Automatica, Vol. 24, No. 1, pp. 87–94, 1988
Printed in Great Britain.

0005–1098/88 $3.00 + 0.00
Pergamon Journals Ltd.

Brief Paper

Calibration and Information in Expert Resolution; a Classical Approach*

ROGER COOKE†, MAX MENDEL‡ and WIM THIJS§

Key Words—Expert resolution; expert opinion; subjective probability; calibration.

Abstract—A classical approach to expert resolution is presented using the concepts of calibration and information. Methodological problems with calibration measurements are brought to light and solutions are proposed. An experiment is described in which this approach is shown to have descriptive value.

Introduction

INTEREST in expert resolution is motivated by the increasing use of subjective probabilities in scientific studies, particularly in quantitative risk assessment. The principles of expert resolution are also applicable in situations where probabilistic diagnostic systems must be evaluated as well as in training and selection programs for personnel who may be called upon to give expert probability assessment (Mendel and Thijs, 1983).

The first author to address expert resolution as such was Roberts (1965). Important contributions can be found in Harrison (1977), Morris (1974, 1977), Lindley (1982), De Groot and Fienberg (1983) and in Lichtenstein and Fischhoff (1977) and Lichtenstein *et al.* (1982). Recent contributions reflecting increasing interest among statisticians can be found in Genest and Schervish (1985) and Agnew (1985). De Groot and Fienberg (1986) and Winkler (1986) take up the problem of evaluating probabilistic forecasters, and Kempthorne and Mendel (1987) discuss Bayesian calibration of forecasters. Cooke (1987) provides a theory of combining expert opinions based on the notions developed in the present article.

The theory of expert resolution is concerned with the development of criteria for evaluating and utilizing expert probability assessments. The sources mentioned above approach this problem from a Bayesian perspective, and analyse the way in which a decision maker should process expert probability assessments. The models proposed all require the decision maker to make two types of probability assessments. First he must make a prior assessment of the decision variable of interest, and second he must assess the likelihoods of the expert's responses conditional on the values of the decision variable. Morris (1977) shows how the decision maker can use calibration data to correct for expert bias. As pointed out in Agnew (1985) and Genest and Schervish (1985), these assessment tasks are rather forbidding. Kempthorne and Mendel (1987) draw attention to other problems in Morris' theory. On the other hand, the Bayesian approach enables the decision maker to calculate the precise value of an expert for a particular decision problem in terms of increased expected value.

De Groot and Fienberg (1986) and Winkler (1986) propose using proper scoring rules for evaluating probabilistic forecasters. Their approach is somewhat similar to the ideas presented here, though Cooke (1987) points out several significant differences.

In this article we approach the problem of expert resolution from a classical perspective. An expert probability assessment is treated as a statistical hypothesis in the sense of "objectivist" statistics, and we show how experts can be evaluated from this perspective.

Morgan *et al.* (1979, p. 12) discuss four criteria for evaluating probability assessments (these criteria are attributed to Sarah Lichtenstein). Assessments should be:

—*consistent,* they should not vary with the assessment method, nor over time (assuming the assessor gets no new information),
—*coherent,* they should obey the laws of probability (e.g. Bayes' rule),
—*informative,* they should contain information about actual outcome values of the quantities assessed,
—*well-calibrated,* in the long run assessed probabilities should approximate empirical frequencies of outcomes.

In this article we are concerned with the last two criteria, informativeness and calibration. We introduce two scores for measuring information and calibration of expert probability assessments; a calibration score and an information score. A good expert should be good with respect to both scores.

An experiment is discussed in which the performance of experienced and inexperienced probabilistic assessors can be compared. It emerges that the group performance of experienced assessors is significantly better with respect to both calibration and information than that of inexperienced assessors, on items relating to their field of expertise. At the same time, there is a negative correlation between the individual scores for calibration and information. On general knowledge items the groups are not distinguishable. This suggests that experienced assessors *as a group* are indeed "better experts" than the inexperienced assessors in their field of expertise.

Since the scores used here do not depend on any prior distributions, or on any specific decision problem, the notion of "goodness" is not restricted to a given problem or to a given decision maker. We claim that these scores provide objective criteria for evaluating expert probability assessors. Such criteria can be of practical importance in training and selection, and also in decision situations where the relevant prior distributions are not available (for example, the decision maker may be a group of individuals unable to agree on priors). Naturally, this approach suffers from the same weaknesses as any classical approach; for instance, the

* Received 25 June 1986; revised 12 February 1987; revised 19 June 1987. The original version of this paper was presented at the 2nd IFAC/IFIP/IFORS/IEA Conference on Analysis, Design and Evaluation of Man–Machine Systems which was held in Varese, Italy during September 1985. The Published Proceedings of this IFAC Meeting may be ordered from: Pergamon Books Limited, Headington Hill Hall, Oxford OX3 OBW, England. This paper was recommended for publication in revised form by Editor A. P. Sage.

