Experiences in Elicitation

Joseph B. Kadane; Lara J. Wolfson

The Statistician, Vol. 47, No. 1 (1998), 3-19.

Stable URL:
http://links.jstor.org/sici?sici=0039-0526%281998%2947%3A1%3C3%3AEIE%3E2.0.CO%3B2-8

The Statistician is currently published by Royal Statistical Society.

The Statistician (1998)
47, *Part* 1, *pp.* 3–19

Experiences in elicitation

Joseph B. Kadane†

Carnegie Mellon University, Pittsburgh, USA

and Lara J. Wolfson

University of Waterloo, Canada

[*Read before* The Royal Statistical Society *at a meeting on 'Elicitation' on Wednesday, April 16th, 1997, the President*, Professor A. F. M. Smith, *in the Chair*]

Summary. Elicitation of expert opinion is becoming increasingly important in the elicitation of prior distributions. In this paper, the psychology of elicitation and the currently available methods are briefly reviewed, but the primary discussion is on the distinction between 'general' elicitation methods for a class of problems and 'application-specific' methods which are useful only once. Examples of both types of elicitation are given, along with a discussion about general *versus* application-specific methods, and predictive *versus* structural elicitation.

Keywords: Demography; Expert opinion; Prior probability; Probability assessment; Time series

1. Introduction

Elicitation of opinion (and, where appropriate, losses or utilities) is a critical step for making subjective Bayesian inference operational. Although most of this paper addresses 'how', it is appropriate first to consider 'why' and 'who'.

If one finds the de Finetti–Savage (Savage, 1954) view of statistics attractive, serious consideration must be given to how to obtain prior distributions for parameters (and also likelihoods). Although the subjective approach relieves the analysis of the (unsatisfiable) demand for objectivity, it leaves open the question of how to go about obtaining prior information in a form that is useful for analysis.

In principle, a Bayesian analysis of any subject using any person's prior distributions may be coherent, and hence 'valid'. However, not all coherent analyses have an equal claim on the attentions of the reading public. We believe that expert opinion is particularly interesting, because it offers the possibility that the information it conveys is especially useful to others. To whom expertise in a particular subject is properly ascribed is not a matter about which we can claim special knowledge. Whatever, operationally, one might mean by 'expert' in a particular context, it is reasonable to hope that they will have thought harder, and over a longer period of time, about the subject at hand than others have. Meyer and Booker (1991) and Chaloner (1996) provide further discussion of the merits of expert opinion.

Expertise in a subject-matter is not the same as expertise in statistics and probability. The goal of elicitation, as we see it, is to make it as easy as possible for subject-matter experts to tell us

†*Address for correspondence*: Department of Statistics, 232 Baker Hall, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA.
E-mail: kadane@stat.cmu.edu

 0039–0526/98/47003

what they believe, in probabilistic terms, while reducing how much they need to know about probability theory to do so.

We give here a brief overview of some of the heuristics to be aware of in probability elicitation; for more on this topic, see Wright and Ayton (1994), Wilson (1994), Chaloner (1996), Hogarth (1975) and the other references cited. The pitfalls to be aware of are

(a) *availability*—assessors link their probabilities to the frequency with which they can recall an event (Tversky and Kahneman, 1974),
(b) *adjustment and anchoring*—judgments are anchored at some starting value and adjust outwards, usually insufficiently (Winkler, 1967a, b; Alpert and Raiffa, 1982),
(c) *overconfidence*—difficulty in assessing the tails of a distribution (Wallsten and Budescu, 1983),
(d) *conjunction fallacy*—usually is not applicable to expert elicitation; it occurs when a higher probability is assigned to an event that is a subset of an event with lower probability (Mullin, 1986) and
(e) *hindsight bias*—if the assessor has seen the sample data, the elicited opinion may have already been updated on the basis of the data (Morgan and Henrion, 1990).

One method that we have found useful in regard to reducing the above is to ask predictive rather than structural questions. Assuming, say, a regression likelihood we can ask questions about the expert's (probabilistic) view of the dependent variable given various values of the predictor variables. This has the advantage that the expert is being asked about only observable quantities and need not understand what a parameter is. The elicitation methods of Kadane *et al.* (1980) commented on by Kadane (1980) and Winkler (1980) are predictive. By contrast, structural methods would ask directly about one's prior distribution on parameters. We find that economists particularly are open to this form of elicitation, since they are quite used to thinking parametrically.

In most of the statistical literature (such as Wilson (1994), Chaloner (1996), Kadane (1980), Murphy and Winkler (1984), Winkler (1986) and Wolpert (1989), among others), some agreement has been reached on how elicitation should be carried out:

(a) expert opinion is the most worthwhile to elicit;
(b) experts should be asked to assess only observable quantities, conditioning only on covariates (which are also observable) or other observable quantities;
(c) experts should not be asked to estimate moments of a distribution (except possibly the first moment); they should be asked to assess quantiles or probabilities of the predictive distribution;
(d) frequent feed-back should be given to the expert during the elicitation process;
(e) experts should be asked to give assessments both unconditionally and conditionally on hypothetical observed data.

Another distinction among methods is that some are designed for a class of problems, whereas others are designed as special purpose elicitation methods that are useful only once. Our attitude is that, although generality is desirable, it is more important that an elicitation method should solve one problem well than that it should address many. Our paper discusses examples of both. We call the former 'general' elicitation methods and the latter 'application specific'.

In Section 2, we discuss the aspects of general predictive elicitation methods by reviewing elicitation methods that are available for the normal linear model, originally developed in Kadane *et al.* (1980) and further in Kadane and Wolfson (1996) and Wolfson (1995). These methods are extended to deal with an AR(1) time series model in Kadane *et al.* (1996), in a way that includes

both predictive and structural components. Elicitations performed using both of these general methods are discussed, and other general elicitation methods in the literature are briefly reviewed.

Section 3 considers the elicitation of prior opinions on vital rates for population projection (Daponte *et al.*, 1997). A brief discussion of some other application-specific elicitations is also given. Section 4 presents a discussion of the relative merits of general *versus* application-specific, and predictive *versus* structural elicitation, along with our recommendations for elicitation in practice.

Although this paper provides a survey of some of the general principles of probability elicitation and discusses some of the advantages and disadvantages of various approaches to elicitation, as well as a discussion of some tools for assessing elicitation methods, the paper is by no means the definitive discussion of elicitation. As Lindley *et al.* (1979) pointed out,

> '... 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.'

2. General elicitation methods

The main feature of an elicitation protocol that earns it the name 'general' is that the method, without any modifications, can be applied to a class of problems.

2.1. Normal linear model

Perhaps the single broadest class of models in statistics is the normal linear model, shown here with a conjugate prior structure:

$$\left.\begin{aligned} Y|X, \beta, \sigma^2 &\sim N(X^{\mathrm{T}}\beta, \sigma^2), \\ \beta|\sigma^2 &\sim N(\mathbf{b}, \sigma^2 R^{-1}), \\ 1/\sigma^2 &\sim \chi^2_\delta / w\delta. \end{aligned}\right\} \tag{1}$$

The hyperparameters of the prior distribution to be elicited are $\mathbf{b}$, R, w and δ. Kadane *et al.* (1980) presented a method for eliciting these hyperparameters, using the predictive approach, which is available (with some modifications by Wolfson (1995) on Statlib as `elicit-normlin.f`). The principle on which the elicitation method is based is that the assessor will be able to make probability judgments about quantities from the predictive distribution, and these judgments can be used to derive values for the hyperparameters. Also, to deal with potential 'elicitation errors', averaging is used to obtain many of the elicited hyperparameters.

