DISCUSSION

J.M. DICKEY (*University College of Wales Aberystwyth*):

I find the paper by Dr. Leonard stimulating. Many of us would agree with the statement that there is much more to real inferences than is modeled by Bayes' theorem: for example, that a given subjective-probability distribution might be usefully conditioned on new data at a particular time, but that to continue formally updating it to a sequence of new data over a long period without rethinking the probability model would be foolish. It would even be foolish to rely on Bayes' theorem on a single occasion if one closed one's mind regarding the assumptionsused.

But, of course, it is not necessary to close ones mind, nor ones eyes and ears. Bayesian theory does not require that, although it may seem so to some authors because of the silence in Bayesian theory on the subject of how to think up new models. The implications of coherence for the subject of learning from data have to do with what attitudes to take regarding contingent bets, how to reason now about the information in future data. The axioms of coherent potential behaviour do not imply that, after the data is in, one should actually follow the previous plan in updating ones opinions. That is, probability conditioning (for example, Bayes' theorem) is not necessary for real opinions, but it provides a point of reference, a rational yardstick, a standard relationship between prior and posterior opinions. If ones opinions do not obey probability conditioning, then one looks for a reasonable probability model under which they do, or which implies opinions that one can reasonably adopt.

How should a Bayesian statistician look at his data to see **whether** he will need to think up new models? Karl Popper (1972) imagines scientific research as a continuing process of using experimental data to test the validity of theories which are then revised when rejected by tests. Popper's nonBayesian conception also suffers from silence on the subject of how to think up new models. Also, it inherits a defect from traditional statistical "data analysis" on the subject of how to decide whether a new model is needed. This defect in traditional tests for validity-checking of models was pointed out by Berkson (1938). In practice, no model ever tested is exactly true, and any prespecified model will be rejected for a large enough (fixed-size) sample. This makes acceptance of models largely a question of the size of samples taken. (See Kadane and Dickey, 1979, for a Bayesian discussion of this problem). Another question, for which a traditional statistician's answers can only be highly subjective when no alternative models have been suggested, is the more general question of which validity tests to perform. Which *experiment* to perform also remains largely subjective.

So traditional theory and Bayesian theory are both limited in the scope of their application. I think it is a mistake, though, to say that coherence implies complexity or that coherence misleads. Do the rules of logic or arithmetic mislead? Nor does IMP to my mind "oppose" coherence, unless Dr. Leonard insists on tying IMP to the Freudian notion of Id. I agree that IMP seems complex, but I call on Dr. Leonard and others to develop theory to shed light on its mysteries.

Dr. Leonard reminds us of the old question of discrimination methods versus regression analysis. It is really simpler for the statistician to specify $p\ (\mathbf{x}|\Lambda)$ than $p\ (y|\mathbf{x})$? I note that he suggest the use of estimated sampling probabilities to approximate predictive probabilities, while Aitchison and Dunsmore (1975) recommend the use of predictive probabilities to estimate sampling probabilities.

Finally, it is claimed that the Bayes factor is sensitive to the choice of conditional prior density, and increasingly so for increasing sample size. Of course, in practice the Bayes factor goes to zero or to infinity as sample size increases. A very small or very large Bayes factor is strong evidence for or against the more complicated model, respectively. So it remains to be shown that the "sensitivity" happens before the evidence becomes too strong to be refuted by the changes in the Bayes factor wrought by reasonable perturbations in the prior density.

My comments on Professor Novick's paper joint with Dekeyrel and Chuang would seen to apply with equal force had the paper been concerned with probability assessments, rather than utility assessments. (Utilities are equivalent to probabilities in technical senses, and this equivalence is exploited in their assessment methods). Therefore, I should like the authors to consider my comments with an eye to the possibility that I have failed to appreciate properties inherent only to utilities. Perhaps they would bring out the important differences in their reply to this discussion.

The methods given are ingenious and rather elegant. A person wishing to use them to assess his own utilities would, I feel sure, need to spend appreciable time and effort learning to use them as effective tools. The worry, of course, is that in so doing the person may acquire bad habits or "biases" that would connect up his different uses of the tool, rather than connecting together the tool and his underlying utilities.