† Department of Mathematics, Delft University of Technology, The Netherlands.
‡ Department of Mechanical Engineering, Massachusetts Institute of Technology, U.S.A.
§ Department of Mechanical Engineering, Delft University of Technology, The Netherlands.

evaluation of an expert is not nearly as precise as in the Bayesian analysis.

The first section of this paper presents the basic concepts of expert resolution. The second section considers implementing this theory in psychometric measurement procedures. The third section analyses the experiment, and the final section presents discussion and conclusions.

1. *Theoretical considerations*

Suppose that we are required to estimate the mean time to failure in days of a new system component which cannot be subjected to destructive experimental tests. Our only way of obtaining quantitative data is to ask the opinion of experts acquainted with similar kinds of components. Given the amount of uncertainty inherent in predictions of this sort, the experts may feel uncomfortable about giving point predictions, and may prefer to communicate something about the range of their uncertainty. The best they could do in this respect would be to give their subjective probability mass functions (or density functions in the case of continuous variables) for the quantity in question. In other words, they could provide a histogram over the positive integers such that the mass above the integer i is proportional to their subjective probability that the mean time to failure is i days.

The mean time to failure will eventually become known, and when it is known, we may want to pose the question how good was this expert's assessment. We shall discuss two general features of subjective probability mass functions which are relevant to performing this kind of evaluation, namely, *information* and *calibration*. We assume that subjective probability mass functions over the integers can be solicited from each of several experts, for a large number of uncertain quantities.

Information. The information associated with a probability mass function P over the integers is defined as

$$I(P) = \sum_{i=0}^{\infty} P(i) \log (P(i)) \qquad (1)$$

where $P(i)$ is the probability assigned to the integer i. $I(P)$ attains its maximum value (zero) if $P(i) = 1$ for some i. Roughly speaking, smaller values of $I(P)$ correspond to a "flatter" mass function. For mass functions concentrated on a finte number N of integers, $I(P)$ attains its minimum value $-\log(N)$ when P is uniform. Notice that the range of the information function depends on the set of possible values on which the distribution is concentrated.

For functions having a small number of "peaks" (as is usually the case for subjective probabilities) $I(P)$ is a good measure of the degree to which the density is "spread out". Unlike the variance, $I(P)$ is not sensitive to the mass in the "tails" of the function P.

Obviously, high information is a desideratum in expert probabilistic assessment. Other things being equal, we should prefer the advice of the expert whose probability functions have the highest information. We shall see, however, that other things are usually not equal.

Suppose we have a set of experts and a set of uncertain quantities, and suppose we solicit a subjective probability mass function from each expert for each quantity. For each quantity we can meaningfully compare the information in each expert's distribution. However, this is not enough. We also need a way of scoring the experts with respect to information for the whole set of uncertain quantities. If the uncertain quantities have the same *intrinsic range* (for example, the uncertain quantity might be a percentage, or a truth value), then it is reasonable to define an information score for each expert as the information in his joint distribution for all the quantities in question. If the distributions are independent, then the information of the joint distribution is just the sum of the values $I(P)$ for each uncertain quantity.

If the uncertain quantities do not have the same intrinsic range then there are good arguments for not using the "joint information" as an information score. For example, if one of the uncertain quantities can take 10,000 possible values, then the miminal information for this quantity is $-\log(10{,}000) = -9.21$ ("log" denotes the natural logarithm). If the other quantities can take only one of two possible values, the minimal information for these quantities is $-\log(2) = -0.60$. Simply adding the information scores may therefore give inordinate weight to quantities with intrinsically larger ranges. In particular, if we rank the experts according to "joint information", then we may well find that the information rank is largely determined by the rank on the variable with the largest intrinsic range. If we do not wish the intrinsic ranges of the uncertain quantities to influence the information score, then the joint information score is not appropriate.

In the tests of Alpert and Raiffa (1982), which serve as a model for the tests described below, one of the uncertain quantities was "percentage of students preferring bourbon to scotch" (true value 42.5). Another was "number of eggs produced in the U.S. in 1965" (true value 6.5E10). As the intrinsic ranges of these quantities are vastly different, summing the information values for distributions for these quantities was deemed inappropriate. In the following section we discuss a method of scoring information in terms of relative information which is not sensitive to the intrinsic ranges of the uncertain quantities.

Calibration. If an expert consistently provided highly informative mass functions, while the true values always fell in the "tails" of his mass functions, then we would say that his assessments did not have a high degree of correspondence with reality. Roughly speaking, calibration is intended to measure the extent to which a set of probability mass functions "corresponds to reality".