The first stage of the elicitation procedure obtains an estimate for $\mathbf{b}$, the prior mean, and δ, the degrees-of-freedom parameter. Before any probabilistic assessments are made, the assessor is asked to specify the range of each of the covariates in the model. For obtaining assessments, realizations of the covariates $\mathbf{x}_t$ are generated by an algorithm that selects these 'design points' optimally. This is done by dividing the range of each covariate into four equal intervals, providing a discrete partition of the span of the design space. The set χ defines the design space, where each element of χ is a unique realization of the vector $\mathbf{x}_t$. Since at any time a proposed design point may be rejected by the assessor, there is another set χ^{R} of the set of rejected $\mathbf{x}_t$. Initially, χ^{R} is the empty set. A point $\mathbf{x}_{t+1}$ is proposed from the set $\chi|\chi^{\mathrm{R}}$ as the next design point at which assessments should be made if $\mathbf{x}_{t+1}$ maximizes the ratio of the largest and smallest eigenvalues of the design matrix $X_{t+1} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \mathbf{x}_{t+1})$.

The assessor is then to specify the 50th, 75th and 90th percentiles of y_i for each vector $\mathbf{x}_i$ generated through this process. If there are p covariates, then the answers from the first stage will be used to assess $p+1$ hyperparameters (the $p \times 1$ vector $\mathbf{b}$, and δ), so assessments of $y_{i_{0.50}}$, $y_{i_{0.75}}$ and $y_{i_{0.90}}$ are required at a minimum of $p+3$ realizations of the covariates $\mathbf{x}_i$, $i = 1, \ldots, m+3$. The expert has the option of assessing up to $p+3$ additional points, which is recommended if the prior distribution assessed is to be used in a model where very few data have been collected. In all the assessments, the assessor is constrained to give 'coherent' answers (i.e. for all i, $y_{i_{0.50}} < y_{i_{0.75}} < y_{i_{0.90}}$).

From these m initial assessments, the hyperparameter $\mathbf{b}$ is assessed as

$$\hat{\mathbf{b}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y_{0.50}.$$

The expert is informed, by looking at the residuals, which, if any, of her assessed medians deviate substantially from the fitted coefficients given by $\hat{\mathbf{b}}$, and then given the opportunity to change her assessments.

To estimate δ, for each design point $\mathbf{x}_i$ at which assessments were collected, take

$$a(\mathbf{x}_i) = \max\left\{\frac{y_{i_{0.90}} - y_{i_{0.50}}}{y_{i_{0.75}} - y_{i_{0.50}}}, \frac{T_\infty(0.90)}{T_\infty(0.75)}\right\}, \tag{2}$$

where $T_\delta(\alpha)$ is the αth quantile of the standard t-distribution with δ degrees of freedom. The posterior predictive distribution of a future observation, Y^{F}, given data Y^{P} (in this case, Y^{P} is the empty set), is a non-standard t-distribution, with centre $X\mathbf{b}$ and spread $w[\{1/(\delta + p)\}XR^{-1}X^{\mathrm{T}} + I]$, and degrees of freedom $\delta + p$. By subtracting the mean in both the numerator and the denominator of the first term on the right-hand side of equation (2), and by the division, this term is made independent of the centre and the spread (so the fact that w and R are still unknown is not a problem). Then the degrees-of-freedom hyperparameter is determined by finding the value of δ that gives the best fit to

$$T_\delta(0.90)/T_\delta(0.75) = \sum_{i=1}^{m} a(\mathbf{x}_i)/m. \tag{3}$$

The expert assessor is offered the chance to view the tail ratios $\mathbf{a}$ from equation (2), and, if necessary, to revise her assessments.

The second stage of the elicitation procedure is aimed at eliciting the hyperparameters w and R, by asking the assessor to specify predictive quantiles conditionally on observing hypothetical data generated by the program. First, the assessor must decide how many of the design points $\mathbf{x}_i$ from the first stage she wishes to use for the second stage. This number, m', is restricted to be between $p+2$ and m. The smaller m' is, the fewer further assessments are required, so one strategy is to make m' larger when assessing a prior for a problem where data are sparse.

The algorithm works as follows for $i = 1, \ldots, m'-1$: first, at design point $\mathbf{x}_i$, a hypothetical data point is generated, denoted by y_i^0. Conditionally on $y_1^0 \ldots y_i^0$, the assessor is asked to re-evaluate her median and 75th percentile for design point $\mathbf{x}_{i+1}$, yielding $y^*_{i+1,0.50}$ and $y^*_{i+1,0.75}$ and then the medians for $\mathbf{x}_i$, $\mathbf{x}_{i+2} \ldots \mathbf{x}_{m'}$. When $i = m'$, a hypothetical data point is generated, but only the medians for $\mathbf{x}_{m'+1}$ and $\mathbf{x}_{m'}$ are re-evaluated conditionally on all the hypothetical data points.

The assessor is shown what her previous assessments were (feed-back), as well as what the best fitted median was, on the basis of her responses from the first stage. The conditional assessments are used sequentially to build a symmetric positive definite matrix U, where the ith diagonal element is the spread of y_i, given by

$$S(y_i|\mathbf{x}_i) = \frac{(y_{t_0\,75} - y_{t_0\,50})^2}{T_\delta^2(0.75)}, \tag{4}$$

and the remaining $i-1$ elements of the ith row (and thus also of the ith column) are the co-spreads of y_i with $(y_1, \ldots, y_{i-1})$. As this matrix is built, the fact that at each stage, conditionally on $y_1^0 \ldots y_i^0$, y_{i+1} and y_{i+1}^* have a joint bivariate t-distribution can be used to obtain w_i, an estimate of w. Thus there will be m' estimates of w, which can be averaged to obtain the final estimate. Finally, the relationship that $U - wI$ is an elicited version of the matrix

$$S(X^{\mathrm{T}}\beta|X) = \frac{w}{n} XR^{-1}X^{\mathrm{T}}, \tag{5}$$

where $n = \delta + m$ can be exploited to obtain the estimate of R^{-1}. Further technical details of the procedure are described in Kadane *et al.* (1980).

This method was used to elicit the prior distribution of a statistician with substantive expertise in the health care field (Kadane and Wolfson, 1996). What follows is a description of the problem, and of how her prior was elicited.

2.1.1. Example 1

On January 1st, 1994, 10 of 41 area programmes in North Carolina began a waiver programme (under the authority of section 1915 (b) of the Social Security Act) for children who are eligible for Medicaid and in need of mental health services. The goal of the waiver programme, called Carolina Alternatives, was to test the effect of capitation funding within an organized managed care system. A grant from the Kate B. Reynolds Health Care Trust was used to fund the development of the programme.

During phase I of the waiver programme, local mental health centres (called area programmes) in charge of managing care are given a capitated (lump sum) amount for in-patient services for all children within their catchment area. The area programmes will be trying to channel children into out-patient rather than in-patient services via community-based diversion and step-down services. Diversion services aim to prevent children from entering hospitals, and step-down services work to speed up the discharge process for those children who do end up in the hospital. The incentive for the area programmes to restrict the use of in-patient services is that they are allowed to keep savings from the capitation allotment and to use these savings to develop community-based alternatives. In areas not participating in the Carolina Alternatives waiver programme, some patients may receive out-patient rather than in-patient services owing to the availability of out-patient care facilities.

