefully be considered as unbiased with a scale matrix essentially proportional to the matrix U.

Obviously, I consider the paper given here today to be an important pathbreaking contribution to a future in which probability will be used as the language of uncertainty. It gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

Mr J. R. Buxton (Loughborough University): I would like to question the meaning of the probabilities $p(\pi \mid A)$ and $p(q \mid A)$ that occur in the authors' *external* approach. The difficulty with these probabilities is that we are asked to assess the probability of a *present* event, the assessment π or q, conditional on a *future* event. For example, in the case of the authors' Example 2, we are asked to assess probabilities conditional on there being a major energy crisis in the United States in the next decade.

At first sight, the situation looks analogous to that in the Bayesian analysis of a medical diagnosis problem, where we have to assess the probability of obtaining a particular test result, given that our

patient will eventually turn out to be suffering from some suspected disease A. However, in the medical case, we assume that the events "patient will turn out to have disease A" and "patient will turn out not to have disease A" are uniquely linked to the events "patient has disease A now" and "patient does not have disease A now". In other words, the future event that we have to consider is uniquely linked to a set of conditions that are present now. So when we assess p(test result | disease A), we can replace the future event "patient will turn out to have disease A" by a set of conditions that are present in the patient at the time of the test.

Returning to the authors' Example 2, there does not seem to be any comparable way of reducing the future event to a present one, since the occurrence of an energy crisis in the U.S. in the next decade cannot be uniquely linked to any set of conditions in the present state of the world. For example, we can easily imagine chance events that would break any direct link between the present and the future—the leader of a campaign for the development of alternative energy sources may for instance be killed in a car crash, with the result that the campaign weakens, and an energy crisis becomes more probable.

One possible way of removing the necessity to condition probabilities on future events, would be to replace $p(q \mid A)$ by $p(q \mid P_D)$, where P_D is a measure of the disposition or propensity of the current world situation to give rise to an energy crisis in the U.S. in the next decade. If we can assess $p(q \mid P_D)$, we can use Bayes' theorem to update our prior assessment of the world being in a given state:

$$p(P_D \mid q) \propto p(P_D)\, p(q \mid P_D)$$

We can then obtain an assessment for A by using:

$$p(A \mid q) = \sum_{P_D} p(A \mid P_D)\, p(P_D \mid q).$$

I have here assumed that P_D contains everything that there is to be known about the tendency of the world to give rise to the event A, so that, given P_D, A is independent of q, or $p(A/P_D, q) = p(A/P_D)$. The only new probability here is $p(A \mid P_D)$, our assessment of the chance of the event A, given that the current dispositional state of the world is P_D. If we interpret P_D as an objective single case *probability*, we may well be prepared to use P_D itself for our assessment of $p(A \mid P_D)$.

Although the introduction of the quantity P_D complicates the structure of the model, I feel that the revised model is closer to reality. In particular, it seems to me to be more natural to assess the probability of q relative to current dispositions as in $p(q \mid P_D)$, rather than relative to a future event.

Dr T. LEONARD (University of Warwick): I would like to add my congratulations to the authors for a brilliantly imaginative paper. This will be viewed as an extremely important contribution to the formalistic side of Bayesian statistics since the authors have developed a wide-ranging theory for the sorts of inductive thought processes which are used to assess subjective probabilities. I think, however, that it would be interesting to compare this approach in practical terms with more obvious, though informal, methods. For example, instead of the internal approach, N and S could presumably just have a chat, and rather than needing to think up a complex distribution $p(q \mid \pi)$ and obtaining a coherent π for S very dependent upon his own coherent appreciation of the world $p(A)$, N could simply interact subjectively with S, advise him about the laws of probability, and help him in intuitive terms to assess a more reasonable distribution. Instead of the external approach, N could just subjectively update his distribution for A by intuitively judging the merit of the information provided by q. Whilst I think that the authors' methods provide a valuable conceptual background to which statisticians may refer when inductively assessing their prior distribution, it is not clear to me whether the formalisms described would lead to too many advantages if rigidly applied in practice. I am indeed a bit worried that their theory might tend to play a similar sort of role to that of significance tests in classical statistics and serve to constrict the inductive thought processes which are so necessary to practical statisticians. Perhaps I should quote the philosophy "Formalisms impede the natural man" which is in fact a bit of a misquote due to Dickey whilst we were trying to find a way through the snow in the Welsh hills this morning.

Probably the most important applications of Bayesian methods lie in the area of multi-parameter estimation. For example, the paper by Lindley and Smith (1972) probably yielded the strongest ever practical advantages for the Bayesian approach. I do not know whether the authors' methodology could be extended to multi-parameter problems since it is often necessary to model and assess a whole multivariate prior distribution; consider, for example, the mean vector and covariance

matrix of a multivariate normal distribution. The inductive modelling of the prior covariance structure is of critical importance, and in this case I do not think that the statistician should feel constrained by de Finetti-type "coherence". For example, he might wish to interact with the data and use them to stimulate creative ideas about the covariance structure, and once he has induced and modelled this structure he might wish to estimate any unknown prior parameters empirically from the data. My general view is that, in practical situations, the mathematical theory of Bayesianism should be tempered with a bit of pragmatism, and that it will then often possess substantial advantages which might not be obvious if too much formalism is employed. I am somewhat daunted by the prospect of subjective Bayesianism detaching itself from the data-analytic substance of the subject of statistics. I think however that tonight's paper is beautiful in theoretical terms and I particularly admired the analysis in Section 3 using log-odds transformations.

Professor D. R. Cox (Imperial College, London): The calibration of subjective probabilities, mentioned briefly in Section 2.3, can be investigated (Cox, 1958) via an assumed model in which for the ith event A_i, of subjective probability q_i, we have

$$\log\{\text{pr}(A_i)/\text{pr}(\bar{A}_i)\} = \alpha+\beta\log\{q_i/(1-q_i)\},$$

where α and β are to be estimated from relevant data. This allows for simple systematic distortion of the probability scale. My one practical encounter with this was connected with a maintenance problem in which q_i was the subjective probability that a piece of equipment would be operating on a particular day. Data suggested that the q_i's, while useful, were seriously optimistic and that one would need $\alpha<0$ and $\beta<1$. The authors' assumptions all seem to involve a kind of unbiasedness, i.e. that the q_i are subject to "random" rather than "systematic" errors. How crucial is this to their argument?

A second point concerns the circumstances under which discrepant information (here probability assessments) should be used as evidence of substantive disagreement rather than as a basis for pooling. Of course, in a broad sense, this is a widely occurring problem in statistics. The Bayesian formulation is effective for pooling information and some pooling is very desirable, but only provided the information pooled is reasonably consistent at least in some rough sense. If there is a major clash between different sources of information, surely attempts to resolve this by detailed discussion are called for. In the particular context here, if the subjective probabilities are too inconsistent, the reasons for this inconsistency should be isolated and the probabilities revised by rational discussion, rather than by arithmetical manipulation. This seems uncontentious: can it be incorporated into the formalism?

My final point is closely related to one of Professor Moore's comments. The authors follow the most recent work on subjective probability by concentrating on self-consistency, so called coherency. While this aspect is certainly interesting and important, concentration on it seems to me to divert attention from a more central issue, namely that of finding probabilities meaningfully related to the real world. It seems to me astonishing, even as an idealized example, to discuss probabilities of types of death without explicit regard to the extensive empirical data on the problem. I appreciate, of course, the non-trivial difficulties of deriving a probability for a specific individual. Nor do I for a moment suppose that the authors would themselves in practice ignore substantial empirical information. The point, though, is that the theory sanctifies as coherent certain sets of numbers unrelated to the external world, while condemning as incoherent other sets of numbers which, while quite possibly capable of improvement, may be closely related to the external world. The objection is not to coherency as such, but to a primary concentration on it. A central problem of subjective probability seems to me that of how to incorporate various kinds of experience of the real world into probability judgements and of discussing when it is useful to do this quantitatively. Coherency considerations no doubt are a useful tool in this, but no more.