Instead of a "person", the authors refer to a "subject". This latter term has been reserved in the psychological literature to mean the same as "object", in the spirit of conceiving persons other than oneself as machines. One trouble with this conception is that it just does not work well, except at a mere physiological level. Persons do not behave predictably without reference to context, including the histories of their personal attitudes and social settings (Kelly, 1955). Experiments tend to be aimed at discovering simple universal context-free laws of behavior, such as, laws that would favour this assessment tool over that one. What is it that justifies our thinking that isolated laboratory experiments will yield psychological findings of any importance in real-world applications?

In spite of the doubts expressed here, I should like to urge the authors to carry out the experiments envisaged, preferably in real applications.

W.H. DUMOUCHEL (*Massachusetts Institute of Technology*):

Professor Leonard's emphasis on the necessity to develop workable procedures, and to show our colleagues that they do work, is well put, in my opinion. More focus is needed on what we can do, rather than too much concentration on the logical inconsistencies of classical statistics. Strict consistency is often unattainable in the real world. For example, we all know that prior distributions cannot logically depend on the data. Yet Professor Leonard rightly points out that most responsible statisticians, Bayesian or not, will try to obtain a "feel" for the data with plots, etc., before inducing a likelihood function or even deciding on a parameter space. However, I am not so pessimistic as to rule out a useful Bayesian approach to many "global" problems. Often a mixture of two or three models can quite well capture the essentials of even a fairly complicated situation, and thus help derive real-life conclusions from the data. The binomial example of section 4 does not seem convincing to me. The situation is that of choosing between H_0 and H_1 based on the observations of n exchangeable observations of 0 or 1, whose sum is x

$$\begin{aligned}&H_0 : x \sim P_0(x)\\ &H_1 : x| \sim Bin(n,\theta)\\ &\quad x = 0,1,\ldots,n \qquad \theta \sim \text{Beta}\,(\alpha,\beta)\end{aligned}$$

The supposed paradox is that posterior odds ratio of H_0 *vs* H_1 depends importantly on α and β even as $n \to \infty$, especially if x/n is far from $\alpha/(\alpha+\beta)$. But the fact of n being large here does not reasonably imply that the sample information should "swamp" the prior information. When alternative H_1 is true and n is large, the variation in $\theta = x/n$ is negligible conditional on θ, so that the relevant comparison is

$$\begin{aligned}&H_0 : \theta \sim P_0(n\theta)\\ &H_1 : \theta \sim \text{Beta}\,(\alpha,\beta)\end{aligned}$$

Thus the problem is more like that of deciding whether a single observation could have a particular beta distribution, and naturally the parameters of that beta distribution would play an important role in the decision.

On another point, in spite of my own liking for logit probability models, I suspect they are being oversold in section 7. The author's distinction between a probabilistic and a predictive model eludes me. Two possible interpretations are: (1) the full information versus conditional information approach to contingency tables, or (2) the errors in variables problem of regression. But the further discussion doesn't seem relevant to either interpretation. The author seems to imply that multivariate density estimation is simpler and more reliable than more common procedures such as stepwise regression. I would guess that use of one of the various robust regression techniques now widely available would be more fruitful than abandoning the ordinal structure of the response variable in favor of a purely categorical-data approach.

Finally, as an argument against constructing unnecessarily complicated models, the author states in section 8 that modeling a thick-tailed distribution is unnecessary if, even with a normal model, the real-life conclusions are the same with and without inclusion of the outliers in the analysis. This cannot be true in general, as the following example shows. Suppose that a sample of size $n = 100$ has mean 0 and standard deviation 1, with one or two outliers near the value $x = 4$. Suppose further that the real-life problem is to decide whether Prob $(X>4)<.001$. Then a normal model including the outliers would estimate Prob $(X>4)<10^{-4}$, while excluding the outliers would result in a smaller sample standard deviation and an even smaller estimate for Prob$(X>4)$. Yet fitting the data to most families of thicktailed distributions would estimate Prob$(X>4)$ to be near the sample proportion, namely 0.01.

Professor Novick and his co-authors are to be commended for continuing to explore a topic so vital to the practical functioning of the Bayesian method. Until we can show how prior opinion can be elicited in a workable fashion, the subjective Bayesian viewpoint can hardly proliferate. The present paper considers with care and sophistication a simple problem involving a single, ordered attribute, and makes us very conscious of how much harder a more realistic elicitation involving several dimensions and a complex data set would be. The work of Kadane *et.al.* (1979) combined with the present paper provide a start toward computerizing this process.