One aspect of the evaluation of this programme is comparing the length of hospital stay before the implementation of the programme with the length of stay during the programme. So the response variable is CHANGE (i.e. an increase or decrease in the number of days) in the length of a hospital stay for children who are eligible for Medicaid and in need of mental health care, and the independent variables are CA (Carolina Alternatives), an indicator variable for whether the child was in an area that was participating in the waiver programme, PCCA (*per capita* community-based alternatives), a variable ranging from 0 to 1 that measures community-based alternative capacity within the child's area programme, and an interaction between the two. PCCA is important because, if the community services are not available, the effect of the programme on the length of hospitalization is likely to be minimal.

The first stage of the elicitation procedure yielded the elicited values given in Table 1. The hyperparameters $\mathbf{b}$ and δ are estimated from these initial assessments. The assessor is shown the

Table 1. Initial elicitation results

Point	Independent variables		Elicited predictive percentiles		
	CA	PCCA	50th	75th	90th
1	0.0	0.0	0.0	2.1	6.0
2	1.0	0.0	0.0	3.5	10.0
3	0.0	0.5	1.0	3.1	7.0
4	1.0	0.5	5.0	8.5	15.0
5	0.0	1.0	2.0	4.1	8.0
6	1.0	1.0	10.0	13.5	20.0

residuals from fitting **b** and given the opportunity to update any previous assessments. A segment of the elicitation script demonstrating this is shown below.

```
Variable                Fitted Value
intercept                  -2.000
CA                          4.000
PCCA                        6.000

Here are the residuals from the least squares fit. Those residuals that are
large in absolute value are marked with a '1' in the flag column, because they
may be out of line. Please consider if you want to change any of the assessed
medians in light of these residuals.

          CA       PCCA       Residual       Flag
1          .0         .0          2.0        1.0
2         1.0         .0         -2.0        1.0
3          .0         .5           .0         .0
4         1.0         .5           .0         .0
5          .0        1.0         -2.0        1.0
6         1.0        1.0          2.0        1.0
```

In a similar fashion, the assessor is given the chance to update her assessments on the basis of the residuals from fitting δ. In the second stage of the elicitation, the conditional assessments (indicated in bold) in Table 2 were made.

Table 2. Elicited conditional assessments

Point	CA	PCCA	Initial percentiles		New observation	Conditional assessments					Percentile
						1	2	3	4	5	
			50th	75th							
1	0.0	0.0	0.0	2.1	−0.88	**−0.75**	**−0.50**	**0.25**	**4.50**		50th
							3.00				75th
2	1.0	0.0	0.0	3.5	1.37		**2.5**	**1.5**	**6.0**		50th
								3.6			75th
3	0.0	0.5	1.0	3.1	2.62			**3.5**	**5.5**		50th
									9.0		75th
4	1.0	0.5	5.0	8.5	7.37				**6.5**	**3.5**	50th
5	0.0	1.0	2.0	4.1							

The estimates obtained for the hyperparameters of the prior distribution are

$$\mathbf{b} = \begin{pmatrix} -2.0 \\ 4.0 \\ 6.0 \end{pmatrix}, \qquad R^{-1} = \begin{pmatrix} 17.1 & -20.0 & -16.0 \\ -20.0 & 120.7 & 67.0 \\ -16.0 & 67.0 & 157.6 \end{pmatrix},$$

$$w = 1.11,$$

$$\delta = 1.2.$$

Fig. 1 shows the elicited predictive distribution for various scenarios: children residing in area programmes community-based alternative services (PCCA = 0, 0.25, 0.50 or 0.75).

This predictive distribution is a t-distribution, with degrees of freedom given by δ, centre given by $\mathbf{b}$ and spread given by $w[\{1/(3+\delta)\}xR^{-1}x^{\mathrm{T}} + I]$, where x represents the values taken by the independent variables.

Fig. 1 shows that the predictive distribution is always sharper when the child is not in the area programmes that do not participate in the waiver programme (CA = 1). This is consistent

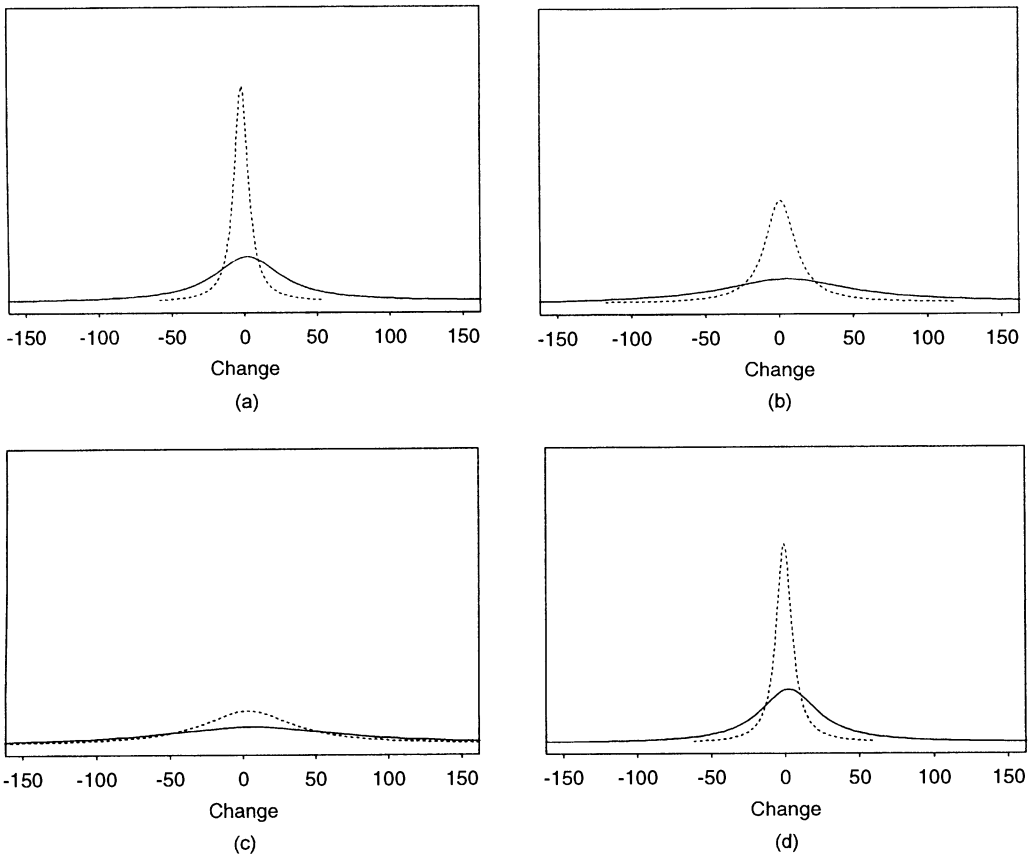

Fig. 1. Prior predictive distribution for change: (a) CA = 1.0, PCCA = 0.0 (———) and CA = 0.0, PCCA = 0.0 (·····); (b) CA = 1.0, PCCA = 0.5 (———) and CA = 0.0, PCCA = 0.5 (·····); (c) CA = 1.0, PCCA = 0.75 (———) and CA = 1.0, PCCA = 0.25 (·····); (d) CA = 0.0, PCCA = 0.75 (———) and CA = 0.0, PCCA = 0.25 (·····)

with what is expected, since the expert has experience with what happens without the waiver programme.