Dr L. D. Phillips (Brunel University Decision Analysis Unit and Department of Psychology): As a psychologist, I am concerned about one assumption in this excellent and original paper. In the internal approach, the authors suppose that the subject has a set of coherent, "true" probabilities, and they justify this by analogy to the measurement of length and to psychological test theory. It seems to me that these analogies are inappropriate. The true distance between two points has an existence apart from the observer, whereas probabilities, as de Finetti insists, do not. Furthermore, the error that adds to a true score to produce an observed score is generally considered

to reflect distortions introduced only by the measurement process, but considerable research on subjective probabilities shows that they are influenced by other factors as well, including aspects of the task, the cognitive style of the assessor, and influences of the group to which the assessor belongs; even the assessor's culture has an impact! I would have thought that Tversky's own work on heuristics in assessing probabilities would be sufficient to discredit the notion of a "true" probability. Confronted with an uncertain event, we experience feelings of uncertainty; if we are required to give a probability assessment, these feelings of uncertainty are relevant, but so are many other factors, and "true" probabilities do not enter in at all. If that is so, then one of the core distributions, $p(q \mid \pi)$, is not psychologically meaningful, and so may be very difficult indeed to assess.

Another point. Although the authors associate the notion of calibration with $p(q \mid A)$, I believe that it is also relevant to $p(q \mid \pi)$, N's opinion of S as a probability assessor, for N would be influenced in his assessment of that distribution by knowledge of the subject's calibration curve. Then σ^2 must be quite large in most practical circumstances, for recent work by Sarah Lichtenstein and myself extending Professor Cox's log-odds model of calibration, has shown that 200–400 assessments for events that subsequently occurred or not are needed to provide reasonable certainty about the subject's calibration.

This point, and the questionable status of "true" probabilities, suggest that the internal approach may not be very useful as a practical method for reconciling probability judgements. As a practising decision analyst, I am more intrigued by the possible applications of the author's interesting finding that substantial improvements in precision can be obtained by extending the conversation to other events.

Dr G. Horsnell (Departments of the Environment and Transport): I welcome this paper as a real contribution to the solution of the practical problems of using subjective probability assessments in policy decisions. The problem of coherence of probability assessments is particularly acute when dealing with very low levels of probabilities—here I am talking about levels of several orders of magnitude smaller than those to which Professor Moore referred. These occur, for example, in assessing the relative magnitude of environmental risks.

The two coherence conditions of Example 1, namely, that the probabilities of all possible and mutually exclusive events should sum to unity, and that the ratios of conditional probabilities and unconditional probabilities in the same subset should be equal are, I suggest, unlikely to be very useful at these very low probability levels.

An alternative approach is to use the stated risks as if they were coherent probabilities, and to derive what might be non-coherent probabilities for more complex events. For example, the statement that an event could happen only once in 1000 years is taken to mean that the event will never happen. In fact, this statement, if it is taken to mean an average of once in 1000 years, yields a chance of 10 per cent that the event will happen in less than 105 years, a 5 per cent chance of it happening in less than 51 years and a 1 per cent chance in less than 10 years. These latter statements are derived by assuming the probability of the event happening in any one year to be constant, and multiple happenings in any year to be impossible.

Clearly, these latter statements give an entirely different appreciation of the degree of risk from that given by the statement "once in 1000 years". If these frequencies are not observed in practice, or not thought likely, clearly the assumption that the event may happen in any one year with equal probability is either untrue or the original statement is uncalibrated and these derived probabilities may assist in the calibration.

I suggest that it might be a more accurate procedure to calibrate derived probabilities and work back up the event chain to calibrate the original unconditional probabilities, rather than starting off the process by attempting to assess these very small probabilities in the first instance.

I would appreciate any suggestions that the authors might have on the development of this approach in rationalizing the present assessments of risks, bearing in mind the very small probabilities with which we are dealing.

Professor D. J. Bartholomew (London School of Economics): There is one feature of the internal approach in situations like that illustrated in Table 1 which seems unsatisfactory. My

difficulty can be illustrated by the following example. Suppose that I assess the probabilities of S mutually exclusive and exhaustive events to be

$$0{\cdot}001, \quad 0{\cdot}250, \quad 0{\cdot}200, \quad 0{\cdot}100, \quad 0{\cdot}279.$$

It is then pointed out to me that these probabilities sum to 0·830 and hence that the assessment is incoherent. If we use the method in Section 4.1, with the probability metric, we have to adjust the probabilities by adding 0·034 to each ($= (1/5)(1-0{\cdot}830)$) to give

$$0{\cdot}035, \quad 0{\cdot}284, \quad 0{\cdot}234, \quad 0{\cdot}134, \quad 0{\cdot}313.$$

The problem with this is that the first event, which I originally regarded as very unlikely, has had its probability increased by a factor of 35! Though still small it is no longer smaller than the others by two orders of magnitude.

Had I been presented with the problem without the benefit of the authors' paper my intuition would have been to scale the probabilities by a factor to make them add up to one. This would reflect the fact that I have more confidence in my assessment of relative probabilities than in their absolute values. I appreciate that the authors' calculation depends on the choice of the least squares criterion and the probability metric and that other choices might give a more reasonable adjustment for my example. However, I am not clear about whether the theory is, or can be made, superior to my intuitive judgement in this case. If a mathematical theory is to be useful in practice the truth of its axioms should be more self-evident than its theorems and I am not sure that this is the case here.

Mr R. C. Bromage: I have a problem that is concerned with subjective probability as defined by Savage and arises in this paper. It is this: a Bayesian believes that probabilities can be put on almost any uncertain event. I say "almost" because I believe that there are some events for which he cannot ascribe probabilities.

To illustrate, consider the following three examples:

(1) That Mrs Thatcher will be Prime Minister on June 21st, 1980.
(2) That the Financial Times Index (FTI) was greater than 200 on June 21st, 1970.
(3) That Shakespeare wrote all the works of literature ascribed to him.

For the first event I can work out my probabilities by comparing gambles involving that event with other gambles, whose probabilities I know. Thus if I felt I was indifferent between accepting a gamble of winning £10 if Mrs Thatcher were Prime Minister on June 21st, 1980, and £0 if not, and a gamble of winning £10 if in a lottery draw one of numbers 1 to 40 were drawn, and £0 if one of 41 to 100 were drawn, then my probability for the first event is the same as for the second, which I choose to be 0·4. Note that in order to avoid accounting for inflation and the like, the second gamble has to be played on June 21st, 1980 rather than now. However, both gambles can be played and paid off.

For the second event I can perform similar tests to obtain my probability that the FTI was greater than 200 on June 21st, 1970. In this case, the gambles can be played and paid off when we find the information on the FTI that we need. Moreover, I assume that this information is obtainable through suitable records.

The third event poses a more difficult problem, because here I believe that the true outcome will never be determined. If this is my belief, it would be a futile exercise for me to try to compare gambles which can never be paid off. How then do I find my probability that Shakespeare wrote "his" own works?

How do these comments relate to the paper? Consider the probability $p(\pi/A)$. In order to work this out, we can try comparing gambles whose prizes depend on π. However, as in the third example above, there is nothing in the theory to say that we will ever know π, the subject's "true" coherent subjective probability.