The author's references to the work of Amos Tversky and his associates are welcome. Certainty bias and anchoring bias are present not only in elicitation problems, and overcoming them can be used as a theme for data analysis in general. Whenever we tell our elementary statistics classes to be more conscious of variation, we are fighting the certainty bias, and when we teach proper methods of estimation we counter the anchoring bias. But Bayesian methodology is peculiarly affected, on a second level, by these tendencies. A stronger potential barrier to solution of the elicitation problem is raised by the work of Shafer (1976) who argues that human opinions are too complicated to be represented by simple probability distributions or utility functions. I would be interested to know if experiments such as the present authors are performing could be designed to test this or similar propositions.

In any event, the "local", "regional", and "ends in" procedures presented here seem reasonable and clever and I am looking forward to the results of the authors' future experiments. There are just a few more specific questions that come to mind:

a) What if the ordering of the states is not prescribed? Would your methods change?

b) Although elicitation of probabilities is formally identical with the elicitation of utilities, the psychological reactions of subjects may differ for the two tasks. Is there any evidence of this?

c) What evidence is there that the regression on the log odds scale is optimal for the coherence checking algorithm? Might some weighted regression be better? Is the standard error of the residuals a useful number?

c) How much real time do these elicitations take? How long for a novice to elicit all the factors for the probabilities in a 2x2 table, and what fraction of·them show noticable fatigue and/or boredom before finishing?

I hope that these questions will help stimulate the authors to continue their interesting work.

J.M. BERNARDO (*University of Valencia*):

I certainly believe that the idea used by Professor Novick of requiring the decision maker to give more than the minimum number of judgments in fitting a personal probability distribution or utility function is important and very useful. I wonder however what is the coherent justification for using least squares in order to force coherence among those judgments

S. FRENCH (*University of Manchester*):

Firstly, perhaps Dr. Leonard will forgive my pointing to an unfortunate omission in his paper. In quoting DeGroot's axiom system for subjective probability, he omits the CP axiom (DeGroot (1970), Chapter 6). It is the CP axiom that introduces the notion of conditional probability and hence justifies the use of Bayes Theorem. Without the CP axiom this system does not pretend to justify Bayesian inference. If Dr. Leonard wishes to criticise the use of axiom systems, he really should cite a whole system.

Turning now to the paper of Novick, Dekeyrel and Chuang, I have two questions that I should like to ask. First, in the fixed state method of assessment the values $U(\theta_o) = 0$, $U(\theta_N) = 1$ are fixed. The values of $U(\theta_n)$ for intermediate n are determined by relations of the form

$$U(\theta_n) = p_n U(\theta_{n+1}) + (1-p_n) U(\theta_{n-1}) \qquad (*)$$

Now, since the paper's very essence is to admit incoherence on the part of the decision maker's statements, it must be admitted that the p_n are "in error". Does this error transmit itself evenly to the determination of $U(\theta_n)$ or does the error on the $U(\theta_n)$ rise steadily from 0 on $U(\theta_o)$ to a maximum on $U(\theta_{N/2})$ before falling away to 0 again on $U(\theta_N)$? There is a relevant passage in Spetzler (1968) in which he discusses the relative merits of three different methods of measuring utility.

My second question concerns the decision makers' role in the resolution of incoherence. For me one of the basic aims of decision analysis is to bring understanding. In particular the process of introspection is not simply one of measuring utilities and subjective probabilities. Rather it is a process that helps the decision maker explore his preference belief structure, discover inconsistencies, think about them and then resolve them. It seems imperative to me that of method of construnting a decision maker's utility function should always refer back to him any discovered inconsistency so that he may reconsider his preferences. Only when all the inconsistencies are of such a slight nature that it is beyond the decision maker's powers of discrimination to resclve them, should an automatic resolution process be invoked. Do I understand that the authors' procedure does in fact do this, namely only use least squares with coherence constraints as a tidying up device having left all the major resoution of inconsistency to the decision maker?

J.B. KADANE (*Carnegie-Mellon University*):

In the discussion, both Dennis Lindley and Bruce Hill strongly criticized Tom Leonard's paper for not being sufficiently Bayesian. In doing so, I think that they have overreacted. When a Bayesian does statistical modelling and data analysis, compromises are often necessary to keep control of the analysis, to separate what is important from what is not.

To associate Tom Leonard's position in this paper with Glenn Shafer's, as did Bruce Hill, is to mix two very different positions, I think. As I understand Shafer's ideas, he rejects Bayes Theorem and the Bayesian paradigm as a theory. This seems to me very different from Leonard's position, which keeps Bayesian theory as essential background for doing statistics. To associate these positions does an injustice to both Leonard's and Shafer's positions.