2.2. AR(1) time series models

Consider the general AR(1) time series model

$$y_t = \rho y_{t-1} + \boldsymbol{\beta}' \mathbf{x}_t + \epsilon_t \qquad (y_0 = 0), \tag{6}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, $\mathbf{x}_t = (x_{1t}, \ldots, x_{kt})'$ and ϵ_t is independently and identically distributed $N(0, \sigma^2)$. In studying macroeconomic phenomena, much empirical research is aimed at distinguishing the case when $\rho = 1$ from the case where ρ is near 1. This is known as the 'unit root' problem (Nelson and Plosser, 1982). There is considerable debate about what prior is appropriate to use. In Kadane *et al.* (1996), the methods for eliciting the prior for the normal linear model are extended to include this special case, with modifications made to account explicitly for the uncertainty that comes with the unit root problem.

The method is a combination of structural and predictive elicitation. Because a single prior distribution over the range of ρ may not be sufficiently flexible to accommodate an appropriate range of views, the prior is a piecewise version of the conjugate prior for the normal linear model, with three components to represent $\rho < 1$, $\rho > 1$ and $\rho = 1$. To illustrate, consider the marginal prior distribution of ρ:

$$f(\rho|\sigma_0^2, \sigma_1^2) = \frac{\omega_0}{\sigma_0} \frac{\varphi\left(\dfrac{\rho - \rho_0}{\sigma_0}\right) I(\rho < 1)}{\Phi\left(\dfrac{1 - \rho_0}{\sigma_0}\right)} + \frac{\omega_1}{\sigma_1} \frac{\varphi\left(\dfrac{\rho - \rho_1}{\sigma_1}\right) I(\rho > 1)}{1 - \Phi\left(\dfrac{1 - \rho_1}{\sigma_1}\right)} + \omega_2 I(\rho = 1), \tag{7}$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and the cumulative distribution function of a standard normal random variable. The hyperparameters ω_0, ω_1 and ω_2 are to be elicited structurally, and then ρ_i, σ_i, $i = 0, 1$, are elicited predictively using the exact method described for the normal linear model.

2.2.1. Example 2

The opinion of a macroeconomist on the projected real domestic gross national product (GNP) for five quarters (i.e. from the second quarter of 1993 to the second quarter of 1994) is given for the model

$$\log(\text{GNP}_t) = \mu + \rho \log(\text{GNP}_{t-1}) + \epsilon_t, \tag{8}$$

where $\epsilon_t \sim N(0, \sigma^2)$, and σ^2 is taken to be stochastic.

The macroeconomist has substantial experience with attempting to forecast GNP, as well as a particular interest in the unit root problem. He also has some training in Bayesian statistics.

The hyperparameters elicited from the expert are given in Table 3.

Fig. 2(a) shows the marginal prior distribution for ρ which is a weighted mixture of two t-distributions, and a point mass at 1. Although most of the prior concentrates in the area where ρ is very close to 1, much of the tail weight lies below $\rho = 1$.

Fig. 2(b) shows what the marginal prior distribution would look like if the distribution for $\rho > 1$ was taken to be the same as the distribution of ρ when ρ is known to be stationary (i.e. $\rho_0 = \rho_1 < 1$, and $\sigma_0 = \sigma_1$). In practice, some experts may be unable to provide different priors for the two ranges of ρ, which is the reason to consider this limiting case. The similarity between Figs

Table 3. Elicited hyperparameters

Given	ω_i	ρ_i	σ_i
$\rho < 1$	0.85	0.98	0.0045
$\rho > 1$	0.05	0.9976	0.0082
$\rho = 1$	0.10	1.0	—

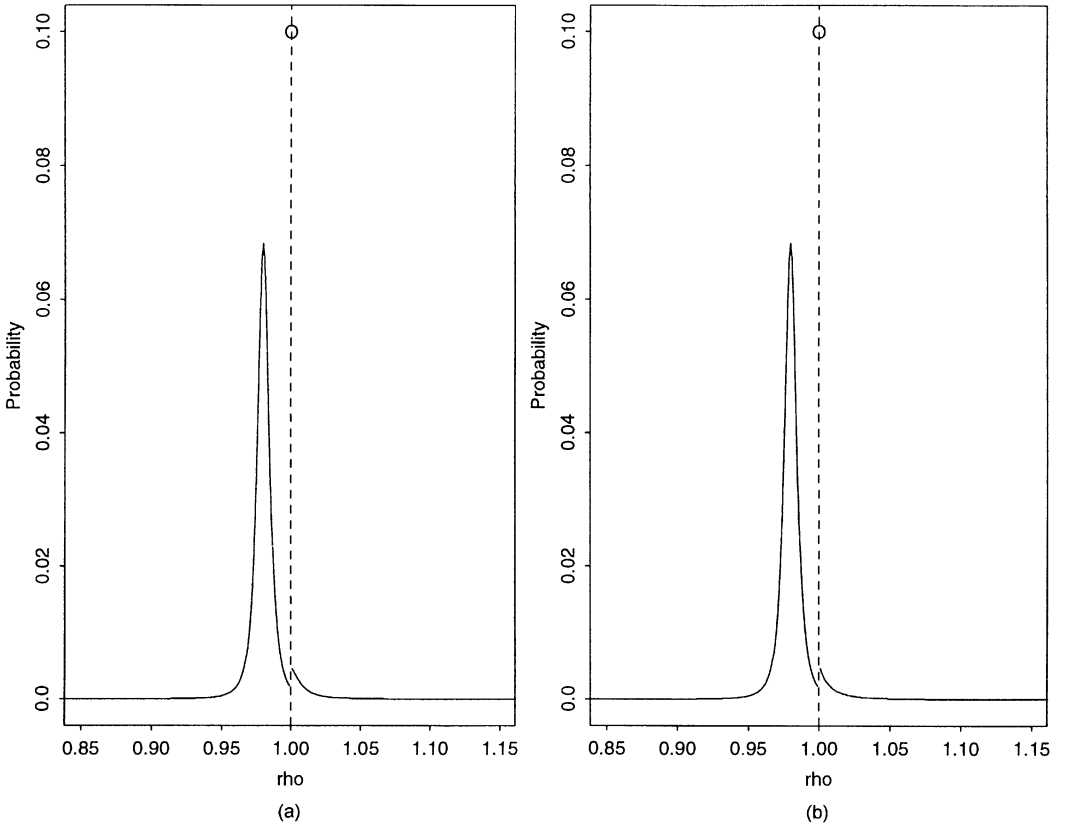

Fig. 2. Prior distributions for ρ

2(a) and 2(b) shows that, in this case, although the two distributions specified were different in Fig. 2(a) and not in Fig. 2(b), very little difference can be seen in the resulting prior distributions. This apparent anomaly is because the prior means in both cases are less than 1, with small variances. The difference in the prior means, however, plays a significant role in the posterior results.

In the unit root problem, where there is concern about whether the root is stationary, a random walk or explosive (corresponding to $\rho < 1$, $\rho = 1$ and $\rho > 1$), it can be useful to calculate the odds that $\rho = 1$. In both cases described above, the prior odds are given by $\omega_2/(\omega_0 + \omega_1) = 0.111$.