Of course, a practical answer is readily available. Instead of comparing gambles we can compare probabilities directly. However, Savage's definition requires the ability to compare gambles, and indeed he argues against defining probability by direct comparison. I believe, therefore, that there is no place in Savage's theory for putting probabilities on some events, such as $p(\pi/A)$, and in general, events whose outcomes will never be known. If we wish to ascribe probabilities to such events theoretical justification must be found outside that theory.

Dr K. McConway (Research and Intelligence Unit, Cleveland County Council): I wish to describe an "adhockery" for reconciling probability assessments for a partition which do not sum to one. Suppose S states probabilities $\frac{1}{4}$, $\frac{1}{4}$ and $\frac{1}{4}$ for the three events A_1, A_2 and A_3 of a partition. He notices that these probabilities do not sum to one, so he uses equation (12) of tonight's paper to revise his probabilities to $\frac{1}{3}$, $\frac{1}{3}$ and $\frac{1}{3}$. At this point he decides he is not interested in A_1 and A_2 separately, but only in $C = (A_1 \cup A_2)$ and A_3, so he gives their probabilities as $\frac{2}{3}$ and $\frac{1}{3}$.

However, suppose he decides to perform this marginalization to a two-fold partition *before* he performs the reconciliation. Now, given that his original stated probabilities do not sum to one, it is not certain that he will state a probability of $\frac{1}{4}+\frac{1}{4} = \frac{1}{2}$ for C, but suppose he does, and lets his probability for A_3 remain $\frac{1}{4}$. Now, if he uses (12) to reconcile these probabilities, he will arrive at $\pi(C) = \frac{5}{8}$ and $\pi(A_3) = \frac{3}{8}$.

The point is that, if an assessor always uses the addition rule for probabilities when coarsening a partition in this way, the coarsening does not commute with the reconciliation given by (12), and it is easy to see that the same holds for other reconciliation methods given in the paper. Consider a partition $A_1, \ldots, A_n$ for which S states probabilities $q_1, \ldots, q_n$. If reconciliation is performed first, the reconciled probability of A_i is some function $\tilde{\pi}_i(q_1, \ldots, q_n)$. On coarsening to the two-fold partition A_i, $\bigcup_{j \neq i} A_j$, one arrives at a distribution defined by $\tilde{\pi}_i(q_1, \ldots, q_n)$. If, instead, the coarsening is performed before reconciliation, the stated distribution on the two-fold partition is defined by q_i and Σq_j $(j = 1, \ldots, n)$, so that the required commutativity implies that

$$\tilde{\pi}_i(q_1, \ldots, q_n) = f_i(q_i, \Sigma q_j),$$

where f_i is some function possibly depending on i. Suppose, however, that $f_i = f$ say, independent of i; then again using the commutativity we have that

$$f(q_i, \Sigma q_j) + f(q_k, \Sigma q_j) = f(q_i + q_k, \Sigma q_j),$$

so that $f(\cdot, \Sigma q_j)$ satisfies the well-known functional equation $g(a)+g(b) = g(a+b)$, and hence

$$f(q_i, \Sigma q_j) = \alpha(\Sigma q_j) \cdot q_i,$$

where $\alpha(\cdot)$ is some function, and the fact that $\Sigma \tilde{\pi}_i = 1$ $(i = 1, \ldots, n)$ shows that

$$\tilde{\pi}_i(q_1, \ldots, q_n) = \left(1 \Big/ \sum_1^n q_j\right) \cdot q_i,$$

as advocated by Professor Bartholomew.

I must emphasize that I am not putting this argument forward as a criticism of the excellent and stimulating work in tonight's paper, but merely to point out an alternative that should perhaps be considered.

Professor Cedric A. B. Smith (University College London): Lindley, Tversky and Brown are to be congratulated on boldly taking the bull by the horns and breaking the ice of a thorny issue, and we must be grateful for what light they can shed on the matter. But have they correctly crystallized our thoughts?

What motive can one have for making one's probability assessments "coherent" in their sense of the word? If by "subjective probability" we mean psychological feelings about degrees of belief, "incoherence" means in effect untidyness, which is uncomfortable, but not a complete disaster. On the other hand, if probability refers to the betting odds one is willing to contemplate, one can be "incoherent" in two different ways. If you are willing to offer odds up to 2 : 1 that tomorrow it will be sunny (rather than cloudy), and at the same time odds up to 2 : 1 that it will be cloudy (rather than sunny), I can certainly extract cash from you by taking up both bets with equal stakes. But if you are cautious, offering only odds of 1 : 2 of sun against cloud, and 1 : 2 of cloud against sun, which can be interpreted as meaning that your assessments of the probabilities of these events add to less than 1, you cannot have a book made against you in this way, and therefore have less motive for changing. You might lose because the low odds offered would not attract takers, and so would deprive you of the opportunity to make money. But raising the odds could be unwise.

In other words, although the idea of making probability assessments cohere sounds attractive, it is less clear that it directly confers any tangible benefit. Conceivably the benefit may be indirect, in that someone giving incoherent probability assessments may lack credibility. The problem would then be partly of maximizing other people's confidence in oneself, an interesting but complicated exercise.

Professor J. F. C. KINGMAN (Oxford University): I would like to draw attention to a recent mathematical result, which seems to me of considerable importance for subjectivist statisticians. Such a statistician, faced with observations $X_1, X_2, \ldots, X_n$, must arrive at a joint distribution to represent his prior beliefs. If he supposes that the order of the observations is irrelevant, and if moreover he can contemplate arbitrarily large values of n, then the celebrated theorem of de Finetti tells him that he must use a parametric model of the form $X_i = h(\theta, \varepsilon_i)$ for some function h, where $\theta, \varepsilon_1, \varepsilon_2, \ldots$ are independent and identically distributed. Moreover, θ can be expressed as a function of the future observations $X_{n+1}, X_{n+2}, \ldots$; in this sense it is a "real thing" (Smith, 1978).

The assumption of exchangeability is inappropriate, however, when more complicated experimental designs are involved. For example, in a two-way layout we would have observations X_{ij} $(i, j = 1, 2, \ldots)$, and it might be natural to make the weaker assumption that the joint distribution of the X_{ij} is unchanged under any permutation of the i and any permutation of the j. Is there an analogue of de Finetti's theorem under this condition of "marginal exchangeability"? The relatively simple case when the joint distribution is normal has been solved by Dawid (1977), but a general solution has now been supplied by Dr David Aldous of Cambridge. He shows that any marginally exchangeable distribution arises from a parametric model of the form

$$X_{ij} = h(\alpha, \beta_i, \gamma_j, \varepsilon_{ij})$$

for some function h, where the random variables $\alpha, \beta_i, \gamma_j, \varepsilon_{ij}$ are all independent and identically distributed. This is, of course, a non-linear generalization of the Bayesian version of the usual linear model.

This result of Aldous is much deeper than de Finetti's, as may be seen from the fact that, in general, the parameters cannot be expressed as functions of the X_{ij}. The proof is at present very complicated, but there is reason to hope that the techniques developed can be applied to more general experimental designs.

The following contributions were received in writing, after the meeting.

Professor G. A. BARNARD (University of Waterloo): After reading the galleys of the Lindley, Tversky and Brown paper, I am moved to comment "c'est magnifique, mais ce n'est pas la statistique". A statistician's training and creativity are focused on developing skill in taking complex sets of empirical data and simplifying them in such a way as to make clear the direction in which the data should lead us to modify our probability assessments and, to some extent, by how much these should be modified. In developing these techniques, the statistician makes use of empirically verified generalizations concerning the typical shapes of distributions of repeated measurements, the conditions under which such repeated measurements can be treated as relatively independent of each other, and so on. He does *not* study the factors, such as the fear of death, which may tend to bias an individual's subjective assessment of the probabilities of deaths from various causes. Such studies are entirely legitimate, and indeed important, but the skills called for are much closer to those of the psychoanalyst than of the statistician. It seems to me obvious that if we accept it as important that a person's subjective probability assessments should be made coherent, our reading should concentrate on the works of Freud and perhaps Jung rather than Fisher and Neyman.