D.V. LINDLEY (*University College London*):

I find myself in almost total disagreement with the views expressed in Leonard's paper. Coherence becomes more important the bigger the situation, not less. If only one uncertain event is assessed, then coherence does nothing more than assert that the descriptive number lies between 0 and 1. With two events, A and B, coherence begins to play a more important role: for example, $p(AB) = p(A)p(B|A)$. The more events, the more opportunity there is to exploit coherence and the more necessary it becomes to do so.

Perhaps it is this fallacious view that leads to Leonard attaching importance to Axiom 5. All this axiom does is to tie probability to a numbering system: the multiplication and addition rules, the rules of coherence, are really contained in the earlier, important axioms and his omitted axiom of called-off bets. Probability is not just a number between 0 and 1: it is a number obeying two important rules of combination.

A. O'HAGAN, (*University of Warwick):*

Dr. Leonard's *IMP*s are of course an over-complication, providing no real insight into the processes of practical statistics. But if we regard them as merely a thin excuse for presenting a miscellany of ideas - his sections 4 to 12 then there is much food for thought in his paper. I would like to examine just a few of the snapshots in Dr. Leonard's album.

His skewed-normal distribution of section 8 is ingenious, but I wonder if some of his criteria (i) to (vii) were chosen *a posteriori*. I commend to him the skew distribution derived in O'Hagan and Leonard (1976), for which I think we could draw up an equally impressive list of criteria. For instance it is more tractable than the skewed-normal.

The sensitivity of the Bayes factor (4.3) to the prior hyperparameter a in his binomial example of section 4 could be quite worrying. Some insight is obtained initially by ignoring θ. Since $D(\alpha,\beta)$ is simply the prior (marginal) probability of the frequency x under the binomial model, we are just comparing the two simple hypotheses given by the distributions $p_0(x)$ and $p_1(x) = D(\alpha,\beta)$. The observed value of x discriminates strongly between the hypotheses if the ratio $R_x = p_0(x)/p_1(x)$ is very large or very small. Dr. Leonard introduces a third hypothesis, that x has distribution $p_2(x) = D(\alpha + 1,\beta)$ and observes that it may be possible to find an x which does not discriminate strongly between p_0 and p_1 but does discriminate strongly between p_0 and p_2. He does this by showing that the ratio (4.5) can give an x that discriminates strongly between p_1 and p_2. His thesis is that this odd because p_1 and p_2 are very similar. But with most parametric families of distributions we can find observations discriminating strongly between any two members of the family, however close their parameter values may be. Consider for example the distributions $N(0,1)$ and $N(\epsilon, 1)$: however small $|\epsilon| > 0$ is, as x tends to infinity the likelihood ratio

$$\exp\{-\tfrac{1}{2}x^2 + \tfrac{1}{2}(x-\epsilon)^2\} = \exp(-x\epsilon + \tfrac{1}{2}\epsilon^2)$$

tends either to zero or to infinity. Almost all the parametric families in common use have monotone likelihood ratios (see Lehmann (1959)) and in most cases the likelihood ratio is unbounded. In fact, since Dr. Leonard's beta-binomial has a bounded likelihood ratio (for given sample size), he has chosen one of the less convincing examples of "sensitivity". Other examples may be constructed similarly "— $p_1(x)$"— is formed from a prior distribution for a scalar parameter θ indexed by a prior hyperparameter ϕ, and a sampling distribution for x given θ. Whenever these two distributions have monotone likelihood ratios, e.g. any two exponential-family distributions (Lehmann, p. 70), then $p_1(x)$ will have a monotone likelihood ratio in ϕ (Lehmann, p. 343 problem 7).

Therefore, Dr. Leonard's sensitivity problem arises whenever we deal only in exponential families. Having seen the "problem" in the above terms I feel that it is not as unreasonable as he implies, but I do think that it is important to recognise that nearly all commonly used distributions will lead to this kind of behaviour and that radically different behaviour is possible using distributions with non-monotone likelihood ratios. In O'Hagan (1979), and more explicitly in a follow-up paper submitted to the Annals of Statistics, I have made this point in connection with a different kind of behaviour which always results from using distributions with monotone likelihood ratios, and not otherwise. In his section 9, Dr. Leonard criticises exponential families on even more fundamental grounds. It is time that we looked very seriously beyond the convenient, tractable exponential families because they are severely limiting the kinds of inference that we can make.