Data on real domestic GNP from the third quarter of 1993 to the third quarter of 1994 were

obtained from the Statistics in Education database server. The posterior odds of a unit root may be calculated as

$$\text{odds} = \frac{P(\rho = 1|\text{data})}{P(\rho < 1|\text{data}) + P(\rho > 1|\text{data})}. \tag{9}$$

The posterior odds of a unit root when the distributions for $\rho < 1$ and $\rho > 1$ are different are 2.098; when the distributions are the same, the posterior odds of a unit root are 0.489. The posterior distributions of ρ for these two cases are shown in Figs 3(a) and 3(b). The likelihood has been fairly informative; the majority of the weight has been shifted away from the stationary root into the unit and explosive roots.

2.3. Other methods

In addition to the method of Kadane *et al.* (1980) for eliciting a prior for the normal linear model, Garthwaite and Dickey (1988, 1992) have proposed an alternative method. They focused their attentions on modifications to allow for the *a priori* selection of covariates. A comparison of the methods is given in detail in Wolfson (1995). Both methods use the predictive approach, however, and to the end-user, aside from a few differences in how the questions are asked, the distinction lies mainly in how the hyperparameters are estimated. Kadane *et al.* (1980) used averaging to diminish the effect of elicitation errors, but Garthwaite and Dickey (1988, 1992) could circumvent a problem in estimating the hyperparameter w that is found in Kadane *et al.* (1980). In general, if the object of eliciting the prior is to provide information about covariate selection, then the methods of Garthwaite and Dickey are preferable. If, however, there are indicator variables in the covariate set, or there are interaction terms or polynomial regression coefficients, then the methods of Kadane *et al.* (1980) should be used. Garthwaite (1994) has reported an experimental study

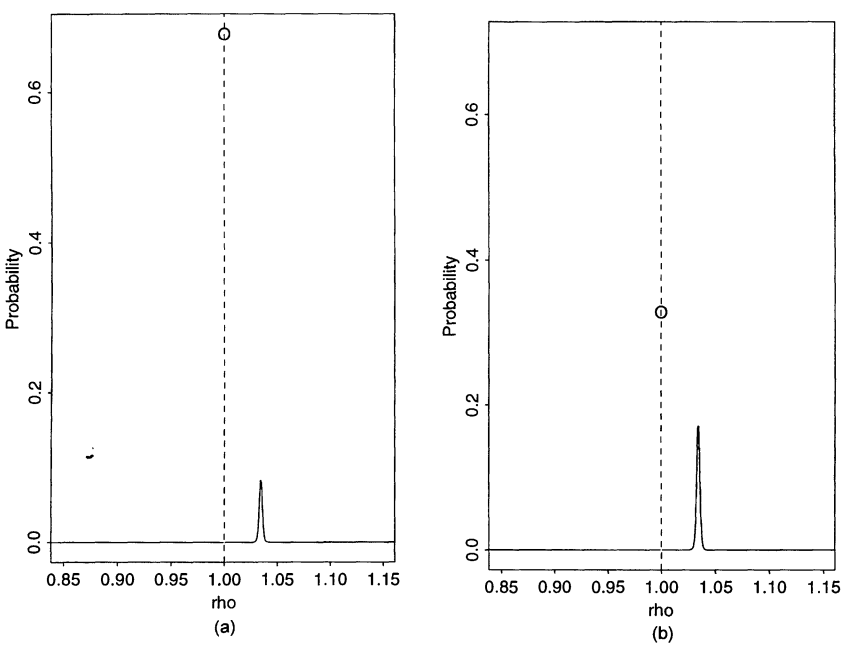

Fig. 3. Posterior distributions for ρ

with some statistical students using a modification of the elicitation methods of Kadane *et al.* (1980).

Other general methods for elicitation are found in Chaloner and Duncan (1983) and Gavasakar (1988), where methods are described for eliciting the prior for the beta–binomial model, and Chaloner *et al.* (1993) have proposed a graphical elicitation method for the Cox proportional hazards model. Some other advances in the general methods have been made by Black and Laskey (1989) and Laskey and Black (1989) for Bayesian analysis of variance, for the Dirichlet–multinomial distribution (Chaloner and Duncan, 1987), multivariate t- and matrix t-models (Dickey *et al.*, 1986), by Gokhale and Press (1982) for the distribution of the correlation coefficient in a bivariate normal distribution and by Singpurwalla and Song (1987), who proposed a method for eliciting information about the mean and shape parameter of a Weibull distribution, including the analyst's opinion about the expertise of the expert.

3. Application-specific elicitation methods

To illustrate the nature of application-specific methods, we report here the experience of Daponte *et al.* (1997) in eliciting the opinions of a demographer on vital rates for the Iraqi Kurdish population during the period 1977–90. The elicitation is a combination of the demographer's expert opinion and information found in the literature on the Iraqi Kurds. Because of the lack of reliable census data from Iraq after 1977, the goal of the analysis of the Iraqi Kurdish population is to examine what this population would have been before the Gulf war if the Anfal, a programme of state-sanctioned violence against the Kurds, had not taken place. The part of the project that we report here is how to elicit the information that is necessary to project the population during that period; to do so, the fertility rates and mortality rates during that period are needed. A combination of structural and predictive questions is used to elicit the opinion in a way that is specific to the problem of demography.

3.1. Example 3

3.1.1. Fertility rates

Fertility rates are needed separately by region (rural and urban) for the 'base' (1977) and 'end' (1990) years of the projection period, by asking the assessor to specify the mean of her distribution and the 2.5% and 97.5% quantile.

To specify the mean of the distributions, the assessor used her knowledge of the current literature available from the United Nations to find the mean total fertility rate (TFR) for 1974 and 1988. Using the TFR at two time points, B and E (respectively tfr_B and tfr_E) a logistic interpolation (Arriaga and Associates (1993), appendix 8-4) with standard demographic asymptotes of la = 2 (lower) and ua = 8 (upper) is employed, where

$$x_1 = \log\left(\frac{\text{ua} - \text{tfr}_B}{\text{tfr}_B - \text{la}}\right), \tag{10}$$

$$x_2 = \log\left(\frac{\text{ua} - \text{tfr}_E}{\text{tfr}_E - \text{la}}\right), \tag{11}$$

$$w = \frac{x_1 - x_2}{B - E}, \tag{12}$$

$$a = x_1 - wB. \tag{13}$$

Then the TFR at a time T is obtained as

$$\text{tfr}_T = \frac{\text{ua} + \text{la} + \exp(a + wT)}{1 + \exp(a + wT)}. \tag{14}$$

This yielded overall TFRs of 6.96 in 1977 and 5.97 in 1990. The same literature suggested that overall rural fertility during that time period should be 20% higher than urban fertility, and the elicited means for each of the four required TFRs are obtained by factoring in this difference.

Whereas a traditional demographic technique was used to assess the mean TFRs, the variance of the TFRs was assessed from a more subjective viewpoint, by asking the assessor to provide her 2.5% and 97.5% quantile for each of the four TFRs. In this case, perceived sources of uncertainty include initial mismeasurement of the TFRs, the appropriateness of using Iraqi TFRs for the Kurdish population, uncertainty in whether the rural–urban projected difference is correctly estimated and uncertainty in how well the projection of the TFR by using a logistic function reflects fertility in Iraq during the projection period.