It is far from obvious that in rendering a set of probability assessments coherent we improve them. In assessing the probability of an event such as the occurrence of an energy crisis in the United States, we need to take into account the willingness of people to change habits and to make other adjustments, involving matters on which our feelings may run very deep. On the other hand, there will also be involved assessments of the likely yield of wells in various parts of the world and so on, concerning which the emotional aspects have a much smaller role. The procedures which are suggested by the present authors for rendering our assessments coherent could well have the effect of spreading the errors in probability assessments arising from emotional involvement in one section of our thinking to the rest of our thought, and to this extent "infecting", as it were, those parts of our attitudes which otherwise would have been more nearly correct.

The paper contains a number of extraordinary statements from which I select just one: in the sixth paragraph we are told that "subjective judgements constitute the major data base for the measurement of uncertainty". The authors would presumably justify this extraordinary remark by appeal to their very special interpretation of the meaning of "measurement" and their very special

interpretation of "uncertainty". But this should lead no one to suppose that in the ordinary accepted meanings of the term, subjective judgements constitute a "data base" for anything at all.

Dr Derek W. Bunn (Oxford University): I should like to add my appreciation to the authors of this paper for producing a most original and thorough analysis of the Coherence Ideal in practical decision-making. The formulation in terms of an Incoherent S and a coherent N is evidently a powerful construct in the pursuit of this basic principle of rational decision analysis. At least one of the authors (R. V. B.) has, on several occasions, reflected the unattainable target of total coherence in referring to "the infinite and impeccable pains" required in the analysis of practical decision problems.

The internal approach suggested in the paper has practical relevance at two levels. For an individual working on his own, knowledge of how the ideal thinker (N) would go about modifying his assessments may well improve his thinking process in terms of coherence. The assessments involved would, in this case, require considerable introspection and further research is clearly required to demonstrate the feasibility and usefulness of this mode of thinking. At the other level, it is common practice for an analyst (N) to elicit probabilities from the decision-maker (S) and help to reconcile his assessments for him. In this case, however, the analyst should not introduce any of his own opinions, $p(A)$, $p(\pi)$, etc., since his function is explicitly not one of a consulting expert (if he were being consulted as an expert, then the roles of N and S would be reversed and the formulation would be the external approach). The relevance of the model would, in this case, have to relate to the "flat" $p(\pi)$ of Section 4.

The least squares procedures for reconciliation have evident usefulness, although it is worth noting that the type of incoherence shown in the medical example is really not so common in practice. Research on probability assessment procedures has now persuaded most analysts to assess in terms of relative likelihoods (e.g. cancer twice as likely as a stroke, etc.) rather than in terms of direct probability metric. In this case, the assessments are normalized as part of the procedure. Whether it is in fact better not to use relative likelihoods (because of its susceptibility to anchoring bias, perhaps) and use a direct procedure followed by reconciliation in the style presented is another issue for further investigation.

With regard to the external approach, this is a useful contribution to the expert use problem where the decision-maker (N) uses the advice of an expert (S). This is distinct from the consensus problem, since S does not have to agree with the final $P(A \mid q)$ of the decision-maker. It would be interesting to see some analysis of the authors' model addressed to the reconciliation of the consensus problem. Clearly, in the consensus case, the first individual would derive a $P_1(A \mid q_2)$ on the basis of the second individual's stated q_2. Likewise there would be a $P_2(A \mid q_1)$ from the second individual's perspective. For concensus, all $q_1, q_2, P_1(A \mid q_2), P_2(A \mid q_1)$ would be equal.

Related to this work, the authors may find deGroot (1974), Bacharach (1975), Seaver (1976), Bunn (1978) and Harman and Press (1978) useful references.

Professor M. H. deGroot (Carnegie–Mellon University, Pittsburgh, Pa 15213 U.S.A.): My congratulations to the authors on a stimulating paper. They are working in a new area and the ideas they present give us much to think about. Even the analogy in the first paragraph is stimulating and controversial. The human mind can be used not only to assess uncertainty, but also in a similar way to assess length, weight and time. A person's judgement as to which of two events is more likely to occur is similar to his judgement as to which of two objects is longer or heavier, or which of two actions takes more time. Using a ruler, scale or clock to help in making these judgements corresponds to performing an experiment and collecting relevant data, just as experimental data are used to help in the assessment of uncertainty.

Is it reasonable to assume that a person who declares an incoherent set of probabilities q actually has in his heart some true coherent set π? In order for a person to actually have subjective probabilities, he must have a consistent ordering of the relative likelihoods of an uncountably infinite number of events. A critical question is how well we can train people to be consistent in their judgements and coherent in their stated probabilities. Can the process of reconciliation be useful in this training?

It has often been suggested that a good weather forecaster will be calibrated, but anyone can get himself calibrated by making forecasts q that intentionally misrepresent his true probabilities π.

Suppose, for example, that a forecaster finds that there have not been enough rainy days among those on which he has stated the probability of rain to be 0·6. To get calibrated on these days, he simply waits for some days on which he is virtually certain of rain and announces his forecasts to be $q = 0{\cdot}6$. A forecaster with just a minimum knowledge of meteorology can stay calibrated over all value of q by using this technique and making few forecasts of extreme values of q.

It is indicated in the paper that a subject may be more certain about some of his probabilities than about others. I agree with this statement, but I am not sure that I agree with the authors' interpretation. The ranges of q_i in Table 1 are said to represent the stability of an assessment and "the degree to which it varies upon further reflection". If that is all these ranges represent, then why not have the subject be a little more careful and reflect a little further in the first place in order to reduce or eliminate these ranges? I believe these ranges represent the subject's view of how his assessments might vary in the light of further *data*, not further reflection. Suppose that the proportion of red balls in a certain box is unknown to me and that tomorrow one ball is to be drawn from the box at random. I would typically assign probability 1/2 to the event that it will be red. I know, however, that if I could observe the colour of a few balls or even one ball selected at random from the box tonight, I might very well change my probability. On the other hand, without this or other relevant information, no amount of reflection during the night is likely to make me change from 1/2.

Dr S. FRENCH (University of Manchester): The authors are indeed to be congratulated on a most interesting paper. It points forward to many areas of research and application. However, before it is applied I should like to offer a word of warning.

Ideally a decision analysis does not primarily prescribe the optimal decision, although this is, of course, one of its objectives. Surely its most important objective is to bring the decision-maker (d-m) an understanding of how the balance of his beliefs and preferences is made up. The aid of decision analysis is sought mainly in cases where the d-m cannot initially see where this balance lies.

As a simplification we may divide an analysis into six stages.

(i) Explain the principles of rationality (coherence) upon which the theory is founded and only proceed if the d-m accepts them.
(ii) Elicit a representative subset of d-m's beliefs and preferences and check for inconsistency.
(iii) Explain any inconsistencies to d-m and let *him* resolve them.
(iv) Fit coherent probabilities and utilities to d-m's beliefs and preferences.
(v) Find option of greatest expected utility.
(vi) Perform sensitivity analysis.

It is stages (i), (iii) and (vi) that bring the d-m understanding.

Speaking very simply, if the methods of this paper are applied within stage (iv) alone, then I applaud their use. Of course, the distinction between stages (iii) and (iv) is by no means as clear as I have made it here, and in reality there may still be some inconsistencies to resolve in stage (iv). This this method will be most useful here. However, if the method is being proposed as a complete replacement for stages (iii) and (iv), then I caution strongly against its use. For by removing stage (iii), you remove an operation in which the d-m can gain understanding and revise his feelings in the light of this understanding. To remove this stage surely devalues the analysis.