A.F.M. SMITH *(University of Nottingham):*

Leonard seems to be making two rather strong attacks on the axioms. If I understand him correctly, he states that:

(i) the straightforward claims set out in 2*a*) and 2*b*) are much more directly *compelling* to clients than are the axioms; and, in any case, they are more *honest*;

(ii) the axioms are tautologous.

Let us first consider (i), and recall that statement 2*a*) invokes the phrases "much more reasonable", while statement 2*b*) refers to "superior practical results". Does Tom Leonard really believe that these particular phrases can (honestly) command general acceptance as having directly obvious meanings that require no further analysis? And if someone refuses to accept these as primitive terms of reference, I think I know where Tom Leonard would eventually end up in attempting an unambiguous explication of "reasonable" and "superior" - back at this axiom system!.

The criticism in (ii) seems most peculiar!. *Theorems* deduced from the axioms are, *of course,* "contained in" them in the sense Tom Leonard presumably intends. But, surely, the (for us) rather profound methodological implications - the likelihood principle, the need to integrate out nuisance parameters - are *in no way* obviously "contained in" the axioms in the sense that they are directly intuited (or guessed, even) by someone who contemplates the axioms?

T.W.F. STROUD *(Queen's University Canada)*:

Leonard's article presents a refreshing relief from doctrinaire approaches which begin with a statement of the statistician's model and his prior beliefs about the parameters of the model. In fact, the statistician always has to begin with a real-life process and, hence, any model concerning this process (and, consequently, any prior distribution on the parameters of such a model) must be regarded as very tentative.

Sections 4 and 5 focus on some important facts often overlooked by Bayesian statisticians. In Section 4 it is pointed out that probabilities associated with choosing *between* models may be quite sensitive to the choice of prior distributions *within* models. Because inference *within* a model is insensitive to prior information when samples are large, it is easy to think that in large samples the prior doesn't matter. But the thing which makes the prior not matter is the likelihood, which is completely model-based. The example presented in Section 4 shows that, in situations where the prior mean within the binomial model ξ is very different from the sample mean p, the information in the data which is ancillary to the binomial model (which is what we need for testing the model) may *not* swamp out the prior in moderately large samples.

In Section 10, which deals with problems involving hyperparameters, the method of maximizing the marginal likelihood is advocated as an alternative to specifying "complicated and possibly confusing" prior distributions on the hyperparameters. Whereas in many problems maximizing the marginal likelihood gives virtually the same answer as integrating over a locally uninformative prior on the hyperparameters, no justification has been given that the former procedure is anything but a convenient approximation to the latter. In some cases, the approximation may be poor. For example, in the normal one-way classification shrunken estimates of the group means toward the grand mean may be obtained by putting a conjugate prior on the exchangeable group means and estimating the hyperparameters in this prior by maximum likelihood (Stroud, 1980). But if the number of groups is small (say 3 or 4), this procedure shrinks too much toward the grand mean because the likelihood function of the between-within variance ratio is skewed, causing the mode to underestimate this variance ratio. A similar problem exists if one uses a prior on the hyperparameters but then resorts to substituting the posterior modal values of hyperparameters, rather than integrating over them. In such cases where skewness causes a problem one should either integrate out the hyperparameters or devise a technique for suitably adjusting the modal estimates in the direction of the skewness.

REPLY TO THE DISCUSSION

T. LEONARD *(University of Warwick)*:

Many thanks to the discussants for their helpful contributions which seem to provide a good representation of current Bayesian thought about the area of Statistics. Since the conference Dennis Lindley and I have corresponded in detail about the axioms, and this has helped us to clarity our ideas in this area.

A positive contribution of this correspondence was an indication that my Axiom 5*a* is not needed in the very strictest mathematical sense, as De Groot utilizes the mathematical properties of random variables to their fullest extent (they are A-measurable functions from the parameter space to the real line). However, if the outcomes of the auxiliary experiment were simply regarded as numerical values, then my Axiom 5*a* would be needed to link the auxiliary experiment with the parameter space: it is this interpretation which the probability assessor would utilize when actually carrying out the suggested procedure. Moreover, my axioms 5 and 5*a* are equivalent mathematically to the combination of De Groot's Axiom 5, and his assumption of A-measurability of the random variable. Therefore my comments are relevant whichever interpretation is used; it is my firm understanding that the combination of the first four axioms with the assumptions surrounding the fifth axiom should be viewed in an inductive sense as virtually as strong as the final result. I would however like to thank Dennis for indicating the desirability of clarification of this mathematical point.