To get an idea how a calibration score could be defined, suppose for the sake of argument that an expert gives the same probability mass function P for a large number n of physically unrelated uncertain quantities. By observing the true values for all these quantities we generate a sample distribution S' with $S'(i)$ equal to the number of times the value i is observed, divided by n.

It might appear reasonable to say that the expert is mis-calibrated if $S' \neq P$. Upon reflection, however, this is easily seen to be quite unreasonable. Suppose the true values represent independent samples from a multinomial random variable with distribution P. P certainly "corresponds to reality" (by assumption), but in general we will not have $S' = P$, as statistical fluctuations will cause P and S' to differ. In line with the intuitive definition of calibration given in the introduction, we might say that the expert was well calibrated if $S' = P$ in the long run. The problem with this, as Keynes was fond of saying, is that in the long run we are all dead. This definition gives us no way of measuring calibration for finite samples. We shall see shortly that "$S' = P$ in the long run" is a necessary but not a sufficient condition for calibration.

Roughly speaking, we want to say that the expert is well calibrated if the true values of the uncertain quantities can be regarded as independent samples of a multinomial random variable with distribution P. This entails that the discrepancy between S' and P should be no more than what one might expect in the case of independent multinomial variables with distribution P. We therefore propose to interprete the statement

"the expert is well calibrated"

as the statistical hypothesis:

$H(P)$ = "the uncertain quantities are independent and identically distributed with distribution P".

We want to define a calibration score as the degree to which the data supports the hypothesis $H(P)$. A procedure for doing this is described below.

The "discrepancy between S' and P can be measured by the *relative information of S' with respect to P*, $I(S', P)$:

$$I(S', P) = \sum_{i=0}^{\infty} S'(i) \log [S'(i)/P(i)]". \qquad (2)$$

Of course, $I(S', P)$ is not a metric, as $I(S', P) \neq I(P, S')$.

$I(S', P)$ may be taken as a measure of surprise which someone would experience if he believed P and subsequently learned S'. $I(S', P) = 0$ if and only if $P = S'$, and larger values correspond to greater surprise. Obviously, large values of $I(S', P)$ are critical for $H(P)$. We interprete the "degree to which the data supports the hypothesis $H(P)$" as the probability under $H(P)$ of observing a discrepancy in sample distribution S at least as large as $I(S', P)$ on n observations:

$$\text{Prob}\{I(S, P) \geq I(S', P) \mid H(P), n \text{ observations}\}. \quad (3)$$

This probability can be used to define statistical tests in the classical sense. Of particular interest is the following. If P is concentrated on a finite number M of integers which include all observed values, then as the number n of observations gets large, $2nI(S', P)$ becomes chi-square distributed. The number of degrees of freedom is $M - 1$ (see Hoel, 1971). The natural logarithm must be used in (2). Expanding the logarithms in (2) via a Taylor series and retaining the dominant terms yields the familiar chi-square statistic for testing goodness of fit between the sample distribution S' and the "theoretical distribution" P.

We call the above conditional probability *the expert's calibration score for the n observations*, and we propose to use this quantity to measure calibration in expert probability assessments. Good calibration corresponds to a high calibration score and is also a desideratum in expert resolution.

We can now understand why asymptotic convergence of S' to P is not sufficient for good calibration. Suppose that P is concentrated on six values so that the number of degrees of freedom of the chi-square distribution is five, and suppose that as the number of observations n goes to infinity, the expert's calibration score converges to 1%. From a chi-square table we conclude that $2nI(S', P)$ converges to 15. This entails that $I(S', P)$ converges to zero, and hence, that *S' converges to P*. However, for all n greater than some n_0, the hypothesis $H(P)$ would be rejected at the 5% significance level.

The basic principle of a classical approach to expert evaluation can now be outlined: good experts should have good information scores and good calibration scores. This theory is normative in the sense that it prescribes how experts *should* perform. In the third section we present evidence that experienced probability assessors do indeed perform better as a group with respect to both scores than inexperienced assessors, for items relating to their field. A few remarks are in order.

—There is a natural antagonism between good calibration and high information. Choosing a "tighter" function P will produce a higher information score, but will make it more "likely" that the true values will fall in the tails, producing a low calibration score. *A priori,* one would expect a negative correlation between the information and calibration scores, and this indeed is found.

—There is no *a priori* reason to expect the group of experienced assessors to perform better in their field of expertise than the inexperienced group in both scores. This requires an explanation. Our explanation, of course, is that experience teaches assessors to be better experts in the above sense. If this explanation is correct, then we should expect that outside their field of expertise the difference between the two groups on both scores should be smaller. This prediction is also borne out in the experiment discussed in Section 3.