Put somewhat differently, the paper postulates a model in which, whatever he *says*, underneath it all the d-m is coherent. I do not believe this assumption holds even approximately until stage (iii) has been completed.

Professor I. J. GOOD (Virginia Polytechnic Institute and State University): The problem of reconciling incoherent probability judgements has presumably never been overlooked by subjectivists, but it has not been given extensive formal treatments as in today's scholarly paper. When the judgements are your own, the problem is essentially the same as that of arriving (approximately) at your own *mature* probability judgements. This problem is the main one in the Bayesian philosophy and one I have thought about for many years. For example, Good (1952) discusses hierarchies of probabilities and it includes the claim that "the higher the type the less the wooliness matters"; also that "It would probably not be necessary to introduce Bayes solutions of type 4". The paper also discusses how to pay people for probability judgements, using a logarithmic payoff so as to encourage good judgements. The idea can of course be applied in multiple-choice examinations. When the logarithmic payoff is used, the subject should be warned that nothing is impossible.

(Alternatively, asserted zero probabilities can be replaced by some $\varepsilon > 0$.) There was one earlier paper on this topic by Brier (1950), in a meteorological journal. He used a quadratic payoff function, but I did not know of it until much later, and I argued in favour of a logarithmic payoff. Jacob Marshak considered that the topic opened a new field of economic theory, and the topic now has a "literature" of a hundred papers or so.

Allowing a hierarchy of probabilities has the advantage that it makes it possible to formalize the process of saying that one of your judgements is more reliable than another one, and thus gives a clue on how to resolve incoherences. It can also be usefully applied to multinomials and contingency tables (for example, Good, 1967, 1976) and leads to compromises in which hyperparameters are estimated instead of being assigned hyperpriors (for example, Good, 1967; Good and Gaskins, 1979).

An attempt to codify the principles for arriving at coherent judgements, in a manner analogous to the laying down of axioms, is made in Good (1971). When A. M. Turing read the first draft of Good (1950), he especially liked the Device of Imaginary Results (pp. 35, 70 and 81).

Recently I proposed a method for reconciling the judgements of several experts concerning the probability distribution of mineral deposits (Good, 1979). There is here space only to give the citation.

Professor J. B. KADANE (Carnegie–Mellon University): The authors should be thanked and congratulated for an apparently novel application of Bayes' theorem to the problem of elicitation of opinion. They distinguish between two cases, which they call the internal and external approaches, depending on whether emphasis is placed on the investigator N's marginal opinion of S's expertise $p(\pi(\cdot) \mid q)$ or on N's opinion of the world $p(\cdot \mid q)$, respectively. Of course, in general, the joint distribution $p(\cdot, \pi(\cdot) \mid q)$ might be of interest.

This paper makes an important contribution to the philosophy of Bayesian statistics, as it clarifies the question raised by Savage (1954, pp. 56–60) about the meaning of uncertain probabilities, which he regarded as one of the two most vexing unsolved problems for Bayesians.

I wonder, however, how practical the methods of this paper may prove to be for elicitation. The authors propose to replace the already difficult task of specifying values for the hyperparameters of a family of distributions representing S's opinion, with several substantially more difficult tasks. After specifying N's joint distribution on all the quantities, the authors place particular emphasis on estimates of the π's rather than on the full posterior distribution. I found this suggestion surprising in view of the apparent bimodality of the posterior distribution reported in Section 3.

One major reason for a careful study of the elicitation problem would be to learn how to design better elicitation experiments. In their treatment of a design question in Section 4.2, the key assumption is on the correlations among $q(A)$, $q(A \mid X)$ and $q(A \mid \bar{X})$. To continue the example of the paper, let X be the event that a team wins the toss at the beginning of the contest, and A be the event that the team wins the contest. Suppose (and nothing in the paper contradicts this assumption) that in my opinion the toss at the beginning is irrelevant to the winning of the game. Then I think $q(A) = q(A \mid X) = q(A \mid \bar{X})$, but the precision of $\pi(A) = q(A)$ should not be increased by knowledge of $q(A \mid X)$ and $q(A \mid \bar{X})$, because there is no "new" information in them. In other words, the correlation among $q(A)$, $q(A \mid X)$ and $q(A \mid \bar{X})$ is then one. Thus you need to know how I am thinking about $q(A)$, $q(A \mid X)$ and $q(A \mid \bar{X})$ to know what correlation to assign. This suggests a difficulty in relying on the analysis of Section 4.2 for advice in designing elicitation experiments.

I hope that the authors plan to extend their fascinating beginning by exploring the tractability of their approach in examples more realistically complex, bringing to bear whatever insights modern psychology may have to offer in both the design and analysis of elicitation.

Dr PETER A. MORRIS (Applied Decision Analysis, Inc.): The authors have undertaken the ambitious task of creating a general framework for resolving incoherent probability assessments. This is an interesting paper that raises a number of challenging issues. Unfortunately, time and space allow only a limited discussion, so I will restrict my comments to a subset of the issues raised by the paper.

First, it seems to me that the efficacy of the methodology depends on the underlying model of an assessor's incoherence. The approach in this paper is consistent with the "noisy measurement" model that holds that probability assessments are simply inaccurate measurements of an assessor's true state of information. Based on this model (which is similar to Tani's "authentic probability"

model presented in *Management Science*, October 1978), the framework proposed, based on the physical measurement analogy seems appropriate. However, another possibility is that the inconsistency is the result of a purely logical error that the assessor would recognize if confronted with it. In this case, any sort of processing rule that weights the different assessments would result in an answer containing a systematic error that might have been avoided by more careful thinking and modelling. For example, consider the case of resolving inconsistencies in several assessments of the length of the hypotenuse of a right triangle. One type of error results from measuring the hypotenuse directly using an inaccurate measuring tool; another type of error results from measuring the sides and misapplying the Pythagorean theorem. In the first case, the more measurements, the better the estimate of the hypotenuse, which corresponds to the analogous probability result in this paper. However, in the case of a logical error, no amount of measuring will result in a reasonable estimate if the measurements are processed using erroneous logic.

Another observation is that the results seem to be more general than implied: I can see no reason why the same logic cannot be applied in situations in which there is no incoherence, since, even if a set of probabilities obey the axioms, they may still be noisy measures of the assessor's true state of information. In this case, the investigator is essentially treating the assessor as an expert. In fact, the example given in the paper of resolving a set of conflicting expert judgments is a case in point, since expert disagreement does not imply incoherence or inconsistency. This raises some interesting issues concerning the legitimacy of a decision-maker acting as his own investigator, and changing his initial assessment according to his opinion of himself as a probability assessor.

One last observation is that I believe Savage had at least one other issue in mind in his quoted work that is not addressed by the methodology of this paper. The authors addressed the issue of incoherence by constructing a procedure for guaranteeing adherence to the axioms of probability and decision theory; however, another deeper issue is the problem of finding a framework for understanding why many persons choose *knowingly* to violate the axioms by acting more confidently on the basis of one over another numerically equivalent coherent assessment. In this case, even if a set of assessments are coherent, or modified to be coherent, a basic problem remains. I believe Savage was suggesting that, in some cases, he found himself wanting to violate the axioms for reasons he admitted he could not articulate. This is a more basic problem than the one the authors addressed of figuring out how to modify assessments to meet the axioms.

Professor R. L. WINKLER (Indiana University): I would like to congratulate the authors on an excellent paper which attacks an important problem that has not received the attention it deserves. The paper contains useful insights on the reconciliation of incoherent probabilities, and I expect it to stimulate further work, both with respect to incoherent assessments and with respect to more general questions concerning the modelling of experts.