The elicited TFRs are shown in Fig. 4, along with the elicited means and standard deviations. Correlations were elicited structurally, by asking the expert to assess them directly, with ρ(rural, urban) = 1 and ρ(1977, 1990) = 0. Although these numbers seem somewhat simplistic, the expert judged that they best reflected the situation as she knew it.

3.1.2. Infant mortality rates

It is assumed that infant mortality rates follow a joint normal distribution, so the joint infant mortality rates (i.e. for both sexes, and rural and urban areas combined) were elicited by asking for the mean and the 2.5th and 97.5th percentile of the distribution for both the base and the end years, yielding

$$\begin{aligned} \text{JMR}_{1977} &\sim N(70.4,\ 6.0^2), \\ \text{JMR}_{1990} &\sim N(39.0,\ 4.0^2). \end{aligned} \tag{15}$$

To model the sex differential in infant mortality, defined as $\Delta_{\text{MF}} = \text{IMR}_{\text{Female}} - \text{IMR}_{\text{Male}}$, the assessor was asked to provide the 50th, 75th and 97.5th percentiles of her distribution of the differential, conditional on JMR (Table 4). The values indicate that both the mean and the standard deviation of the resulting normal distribution are linear in JMR, so $\Delta_{\text{MF}}|\text{JMR} \sim N\{0.1\,\text{JMR} - 0.9,\ (0.059\,\text{JMR})^2\}$, for both 1977 and 1990.

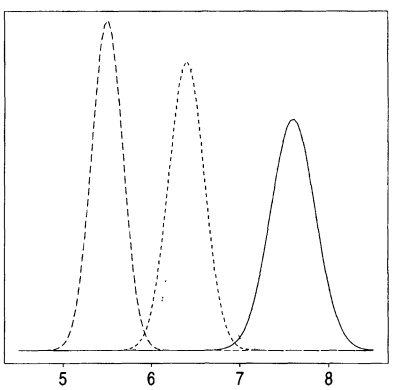

Fig. 4. Elicited TFRs: – – – –, urban, 1990, $\mu = 5.5$, $\sigma = 0.175$; - - - - - - -, urban, 1977, $\mu = 6.4$, $\sigma = 0.2$; · · · · · · ·, rural, 1990, $\mu = 6.5$, $\sigma = 0.2$; ———, rural, 1977, $\mu = 7.6$, $\sigma = 0.25$

Table 4. Elicited quantiles of sex differential given infant mortality rates

JMR	*Quantiles*		
	50th	*75th*	*97.5th*
90	0	4	12
70	−2	1	7
50	−4	−2	2
30	−4	−2	2

Finally, the distribution of the rural–urban differential Δ_{RU} was obtained by asking the assessor to give values for the mean and variance of the appropriate normal distribution; her response yielded $\Delta_{RU} \sim N(0.1, 0.05^2)$.

These distributions are combined with the following quantities (assumed to be fixed throughout the projection period) to determine infant mortality rates IMR for each sex, within each region: expected sex ratio at birth $ESR_0 = 1.05$; portion of population living in rural areas, $r = 0.51$; portion of population living in urban areas, $u = 0.49$.

So, with $r + u = 1$,

$$IMR_F = JMR - \left(\frac{ESR_0}{1 + ESR_0}\right)\Delta_{MF}, \tag{16}$$

$$IMR_M = JMR + \frac{1}{1 + ESR_0}\Delta_{MF}, \tag{17}$$

$$IMR_{J,RURAL} = \frac{1 + \Delta_{RU}}{1 + r\Delta_{RU}} IMR_J, \tag{18}$$

$$IMR_{J,URBAN} = \frac{1}{1 + r\Delta_{RU}} IMR_J, \qquad J \equiv F, M. \tag{19}$$

3.2. Other application-specific methods

Cooke (1991), Morgan and Henrion (1990) and Meyer and Booker (1991) have extensive discussions of elicitation, and each has a number of examples of application-specific elicitations. DuMouchel (1988) conducted an elicitation for multiple comparisons of parameters measured on the same scale in the context of a normal linear model. In Kadane and Schum (1992, 1996), prior probabilities with regard to legal evidence are elicited. In Berry *et al.* (1992), priors are elicited for use in the design of a sequential vaccine trial. Kadane and Hastorf (1988) elicited prior distributions for a problem in paleoethnobotany. Flournoy (1994) elicited prior distributions to be used in the design of a clinical trial from physicians by having them sketch upper and lower response curves; Freedman and Spiegelhalter (1983) elicited probability distributions from physicians on the magnitude of the effect (in terms of the probability of recurrence) of a drug used after surgery for bladder cancer; Freedman and Spiegelhalter (1992), Spiegelhalter and Freedman (1986, 1988), Spiegelhalter *et al.* (1993, 1994) and Kadane (1994, 1996) are other examples of papers where elicitation is used in clinical trials. Goldman *et al.* (1988) developed a computer program to use expert opinion to predict myocardial infarction. Kirnbauer *et al.* (1987) used elicited opinions combined with historical information as a prior distribution for predicting floods.

Elicitation in meteorology is described in Murphy and Winkler (1977, 1984) and Winkler and Murphy (1968). Raftery *et al.* (1995) elicited prior information for the population dynamics of bowhead whales.

4. Discussion

Prior elicitation is the process of using expert opinion to construct prior distributions. The purpose is to yield a useful prior that captures the main features of the expert's opinion, integrating her experience and her knowledge of the literature. In this paper, we have discussed both general elicitation methods and application-specific methods. The choice of whether to use either the former or the latter relies on the statistician's determination of what a 'useful' prior would be.

For example, we return to the elicitation model given in Section 2.2.1 where the prior distribution, although closed under sampling, is not the usual conjugate prior family. The piecewise nature of the prior, however, provides useful information that the usual conjugate prior family would not have.

Many statistical models, such as the normal linear model, are common to more than one field of application. For such models, general elicitation methods are very desirable because of their portability, and because they are relatively simple to use, once they have been implemented. The elicitation program for the normal linear model that we have made available on Statlib does not require the user to have much knowledge about elicitation and thus allows many practitioners using the linear model to elicit a prior. For complicated models that are specific to the application, such as example 3 in this paper, the elicitation method is particular to the problem at hand, and, although the method can be used within demography, it does not generalize well to other applications. Additionally, it requires a greater understanding of the underlying structure of the model and the principles of elicitation.

We believe that both predictive and structural elicitations are useful in different contexts. We have provided examples of predictive elicitation (example 1), structural elicitation (example 2) and a hybrid of both methods (example 3) in this paper. Which method should be used is determined by examining the nature of the problem and whether or not the parameters have intrinsic meaning to the expert. In the first example, the expert does not usually consider the correlation between regression coefficients, so the predictive method of elicitation is particularly apt since it allows an indirect assessment of the correlations. When the expert is accustomed to thinking directly about the parameters of the distribution, as was the case in the third example (and in assessing the probability of ρ being in particular regions in the second example), it is more expedient to assess these structural components directly.