It still seems completely obvious to me that the axioms are not really proving much, but simply describing a way of thinking. During my correspondence with Dennis he suggested various sensible changes to the axioms, but despite about half-a-dozen intuitively appealing suggestions at least one of the axioms always turned out upon close scrutiny to be similar in strength to De Groot's fifth axiom. It is interesting that whilst recently teaching utility theory, I decided to play the role of a formal Bayesian, but this approach was quickly shown to be deficient by a series of simple and unprompted questions from my students; these were much on the same lines as the points I have raised here about subjective probability.

Dennis seems to have dodged the real issue - my main point is that coherence is less important and even constrictive in practical situations where the objective is to extract real-life conclusions from a data set. Probably we Bayesians should leave our ivory towers once in a while and work in a Statistical Laboratory analyzing real data. We might then learn that modelling is the really important part of statistics; analyses which proceed conditionally upon the choice of model are enjoyable but do not provide the complete answer.

I would like to thank Tony O'Hagan for his comments. I don't think that my *IMP*'s are an over-theoretisation - in fact there're not really a theoretisation at all! They are just a way of thinking, or perhaps a term to describe what most of us have been doing anyway. My point is that thinking about the problem in order to extract a model or a conclusion is much more important than trying to be formally coherent. Tony's comments on the sensitivity problem are helpful and interesting. His work on outlier behaviour would be useful if it were possible to find families of distributions with thick tails which are both meaningful and analytically tractable, for example, in multivariate situations.

I'm a bit confused by Simon French's comments. I didn't use the conditional probability axiom because I was just discussing straight-forward probability. I think however that my main points would extend to this situation.

Tom Stroud's thinking seems to be on similar lines to my own - we should probably form a clique of pragmatic Bayesians (this may be a good time to announce the foundation of the Bayesian-Fisherian school of statistics!). It is possible to justify estimating hyperparameters by their marginal likelihood estimates when the number of first-stage parameters is greater than about ten, because the estimates will then approximate the Bayes estimates under a wide range of loss functions. When the dimensions are smaller the estimates are less precise but still fairly sensible. A more sophisticated estimation procedure would in this case probably not be justified in view of the small amount of information available about the hyperparameters.

Bill DuMouchel's comments are very helpful and I'm glad that he supports the main theme of my paper. I remain a bit pessimistic about a mixed model approach since it would not be particularly meaningful or easy to check out each of the candidate models against the data or to think in a lucid way about the complicated analysis employed. It is interesting that he indicates that the binomial hypothesis testing problem is similar to deciding whether a single observation could have a particular beta distribution - this really supports my argument since it tells us that the standard Bayesian procedure for this situation can't properly distinguish between the two hypotheses.

My distinction between probabilistic and predictive models is a practical one. For many data sets the explanatory variables are extremely noisy so that it is virtually impossible to find a least squares model via standard procedures like stepwise regression, and therefore difficult to get reasonable numerical predictions of further dependent variables. However the data may still be rich in a content of a probabilistic nature, in the sense that they indicate how much the statistician should adjust his probabilities about the dependent variables, in the light of knowledge of the explanatory variables. In such circumstances, where we just can't find a reasonable least squares model, we can often still arrive at useful conclusions by modelling the distributions of the important explanatory variables.

I am not arguing completely against the use of thick-tailed distributions, but simply saying that if we look at the data and think about the problem then we can sometimes avoid this extra complication. In the example Bill discusses, I guess that most of us would prefer a much smaller value for prob($X > 4$) than 0.01.

Adrian Smith feels that my implication that the axioms of coherence are tautologous is most peculiar. This is probably because, like Dennis, he is thinking deductively rather than inductively - if we constrain ourselves to Bayesian formalism then statements by more open and inductive thinkers will very often appear to be peculiar. As I see it, if we look at the axioms and judge intuitively the strength of what is being assumed, and next look inductively at the strength of the final result, then the two appraisals will be extremely similar. Therefore the fact that the axioms deductively imply the final result does not really give us much - it would be inductively speaking just as reasonable to assume the final result to start off with. It's a pity that neither Dennis nor Adrian have taken this opportunity to look deeply enough at the problem to be able to give a definitive answer to this point.