Rather than comment on details of the paper, I wish to look at the model from a broader perspective. The paper focuses upon the reconciliation of incoherent probability assessments, but models of this sort could also be very useful in situations where all probabilities are in fact coherent. The approach can then be thought of in terms of modelling experts, who may be coherent or not. Moreover, the experts need not all be people; some could be models used to generate probabilities, although a subjective element would still be present (e.g. in the development of such models).

In terms of what the authors call the internal approach, we may wish to make inferences about certain probabilities on the basis of certain other probabilities which are themselves internally consistent. In the notation of the paper, we may be interested in $\pi(A)$ when the available probabilities q pertain to events other than A. It may be difficult for an expert to think about the probability of A but much easier to think about probabilities for some related events. For example, in a Bayesian analysis of a normal linear regression model with several explanatory variables, a prior distribution for the regression coefficients is needed. Even an expert with some experience in probability and statistics may find it difficult to assess probabilities directly for these coefficients. A much easier task is to assess probabilities for observable quantities, such as probabilities for the dependent variable in the regression model conditional upon certain values for the explanatory variables. The problem is to use these assessed probabilities to make inferences about the expert's π for the regression coefficients. For discussions of this problem, see Kadane *et al.* (1978) and Winkler *et al.* (1978).

In terms of the authors' external approach, even if the assessments represented by q are coherent (and even if they accurately reflect π), we must consider whether they can be taken at face value or

whether they need to be adjusted, or calibrated, in some fashion. Furthermore, the possible dependence between S's assessments and N's assessments must be considered, and this issue becomes more critical if assessments are obtained from more than one S. The impact on probability revision of such dependence is studied in the context of a multinormal model in a forthcoming paper of mine, and some preliminary empirical results involving probabilities for point spreads in U.S. football games and weather forecasters' precipitation probabilities indicate that such probabilities exhibit a considerable amount of dependence among assessors, as might be expected because of similarities in training, experience and information. Along these lines, since we see q rather than π, should not q replace π in the form of Bayes' theorem given at the end of the paper and attributed to French? N could assess $p(\pi \mid A, p)$ but then find $p(q \mid A, p)$ from (4) with $p(\pi \mid A, p)$ used in place of $p(\pi \mid A)$. This sort of modelling is similar in spirit to some of the work of Morris (1977), and another useful reference not likely to be seen by many statisticians is Schum (1977).

J. M. B. Moss: One of the authors of the paper has emphasized: "I know of no field where the foundations are of such practical importance as in statistics" (Lindley, 1971, p. 3). It is not enough, however, for the foundations of a discipline to be *roughly* right; rather the attempt should be made to get them *exactly* right, and I thus raise two philosophical queries about the paper.

The first, though verbal, suggests a possibly significant conceptual connection between coherence, in the sense of the paper, and idealist philosophy. The received distinctions between the syntactic, semantic and pragmatic dimensions of the meaning of the language of science were drawn over 40 years ago by the Chicago philosopher C. W. Morris (Morris, 1938a, pp. 84 *et seq.*; Morris, 1938b, p. 70). Applied to the language of inference, these distinctions fail to accord with those in Section 1 of the paper between the "pragmatic, semantic and syntactic" criteria for evaluating subjective probability assessments. Moreover, the authors' elucidations fit no better with other natural language uses of "syntactic", "semantic" or "pragmatic", except in so far as all conceptual problems in the theory of meaning are to be classified as semantic. What the authors refer to as the pragmatic criteria apply when the statements *correspond* to the way things are, whereas their semantic criteria employ only the *coherence*, in the sense of the idealist theory of truth, of the set of statements. Finally, their syntactic criteria are apparently realized in the laws of probability, which would, however, often be said to be true in virtue of their *semantic* content—though no such account of these laws can be complete, since it fails to explain their applicability. Further, the use of the word "coherent" in Bayesian statistics (Lindley, 1971, p. 6) appears to agree well with its use in idealist philosophy; some may thus be tempted to conclude, by realist arguments, that the coherence of a collection of statements may indeed be a *necessary* condition for their truth or acceptability, but that it will not in general be also a *sufficient* condition.

The second query concerns the nature and status of the coherent subjective probability set π introduced in the paper, which it is "assumed that the subject has, in some sense" (Section 1). Moreover, a task for the coherent observer N is to provide, from what *can* be observed, an assessment of the "unobservable 'true' probability π" (Section 2.1). What, however, is the content of the claim that π, in some sense, really exists? At first sight, the claim appears to belong to experimental psychology, but if so, the questions arise: is the claim (established to be) true, and if so does it—or can it—play a role in the foundations of rational inference? Certainly, it is often supposed that foundational results cannot in principle depend upon any contingent truths. Perhaps therefore the assumption of the existence of a probability set π is *not* empirical, in which case the question arises as to what it is. Does it in fact make any difference whether or not in a particular case π exists? To use the language of psychology, π is at best an "hypothetical construct", but would there result any difference to the work of the present paper if π is no more than an "intervening variable" (MacCorquodale and Meehl, 1948)? If not, this query would in no way affect the practical utility of the authors' work, but solely its theoretical foundations, hence the foundations of personalist Bayesianism. One of the authors appears to hold that personalist Bayesianism is an important part of the truth about the foundations of statistics. ("What shall we do with sampling theory statistics ... with all those methods that violate the likelihood principle? The answer is, let them die". Lindley, 1975, p. 112.) Consequently the status of the crucial assumption that man is coherent, i.e. is *really* a rational animal, needs to be clarified. For example, what consequences for the foundations of statistics ensue if, for many subjects, the hypothesis that π exists should be established to be a false, though empirical, assumption?

The AUTHORS replied later, in writing, as follows.

D. V. LINDLEY: Some contributors to the discussion take a narrower view of statistics than is adopted in this paper: Barnard expresses this most forceably. Our topic is the study of uncertainty generally, whereas a more data-oriented view would restrict statisticians to study only data uncertainty. I believe that statisticians, because of their expertise in probability and because probability is the way to describe uncertainty, can make a contribution to decision-making under uncertainty, as can the psychoanalyst. Some of this uncertainty may be influenced by data and, as Moore has explained, there is no sharp distinction between data uncertainties and others which are more subjective. Our reason for excluding data from the paper was not irrelevance but the additional complexities its inclusion would introduce. For example: it is not easy to see just what the connection is between the data on U.S. deaths (Table 1, column 8) and the corresponding π's or q's for You. Statisticians typically and tacitly assume You are exchangeable with the data but, again as Moore points out, this is not necessarily reasonable.

Cox, with his concern for data, wants us to find "probabilities meaningfully related to the real world". But why is a significance level more meaningful than a posterior probability? Is not Moss correct when he says that "coherence . . . may indeed be a *necessary* condition for . . . acceptability"? The difficulty over pooling that Cox mentions may arise through stating distributions that are too tight, and by failing to reflect on what data might occur and, if they did, what Your reaction would be. But no formalism is so wonderful that You should not sometimes be informal.

Moore and Horsnell raise questions concerning small probabilities. Often an event of small probability occurs because all of a number of events, each of themselves of appreciable probability, occur. Hence a small probability can sometimes be found as a product of appropriate conditional probabilities that may be easier to understand. Generally, the useful idea is to extend the conversation to include other events, as in Section 4.2 or, as Horsnell does, to look at the events in different ways.

Bartholomew's criticism of the additive correction is one that we share and was one reason for going to other metrics which seem less objectionable. To our shame, we had not thought of his simple scaling device. It appears to arise from the use of the metric $F(x) = x^{\frac{1}{2}}$ (in the notation of Section 4.1) for then the minimization equation (before (14)) is $(q_i^{\frac{1}{2}} - \pi_i^{\frac{1}{2}})/\pi_i^{\frac{1}{2}} = \lambda$ and $q_i = (\lambda+1)^2 \pi_i$. This corresponds to the variance of q being proportional to q so that the smaller probabilities have the smaller variances. McConway's contribution is illuminating but there are doubts concerning the commutativity assumption: is there not some loss of information in replacing $q_2, \ldots, q_n$ by Σq_i that might suggest a different reconciliation in the two cases?