Two key issues that often arise in choosing an elicitation method are how to choose between methods and how to validate an elicitation. Three components might be used to 'validate' an elicitation: reliability (Wallsten and Budescu, 1983), coherence (Lindley *et al.*, 1979) and calibration (Morgan and Henrion, 1990; Hogarth, 1975). Since elicitation purports to represent the opinions of the person being elicited, and only that, any attempts at validation should use only other statements of belief by that person as data. The best that can be done, then, is checking the way that different answers of the person go together (i.e. coherence) and overall statements of satisfaction by the person whose opinion is being elicited (thus giving reliability). Averaging over several elicited responses and designing elicitation methods such that non-mathematically coherent responses will be rejected have been used in many of the generally available elicitation methods, including those described in this paper.

Calibration, though, remains the trickiest issue; in the sometimes misguided quest for objectivity, researchers often wish for the elicited prior to be well calibrated in the sense that $p\%$

of all predictions reported at probability p are true. There are two arguments against calibration, one mathematical and one philosophical. The mathematical argument (attributed to Kadane and Lichtenstein (1982) as described in Seidenfeld (1985)) is that there is no function $F(p)$ other than the identity function such that both the pre-calibration and post-calibration probabilities are coherent. The philosophical argument, and in our opinion the more compelling, is that what is being elicited is *expert*, not perfect, opinions, and thus they should not be adjusted. Some comfort for those who are ill at ease with such subjectivity can be found in Dawid (1982), who argued that an expert who is a 'coherent Bayesian' expects to be well calibrated. Other papers that address the issue of calibration are DeGroot and Fienberg (1982), French (1986), Kadane and Lichtenstein (1982), Murphy and Winkler (1977) and Vardeman and Meeden (1983).

As a final note on validation, we direct the reader's attention to the work of Gavasakar (1988), where an alternative method for eliciting priors for the beta–binomial model to that given by Chaloner and Duncan (1983) is proposed. Gavasakar (1988) also introduced a strategy for comparing the two methods, based on a hierarchical model component to model elicitation errors. Both elicitation methods were tested by assuming that the prior distribution had a certain form, and then adding random errors to what the answers should have been, given the specified prior. The results from the elicitation procedures could then be used to compare the estimated hyperparameters with the 'true' hyperparameters and give a sense of under what circumstances the performances of the two methods differ. Essentially, using this diagnostic as a tool to choose between elicitation methods is based on the analyst's putting a prior on the nature of the elicitation errors that his expert will make, or, as Dickey (1980) put it, 'beliefs about beliefs'! (Other works that discuss models for errors in elicitation are French (1980), Logan (1985) and Wolfson (1995).)

Beyond these measures of 'goodness' (Winkler, 1986; Winkler and Murphy, 1968) of an elicitation method, the primary criterion in choosing an elicitation method is practicality. If the expert can answer the questions and feels comfortable, in the end, that to some degree her opinion has been captured, then provided that the method meets the basic mathematical criteria of coherence, and hopefully involves some reliability testing, it is a good method.

The experience of participating in an elicitation protocol often leads the expert to think more carefully about the underlying structure of the problem in a more probabilistic way. Blanck *et al.* (1996) report on their experiences with elicitation. We think that expert elicitation, as a form of knowledge engineering, has an important future for statisticians, and for practitioners in other fields. We think that statistical analysis should be based on full information (Barnett, 1982), which consists not only of data but also of prior information and utilities.

References

Alpert, M. and Raiffa, H. (1982) A progress report on the training of probability assessors. In *Judgment under Uncertainty: Heuristics and Biases* (eds D. Kahneman, P. Slovic and A. Tversky), pp. 294–305. Cambridge: Cambridge University Press.

Arriaga and Associates (1993) *Population Analysis with Microcomputers*, vol. 1, *Presentation of Techniques*. Washington DC: US Bureau of the Census.

Barnett, V. (1982) *Comparative Statistical Inference*. New York: Wiley.

Berry, D. A., Wolff, M. C. and Sack, D. (1992) Public health decision making: a sequential vaccine trial. In *Bayesian Statistics* 4 (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 79–96. Oxford: Oxford University Press.

Black, P. and Laskey, K. (1989) Models for elicitation in Bayesian ANOVA: implementation and application. *Proc. Statist. Comput. Sect. Am. Statist. Ass.*, 247–252.

Blanck, T. J. J., Conahan, T. J., Merin, R. G., Prager, R. L. and Richter, J. J. (1996) Being an expert. In *Bayesian Methods and Ethics in a Clinical Trial Design* (ed. J. B. Kadane), pp. 159–162. New York: Wiley.

Chaloner, K. (1996) The elicitation of prior distributions. In *Case Studies in Bayesian Biostatistics* (eds D. Berry and D. Stangl), pp. 141–156. New York: Dekker.

Chaloner, K., Church, T., Louis, T. A. and Matts, J. P. (1993) Graphical elicitation of a prior distribution for a clinical trial. *Statistician*, **42**, 341–353.

Chaloner, K. M. and Duncan, G. T. (1983) Assessment of a beta prior distribution: PM elicitation. *Statistician*, **32**, 174–180.

———(1987) Some properties of the Dirichlet-multinomial distribution and its use in prior elicitation. *Communs Statist. Theory Meth.*, **16**, 511–523.

Cooke, R. M. (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York: Oxford University Press.

Daponte, B. O., Kadane, J. B. and Wolfson, L. J. (1997) Bayesian demography: projecting the Iraqi Kurdish population 1977–1990. *J. Am. Statist. Ass.*, to be published.

Dawid, A. P. (1982) The well-calibrated Bayesian (with discussion). *J. Am. Statist. Ass.*, **77**, 605–613.

DeGroot, M. H. and Fienberg, S. E. (1982) Assessing probability assessors: calibration and refinement. In *Statistical Decision Theory and Related Topics III*, vol. 1, pp. 291–314. New York: Academic Press.

Dickey, J. M. (1980) Beliefs about beliefs: a theory of stochastic assessments of subjective probabilities. In *Bayesian Statistics* (eds J. M. Bernardo, M. H. DeGroot and A. F. M. Smith). Valencia: Valencia University Press.

Dickey, J. M., Dawid, A. P. and Kadane, J. B. (1986) Subjective-probability assessment methods for multivariate-t and matrix-t models. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (eds P. Goel and A. Zellner), pp. 177–195. Amsterdam: North-Holland.

DuMouchel, W. (1988) A Bayesian model and a graphical elicitation procedure for multiple comparisons. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 127–145. Oxford: Oxford University Press.

Flournoy, N. (1994) A clinical experiment in bone marrow transplantation: estimating a percentage point of a quantal response curve. In *Case Studies in Bayesian Statistics* (eds C. Gatsonis, J. Hodges, R. Kass and N. Singpurwalla), pp. 324–336. New York: Springer.

Freedman, L. S. and Spiegelhalter, D. J. (1983) The assessment of subjective clinical opinion and its use in relation to stopping rules for clinical trials. *Statistician*, **32**, 153–180.

———(1992) Application of Bayesian statistics to decision making during a clinical trial. *Statist. Med.*, **11**, 23–35.

French, S. (1980) Updating of belief in the light of someone else's opinion. *J. R. Statist. Soc.* A, **143**, 43–48.

———(1986) Calibration and the expert problem. *Mangmnt Sci.*, **32**, 315–321.

Garthwaite, P. (1994) Assessments of prior distributions for regression models: an experimental study. *Communs Statist. Simuln*, **23**, 871–895.