I can't see how the likelihood principle follows from the axioms unless coherence is also assumed across an n-dimensional sample space in order to justify the existence of a sampling distribution - an extremely complicated assumption (don't the sufficiency principle and the very complex conditionality principle come into it as well?). The assumption that we can marginalise subjective distributions is barely stronger than the axioms that might be used to justify this procedure.

Further analyses are of course needed to justify statements like "superior practical results", but I think that this has already been done - see for example the work by Adrian and others on multi-parameter estimation, time series analysis, and categorical data. I personally think that the Bayesian approach is "much more reasonable" because it is extremely natural to think in terms of probability distributions when updating information about quantities of interest.

My thanks to Jim Dickey and Jay Kadane for their contributions. On the question of discrimination methods versus regression analysis it is indeed much simpler in many situations to model the distributions of the explanatory variables. Of course, one should always choose the method which best suits the practical situation at hand.

I would finally like to say how much I enjoyed giving a paper in the same session as Mel Novick. His practical implementation on CADA of my early marginalization work on categorical data fits in well with the things I have been trying to say.

M.R. NOVICK *(University of Iowa)*:

The commentary provided by Professors Bernardo, Dickey, Du Mouchel and French, are useful in themselves, but to me they have the added value of opening up for discussion some topics that I might have covered in my original presentation, had time and foresight permitted.

Professor Bernardo notes, with bated foil, that there may be no "coherent justifications for using least-squares in order to force coherence among ... judgments". He is, of course, correct. The only reply is that coherence, like virtue, can be absolute only in contemplation and is more likely to be compelling as we examine the actions of others rather than ourselves. Wisdom must guide us in knowing when small deficiencies in coherence (and virtue) can be tolerated.

The essence of Professor Dickey's critique of our paper is summarized in his question: "What is it that justifies our thinking that isolated laboratory experiments will yield findings of any importance in real-world applications?" Feelings of inadequacy in my ability to contribute anything *new* to the discussion of *that* question compel me to refer Professor Dickey to his biologist, chemist, physicist, psychologist, et. al. friends, some of whom may be willing to take the time to instruct him on the general decline in acceptance of the Kantian view of science and the acceptance since the end of the Dark Ages of the value of laboratory experimentation. For my own part I shall borrow Professor Bernardo's bated foil and ask Professor Dickey, "What is it that justifies *his* thinking that the mathematical derivations *he* presents us without *any* empirical investigation of relevance, will provide us with useful methods of assessing prior probabilities?" Perhaps Professor Dickey and I are both guilty of demanding a higher level of virtue and coherence of others than of ourselves. For my part I speculate that Professor Dickey's work will be very useful but question the appropriateness of his presupposition.

Professor Dickey, however, is not entirely off the mark. We have found that our methods are "successful" only when we go to great lengths, in our laboratory, to simulate practical decision problems. People do not carry around utility functions in their heads and we ought not to view the assessment process simply as a psychological measurement (psychometric) problem. However, we have also found that the nature of the graphic display has significant influence on assessors responses and that the anchoring effect can be reduced by the methods we propose. We also believe that further refinements will be useful.

Professor Du Mouchel's comments are more penetrating and require more detailed response. It is true that human opinions can be very complicated. Part of that complication is due to incoherence which, it is hoped, can be reduced through computer interaction. It is also true that humans attempt to uncomplicate their opinions and decision processes by the use of simplifying heuristics. Unfortunately these heuristics *typically* introduce bias. Our goal is to uncomplicate human opinion by providing alternative heuristics that avoid major biasing effects. This is not a simple task and we make no claim of "complete" success. But if, in education, I had to choose between decision-making with or without the prior probability, utility assessment, and decision-making procedures now available on the Computer-Assisted Data Analysis (CADA) Monitor I would certainly opt to use CADA.

With respect to Professor Du Mouchel's question as to whether experiments could be designed to test whether human opinions are too complicated to be represented by simple probability distributions or utility functions, I would respond that I think rather different experiments are necessary. I personally accept the notion that human opinion is too complicated to be so modelled. The point, however, is that what we seek is *not* a descriptive modelling of what human opinion *is*, but a normative modelling of what a particular human being's opinion "ought" to be. The word "ought" here has a special meaning that must be made precise. A human being's opinion "ought" to be internally coherent and ought to be consistent with contemplated behavior. If contemplated behavior is inconsistent no formal modelling with a probability distribution or utility function is possible. Thus probability and utility assessment procedures do not involve descriptive modelling. They involve a process that changes opinions in some way that results in internal coherence without changing those aspects of contemplated behavior that most clearly represent the person's opinions regarding the real world.