Bromage's point is also a good one that might yield to analysis along the following lines. If the event that cannot be decided is of interest to You, then You will presumably be influenced in Your decision-making by its uncertainty. Consequently, by Your assessment of the value of these related decisions, a test can be made of Your probability for the original event. Thus the uncertainty about the authorship of the plays usually attributed to Shakespeare influences a decision about whether to visit Stratford or Lavenham, where the Earl of Oxford, who may have written them, lived. This is another example of extending the conversation. Buxton's point is another that might be surmounted by introducing additional features, as Barnard suggests. Buxton's resolution is unfortunate in that it introduces a new type of quantity, P_0; though who are we to protest at this when we have included π and probabilities of π to complicate an already complicated problem. Kadane comments on this complexity. Our response is that these elements reflect genuine features of the situation that it seems necessary to include, and moreover reflect features common to many situations: thus the meteorologist's $p(\pi \mid A)$ will do for many days, not just for one, so that, looked at in the context of a collection of decision problems, the complexity is not so great as a proportion of the total.

Aldous' work, reported by Kingman, could be of substantial value in our understanding of the two-way model, and hence of related models used in experimental design. It is not clear what are the implications of the parameters not being expressible as functions of the present and future data, nor why this should arise in the two-way, but not in the one-way, case.

We do apologize to Good for omitting to refer to his useful and much earlier work: and to Cox for forgetting his model. Our approach can incorporate his case and we only used the special case $\alpha = 0$, $\beta = 1$ (in his notation) for illustration.

Morris, French and Leonard refer to limitations on our formalism. They are right to do so. Formality has its place in guiding one along suitable paths of argument: and most of us need some guidance.

DeGroot is surely correct in asserting that an important element in our uncertainty about probabilities is how robust they are likely to be to future, likely data. But reflection can resurrect data. For example, I was once asked to provide my probability distribution for the number of English monarchs since William the Conqueror. (Do I hear someone say, what nonsense, go and consult the almanac?) Reflection involves remembering all their names; that is, resurrecting data that one has temporarily forgotten.

Moss and Dickey comment on our language. I accept their criticism but hope that the three ways of considering the quality of probability assessments will be understood even if incorrectly designated. The research programme that Dickey describes contains valuable material that nicely complements our own: for whilst he is interested in asking the subject (expert) questions meaningful to him, we are concerned with his ability to answer these questions. Thus, in generalization of the ideas in the paper, what is $p(\gamma \mid m)$, where m is the stated, and γ the true, median? Both types of study are needed before a satisfactory interrogation procedure is available. The difficulties in the multivariate case, mentioned by Leonard, are formidable. I agree with Winkler in suggesting that resolution may be useful when the q's are coherent and welcome his modification of French's suggestion.

Bayesians are not bookmakers. A coherent assessor who offers 2 : 1 against an event A is not only prepared to accept a £2 stake and pay out £4 if A occurs, but also to accept £2 and pay out £1 if A does not occur. Thus a Bayesian book can be made against Cedric Smith's proposer of odds. Bookmakers are in the business for money: Bayesians for truth.

Morris interprets the quote from Savage as having a wider meaning than perhaps we gave it. A coherent subject would not *knowingly* violate the axioms, but would correct any violations and feel happy in so doing. For even a single, continued violation would destroy the theory, just as a single plane triangle whose angles totalled 190 degrees would destroy Euclidean geometry.

What emerges from this discussion is a fairly general recognition that the problems of people's probability assessments form a topic worthy of scientific study. What is much less clear is whether the complicated methods of this paper form the basis of the best way of conducting the study. I share this conviction and these doubts. But, with Phillips, I feel that the notion of extending the conversation from the event of interest to other, related events, is likely to be an important, practical tool: and if that is all that emerges from this meeting, it will have been worthwhile.

R. V. Brown: Some of the issues raised in this discussion are addressed in a more general paper on the reconciliation of incoherent judgements including choice and value as well as probability judgements (Brown and Lindley, 1978). Two of these issues deserve brief discussion here—one bearing on the interpretation of π, the other bearing on the practical motivation for our enquiry.

Dr Phillips and others have questioned the existence of π that characterizes the intrinsic "correct" probability of a given subject. We have simply found it a convenient analytic construct, without which it is difficult to gain a hold on our problem. Indeed, its rejection poses serious problems in areas beyond the scope of this paper. For example, it is difficult to develop a systematic paradigm to evaluate improved analysis if there is no benchmark of perfect analysis to compare it with (Watson and Brown, 1978). The analogy is with the well-established use of perfect information as a construct for evaluating imperfect information.

Our interpretation of π is that it is the result of ultimate reconciliation, the end-product of applying infinite and impeccable pains as Dr Bunn has noted. It is the result of bringing to bear additional assessments without limit from the subject's psychological field, and processing them in an appropriate fashion—whatever that fashion may prove to be—provided that the resulting target assessment converges. Whether it does converge appears to depend on propositions both mathematical and empirical that are still to be established.

Suppose the acceptable procedure is a Bayesian updating paradigm such as we have exemplified in our paper. The original incoherent readings are reconciled by a prior and likelihood function from N, which is a partition of the subject's psychological field. In principle, N is incoherent, but his incoherence can be reconciled by raising the argument one stage further, by invoking a higher order N—who may again be incoherent. Without regard to practicality, the question is whether, in principle this process if pursued indefinitely, would produce a stable π. We have heard, and seriously entertained, arguments which say that the process blows up: given the way people are, empirically, incoherence gets worse at each higher stage because increasingly inaccessible assessments are called for. However, Professor Good's earlier reflections ("the higher the type, the less

the wooliness matters", Good, 1952) suggests that the process may still converge, due to the logical structure of the process.

Professor Dickey observes that "one's actual beliefs may not cohere into joint probabilities". Our interpretation of π is constrained to be coherent, and it might be visualized as the product of a two-tier process for removing errors. The first tier removes reporting error, and yields an accurate measurement of elements in the subject's head ("actual beliefs") which may be incoherent. The second tier, with which we have been principally concerned, removes what we might call an integration error and yields a coherent π.

The other general issue is the practical motivation for addressing the reconciliation problem which has been questioned by Professor C. A. B. Smith. It can be discussed in the context of decision analysis consulting work with U.S. Government agencies, which actually stimulated this research effort. In 1977, I worked with a senior civil servant at the Federal Energy Administration to produce probabilistic estimates of energy demand based on survey and other data.

Our general practice was—and is—to look for many divergent approaches to a controversial issue, and thereby deliberately invite incoherence—though at that time we had no formal justification for our intuition. The target quantity, L, was demand for lighting energy in 1978. There were perhaps 20 different ways of assessing it, but the two discussed here will illustrate the points involved. The demand for lighting could be assessed as being the product of lighting in 1968, for which a rather extensive study had been done by Stanford Research Institute, times the change in demand for lighting since that time. Based on available data, we helped the subject assign a joint probability distribution over those two arguments (in fact treating them as independent so the marginal distributions were sufficient). Using straightforward probability theory we derived an implied probability distribution on lighting demand, L', as required.