Garthwaite, P. H. and Dickey, J. M. (1988) Quantifying expert opinion in linear regression problems. *J. R. Statist. Soc.* B, **50**, 462–474.

———(1992) Elicitation of prior distributions for variable-selection problems in regression. *Ann. Statist.*, **20**, 1697–1719.

Gavasakar, U. (1988) A comparison of two elicitation methods for a prior distribution for a binomial parameter. *Mangmnt Sci.*, **34**, 784–790.

Gokhale, D. V. and Press, S. J. (1982) Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *J. R. Statist. Soc.* A, **145**, 237–249.

Goldman, L., Cook, E. F., Brand, D. A., Lee, T. H., Rouan, G. W., Weisberg, M. C., Acampora, D., Stasiulewicz, C., Walshon, J., Terranova, G., Gottlieb, L., Kobernick, M., Goldstein-Wayne, B., Cohen, D., Daley, K., Brandt, A. A., Jones, D., Mellors, J. and Jakubowski, R. (1988) A computer protocol to predict myocardial infarction in emergency department patients with chest pain. *New Engl. J. Med.*, **318**, 797–803.

Hogarth, R. M. (1975) Cognitive processes and the assessment of subjective probability distributions (with discussion). *J. Am. Statist. Ass.*, **70**, 271–294.

Kadane, J. B. (1980) Predictive and structural methods for eliciting prior distributions. In *Bayesian Analysis in Econometrics and Statistics* (ed. A. Zellner). Amsterdam: North-Holland.

———(1994) An application of robust Bayesian analysis to a medical experiment. *J. Statist. Planng Inf.*, **40**, 221–232.

———(ed.) (1996) *Bayesian Methods and Ethics in a Clinical Trial Design*. New York: Wiley.

Kadane, J. B., Chan, N. H. and Wolfson, L. J. (1996) Priors for unit root models. *J. Econometr.*, **75**, 99–111.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S. and Peters, S. C. (1980) Interactive elicitation of opinion for a normal linear model. *J. Am. Statist. Ass.*, **75**, 845–854.

Kadane, J. B. and Hastorf, C. A. (1988) Bayesian paleoethnobotany. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 243–259. Oxford: Oxford University Press.

Kadane, J. B. and Lichtenstein, S. (1982) A subjectivist view of calibration. *Technical Report 233*. Department of Statistics, Carnegie Mellon University, Pittsburgh.

Kadane, J. B. and Schum, D. A. (1992) Opinions in dispute: the Sacco-Vanzetti case. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 79–96. Oxford: Oxford University Press.

———(1996) *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. New York: Wiley.

Kadane, J. B. and Wolfson, L. J. (1996) Priors for the design and analysis of clinical trials. In *Case Studies in Bayesian Biostatistics* (eds D. Berry and D. Stangl), pp. 157–184. New York: Dekker.

Kirnbauer, R., Schnatter, S. and Gutknecht, D. (1987) Bayesian estimation of design floods under regional and subjective prior information. In *Probability and Bayesian Statistics* (ed. R. Viertl), pp. 285–294. New York: Plenum.

Laskey, K. B. and Black, P. (1989) Models for elicitation in Bayesian analysis of variance. In *Computer Science and*

Statistics: Proc. 21st Symp. Interface, pp. 242–247. Alexandria: American Statistical Association.

Lindley, D. V., Tversky, A. and Brown, R. V. (1979) On the reconciliation of probability assessments (with discussion). *J. R. Statist. Soc.* A, **142**, 146–180.

Logan, D. M. (1985) The value of probability assessment. *PhD Thesis*. Department of Engineering-Economic Systems, Stanford University, Stanford.

Meyer, M. and Booker, J. (1991) *Knowledge-based Expert Systems*, vol. 5, *Eliciting and Analyzing Expert Judgment: a Practical Guide*. New York: Academic Press.

Morgan, M. G. and Henrion, M. (1990) *Uncertainty: a Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. New York: Cambridge University Press.

Mullin, T. M. (1986) Understanding and supporting the process of probabilistic estimation. *PhD Thesis*. Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh.

Murphy, A. H. and Winkler, R. L. (1977) Reliability of subjective probability forecasts of precipitation and temperature. *Appl. Statist.*, **26**, 41–47.

———(1984) Probability forecasting in meteorology. *J. Am. Statist. Ass.*, **79**, 489–500.

Nelson, C. R. and Plosser, C. I. (1982) Trends and random walks in macroeconomic time series. *J. Monet. Econ.*, **10**, 139–162.

Raftery, A. E., Givens, G. H. and Zeh, J. E. (1995) Inference from a deterministic population dynamics model for bowhead whales (with discussion). *J. Am. Statist. Ass.*, **90**, 402–430.

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

Seidenfeld, T. (1985) Calibration, coherence and scoring rules. *Phil. Sci.*, **52**, 274–294.

Singpurwalla, N. D. and Song, M. S. (1987) The analysis of weibull lifetime data incorporating expert opinion. In *Probability and Bayesian Statistics* (ed. R. Viertl), pp. 431–442. New York: Plenum.

Spiegelhalter, D. J. and Freedman, L. S. (1986) A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statist. Med.*, **5**, 1–13.

———(1988) Bayesian approaches to clinical trials. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 453–477. Oxford: Oxford University Press.

Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. (1993) Applying Bayesian ideas in drug development and clinical trials. *Statist. Med.*, **12**, 1501–1511.

———(1994) Bayesian approaches to randomized trials (with discussion). *J. R. Statist. Soc.* A, **157**, 357–416.

Tversky, A. and Kahneman, D. (1974) Judgment under uncertainty: heuristics and biases. *Science*, **185**, 1124–1131.

Vardeman, S. and Meeden, G. (1983) Calibration, sufficiency and domination considerations for Bayesian probability assessors. *J. Am. Statist. Ass.*, **78**, 808–816.

Wallsten, T. S. and Budescu, D. V. (1983) Encoding subjective probabilities: a psychological and psychometric review. *Mangmnt Sci.*, **29**, 151–173.

Wilson, A. G. (1994) Cognitive factors affecting subjective probability assessment. *Discussion Paper 94-02*. Institute of Statistics and Decision Sciences, Duke University, Durham.

Winkler, R. L. (1967a) The assessment of prior distributions in Bayesian analysis. *J. Am. Statist. Ass.*, **62**, 776–800.

———(1967b) The quantification of judgment: some methodological suggestions. *J. Am. Statist. Ass.*, **62**, 1105–1120.

———(1980) Prior information, predictive distributions, and Bayesian model-building. In *Bayesian Analysis in Econometrics and Statistics* (ed. A. Zellner). Amsterdam: North-Holland.

———(1986) On "good probability appraisers". In *Bayesian Inference and Decision Techniques* (eds P. Goel and A. Zellner), pp. 265–278. Amsterdam: Elsevier Science.

Winkler, R. L. and Murphy, A. H. (1968) "Good" probability appraisers. *J. Appl. Meteorol.*, **7**, 751–758.

Wolfson, L. J. (1995) Elicitation of priors and utilities for Bayesian analysis. *PhD Thesis*. Department of Statistics, Carnegie Mellon University, Pittsburgh.

Wolpert, R. L. (1989) Eliciting and combining subjective judgments about uncertainty. *Int. J. Technol. Assmnt Hlth Care*, **5**, 537–557.

Wright, G. and Ayton, P. (1994) *Subjective Probability*. New York: Wiley.