I now respond to Professor Du Mouchel's specific questions a) to d):

a) If states are not ordered we begin by ordering them.

b) All of our elicitation procedures require probability judgements (fixed state as opposed to fixed probability). We believe that the direct elicitation of utilities is deceptively easy but subject to a high degree of artifactual bias.

c) We have our intuition and some informed observation to suggest benefit from the log-odds scale for the regression of probabilities. I have very high personal probability that this is very much better than least-squares in the original metric. However, I would think that somewhat less weight on the extreme values might be useful. Dennis, Lindley and I have often debated the relative benefits of log-odds and root inverse sine transformations.

d) For most problems that we have adressed to date elicitations are handled quickly, with perhaps 10% of subjects showing boredom, fatigue, or uncorrectable incoherence. (For some this result may be endemic to the laboratory context which remains somewhat artificial despite our best efforts). The key to success with such methods is the moderate realism of the established scenario and the smoothness of the person/machine interaction. But our degree of success does also vary with the complexity of the model. A nine point unidimensional utility assessment is comfortable. A bivariate utility assessment is more difficult. Higher dimensional assessment is currently beyond our ability. (We have not been impressed by the mathematically convenient but largely unrealistic assumptions that others have chosen to make). The interrogation procedure for multiple linear regression originally programmed following the Kadane et. al. suggestions proved inadequate. However, Dr. James Chen of my staff has now produced an acceptable program which is tolerated by keen investigators, but is still wearisome for most users. Further improvements will need to be made.

Finally, let me adress Professor French's useful queries. Professor Lindley and I showed in our original paper that the value of P_n effected $U(\theta_n)$ most with decreasing effect for more distant values of θ_i. This is, I think, a desirable property, though independence for $i \neq n$ would be preferable.

Professor French's second query gets to the heart of our methods and I am grateful to him for raising the issue because I neglected this vital point in my presentation. (I really ought not assume that everyone is familiar with our CADA project). If I may borrow Professor French's words, the primary function of elicitation procedures on CADA is to help the "decision maker explore his preference belief structure, discover inconsistencies, think about them and then resolve them". We believe that this process is facilitated by conversational language computer interaction. Descriptions of CADA are contained in my article on CADA in the *International Statistical Review*, 1973, my article in the *American Statistician* in 1975 and a second article in the *American Statistician* to appear in November, 1979.

REFERENCES IN THE DISCUSSION

AITCHISON, J. and DUNSMORE, I.R. (1975) *Statistical Prediction Analysis.* Cambridge: University Press.

BERKSON, J. (1938) Some difficulties of interpretation encountered in the application of the chi-squared test. *J. Amer. Statist. Assoc.* **33**, 526-42.

DEGROOT, M.H. (1970) *Optimal Statistical Decisions*, Reading, Mass: Addision-Wesley.

KADANE, J.B. and DICKEY, J.M. (1979) Bayesian decision theory and the simplification of models. *Evaluation of Economic Models.* (J. Kmenta and J. Ramsey, Eds.) New York: Academic Press.

KADANE, J.B., DICKEY, J.M., WINKLER, R.L., SMITH, and PETERS, S.C., 1979. *Tech. Rep.* **150**, Carnegie-Mellon.

KELLY, G.A. (1955) *The Psychology of Personal Constructs.* New York: Norton.

LEHMANN, E.L. (1959) *Testing Statistical Hypotheses.* New York: Wiley.

O'HAGAN, A (1979) On outlier rejection phenomena in Bayes inference. *J. Roy. Statist. Soc. B.* **41**. 358-367.

O'HAGAN, A. and LEONARD, T. (1976) Bayes estimation subject to uncertainty about parameter constraints. *Biometrika* **63**, 201-203.

POPPER, K. (1972) *Objective Knowledge: An Evolutionary Approach.* Oxford: Clarendon Press.

SHAFER, G.: (1976) *A Mathematical Theory of Evidence.* Princenton: University Press.

SPETZLER, C.S. (1968) I.E.E.E. Trans. Systems, Science Cybernetics. SSC-4, 297-300.

STROUD, T.W.F. (1980) Empirical Bayes versions of Stein-type estimators. *Tech. Rep.* Stanford University.