The alternative tack was to decompose lighting as: the number of users, times the bulbs per user, times hours per bulb, times kilowatts per hour. There was a great deal of objective data that bore on each of these components. When our subject assessed his uncertainty about these (generally as a joint distribution, but again, components were treated as an independent), a new implied distribution, L'', on lighting for 1978 was produced. The distributions barely overlapped, the lower tail of L' just touching the upper tail of L''. Our subject was faced with the dilemma of making up his own mind ($\hat{L}$) in the light of these conflicting cues and, more important, of communicating his conclusions defensibly to Congress. He pointed out that some highly controversial legislation was to be based on these lighting estimates. When he presented these estimates on Capitol Hill what was he to do? Was he to say that they had been derived in two different ways and that the answer was X plus or minus 20 per cent, or $3X$ plus or minus 20 per cent, so let us settle for $2X$? The potential for embarrassment was unmistakeable!

In seeking to guide him, our intuition prompted us to consider the space of possible reconciliations, that is coherent sets of component distributions, and select one set which implies a unique $\hat{L}$ distribution. That could be achieved informally in line with Dr Leonard's suggestion, by shifting some of the L' distributions downward and/or some of the others upward. We would presumably take into account the relative "sureness" of the original distributions, as we have quoted Savage as suggesting in our paper. We could iteratively move these distributions around with feedback on degree of coherence until we achieved two sets of component probability distributions implying identical target distributions for L' and L'', and therefore for $\hat{L}$.

We discussed this approach with our subject. He said that it sounded fine and plausible, but if he "jiggled" the numbers so that they came out together as $2X$ plus or minus 40 per cent, would he be able to defend this publicly? If an antagonistic Congressman were to check with acknowledged academic authorities, would he be assured that our procedure conformed to established statistical practice? Professor Barnard will not be surprised that we could give him no such assurance. So we resolved to seek out a reconciliation procedure which could be given a defensible rationale.

Nevertheless, if the subject's motivation is to clarify his own uncertainty, for some time to come his procedure will likely be to jiggle the numbers informally, intuitively bearing in mind that he wants to end up closer to those assessments of which he feels most sure.

References in the Discussion

Bacharach, M. (1975). Group decisions in the face of differences of opinion. *Manag. Sci.*, **22**, 182–191.
Bunn, D. W. (1978). *The Synthesis of Forecasting Models in Decision Analysis.* Basle: Birkhauser Verlag.
Cox, D. V. (1958). Two further applications of a model for binary regression. *Biometrika*, **45**, 562–565.

DAVIDSON, D. (1976). Psychology as philosophy. In *The Philosophy of Mind* (J. Glover, ed.). Oxford: Oxford University Press.
DAWID, A. P. (1977). Invariant distributions and analysis of variance models. *Biometrika*, **64**, 291–297.
DAWID, A. P., DICKEY, J. M. and KADANE, J. B. (1979). Matrix *t* and multivariate *t* assessment. Department of Statistics, University College of Wales, Aberystwyth.
DAWID, A. P., STONE, M. and ZIDEK, J. (1973). Marginalization paradoxes in Bayesian and structural inference (with Discussion). *J. R. Statist. Soc.* B, **35**, 189–233.
DEGROOT, M. H. (1974). Reaching a consensus. *J. Amer. Statist. Ass.*, **69**, 118–121.
DICKEY, J. M. (1979). Beliefs about beliefs, a theory for stochastic assessments of subjective probabilities. International Meeting on Bayesian Statistics, Valencia, Spain, 27 May–2 June 1979. To be published in *Trabajos de Estadistica.*
GOOD, I. J. (1950). *Probability and the Weighing of Evidence.* London: Griffin.
—— (1967). A Bayesian significance test for multinomial distributions. *J. R. Statist. Soc.* B. **29**, 399–431.
—— (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.*, **4**, 1159–1189.
—— (1979). On the combination of judgements concerning quantiles of a distribution with potential application to the estimation of mineral resources. *J. Statist. Comput. Simul.*, in the press.
GOOD, I. J. and GASKINS, R. A. (1979). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Ass.*, in the press.
HARMAN, A. J. and PRESS, S. L. (1978). Assessing technological advancement using groups of experts. In *Formal Methods in Policy Formulation* (D. W. Bunn and H. Thomas, eds). Basle: Birkhauser Verlag.
KADANE, J. B., DICKEY, J. M., WINKLER, R. L., SMITH, W. S. and PETERS, S. C. (1978). Interactive elicitation of opinion for a normal linear model. *J. Amer. Statist. Ass.*, submitted for publication.
KELLEY, G. A. (1955). *The Psychology of Personal Constructs*, Vols 1 and 2. Norton.
LINDLEY, D. V. (1971). Bayesian statistics, a review. In *Regional Conference Series in Applied Mathematics*, Vol. 2. Philadelphia: Society for Industrial and Applied Mathematics.
—— (1975). The future of statistics—a Bayesian 21st century. *Adv. Appl. Prob. Suppl.*, **7**, 106–115.
LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model (with Discussion). *J. R. Statist. Soc.* B, **34**, 1–41.
MACCORQUODALE, K. and MEEHL, P. E. (1948). Hypothetical constructs and intervening variables. In *Readings in the Philosophy of Science* (H. Feigl and M. Brodbeck, eds), pp. 596–611. New York: Appleton-Century-Crofts. (Reprinted from *Psychol. Rev.*, **55**.)
MORRIS, C. W. (1938a). *Foundations of the Theory of Signs.* Chicago: Chicago University Press. Reprinted in 1955 in *International Encyclopedia of Unified Science*, Vol. 1 (O. Neurath, R. Carnap and C. Morris, eds), pp. 77–137. Chicago: Chicago University Press.
—— (1938b). Scientific empiricism. In *Encyclopedia and Unified Science* by O. Neurath *et al.* Chicago: Chicago University Press. Reprinted in 1955 in *International Encyclopedia of Unified Science*, Vol. 1 (O. Neurath, R. Carnap and C. Morris, eds), pp. 1–75. Chicago: Chicago University Press.
SCHUM, D. A. (1977). Contrast effects in inference: on the conditioning of current evidence by prior evidence. *Organiz. Behav. & Human Perform.*, **18**, 217–253.
SEAVER, D. A. (1976). Assessing group preferences and uncertainty for decision making. Technical Report 76–4, Social Science Research Institute, University of Southern California.
SMITH, A. F. M. (1978). Comment in discussion of paper by M. S. Bartlett. *J. R. Statist. Soc.* B, **40**, 167–168.
WATSON, S. R. and BROWN, R. V. (1978). The valuation of decision analysis. *J. R. Statist. Soc.* A, **141**, 69–78.
WINKLER, R. L., SMITH, W. S. and KULKARNI, R. B. (1978). Adaptive forecasting models based on predictive distributions. *Manag. Sci.*, **24**, 977–986.

As a result of the ballot held during the meeting, the following were elected Fellows of the Society.

AHAMAD, Nasir B.
ANAJE, Stephen C.
BAILEY, Steven P.
BARLEY, Lynda M.
CRAIG, Ian S.
DADA, Michael O. A.
CHEN, Gina G.
DAVIES, Colin
EDELMAN, David B.
FORTNEY, William G.
GOLDBLATT, Peter O.
GUARD, Malcolm J.
GUERRERO-GUZMAN, Victor M.
HAJIJUBOK, Zainodin B.
HISCOCK, Sally C.
HULL, Derek L.
INSKIP, Hazel M.
JACOBS, Peter J.
KARAKOSTAS, Konstantinos
KEY, Peter B.
LAMBOY, Warren F.
LANGELAND, Thore
LILLYSTONE, Robert J.
MCNAB, Alan F.
MILES, David
OTUNUBI, Joel O.
PALMGREN, Juni
PANOUSIS, Mary E.
RAVENDRAN, Sivagnanasunderam
STONE, Janice M.
VERRILL, Steven P.
WEBB, William V.
WONG, Wing H.
YAN, Alan Tak-ming