—The assumption of independence underlying the probability (3) is generally unwarranted when referred to the subjective probability functions of the experts themselves. Indeed, subjective probabilities are dependent whenever the subject is prepared to "learn" from the observed values of the quantities involved. However, the individual or group seeking expert advice may well regard the quantities as independent, and may legitimately use the independence assumption in evaluating calibration.

—The above principles for evaluating expert probability assessors should be regarded as a "first pass". The calibration and information scores are "global" properties of assessors. In a concrete problem other aspects may be important as well, such as the existence of a bias in the central tendency or a bias toward "overconfidence".

2. *Implementation*

The approach presented above is not yet very practical since it requires a large number of quantities for which the expert gives the same probability mass function. We discuss two methods of implementing these ideas in terms of psychometric tests. The first method, the fractile method, is the one used in the experiment discussed in Section 3. The second method, the discrete method, corresponds to the most common type of calibration experiments found in the literature. Each method has its strengths and weaknesses. Roughly speaking, the fractile method is better for calibration and the discrete method is better for information. We also draw attention to methodological problems inherent in the design of most calibration experiments done to date.

The fractile method. The fractile method for soliciting subjective probability assessments (Morris, 1977; Alpert and Raiffa, 1982) provides an elegant technique for overcoming the difficulty mentioned above. Instead of soliciting the entire mass function from an expert, we can solicit various fractiles from his mass function. The r-fractile of the mass function P is by definition the smallest value i for which

$$\sum_{j \leq i} P(j) \geq r. \quad (4)$$

In the experiment described in the third section, the 1%, 25%, 50%, 75% and 99% fractiles are solicited for each uncertain quantity. After observing the true values, we note between which fractiles the true value falls. If the subject is perfectly calibrated, we should expect 1% of the true values to fall beneath the 1% fractile, 24% of the true values to fall between the 25% and the 1% fractile, etc. If we number these fractiles with $i = 1, \ldots, 5$, then we can let i denote the event that the true value falls between the fractiles number i and $i - 1$. "6" denotes the event that the true value lies above the 99% fractile. Obviously:

$$P'(1) = P'(6) = 1\%; \quad P'(2) = P'(5) = 24\%; \quad P'(3) = P'(4) = 25\% \quad (5)$$

where $P'(i)$ denotes the expert's subjective probability for the event i, $i = 1, \ldots, 6$. When we solicit the same fractiles for different uncertain quantities, variables with the same subjective distribution are generated. The theory of Section 1 can now be applied in a straightforward manner: by observing the true values for the uncertain quantities, we build up a sample mass function S', where $S'(i)$ denotes the fraction of the true values falling between fractiles number i and $i - 1$. As before, we can now calculate the *expert's calibration score* for the n observations, with P' replacing P in (3).

In measuring the information in the expert's distributions a number of problems must be addressed. In the first place, as we have solicited only certain fractiles from each expert's distribution, we can only approximate his true distribution. In the experiment described below we approximate the experts' distributions by distributing the appropriate probability mass evenly between the various fractiles. The mass functions obtained in this way are the maximal entropy distributions consistent with the constraints imposed by the expert's fractiles.

Secondly, the elicited fractiles enable us to locate only the central 98% of each expert's probability mass. We do not know how far the tails extend below the 1% fractile and above the 99% fractile. In scoring information in the experiment we cut all distributions off at the lowest 1% and highest 99% fractiles in the whole set of responses for the items in question. In other words, we consider these extremes as the 0% and 100% fractile for each distribution,

and this defines the intrinsic range for the item. The contribution to the information score from the tails is small.

We are still left with the problem of different intrinsic ranges discussed in the previous section. Each uncertain item is given in certain units, for example degrees celsius, percentages, kilograms, etc. Let Rg denote the intrinsic range for the item as specified above; that is, the set of integers between the highest 99% fractile and the lowest 1% fractile. Let P denote the maximal entropy approximation to the expert's distribution and let $P'(i)$, $i = 1, \ldots, 6$ denote the distribution given in (5). Let n_i denote the number of units falling between the i-th and the $(i-1)$-th fractiles ($i = 0$ corresponds to the 0% fractile, and $i = 6$ to the 100% fractile, i.e. the lower and upper endpoints of Rg). If j is an integer falling between the i-th and $(i-1)$-th fractiles, then clearly

$$P(j) = P'(i)/n_i.$$

Applying equation (1), the information in P is given by:

$$\begin{aligned} I(P) &= \sum_{j \in Rg} P(j) \log P(j) \\ &= \sum_{i=1}^{6} P'(i) \log (P'(i)/n_i) \\ &= \sum_{i=1}^{6} P'(i) \log P'(i) - P'(i) \log (n_i). \end{aligned} \tag{6}$$

One way of eliminating the effect of different intrinsic ranges for different items would be to rescale all the items so that the number of units in the range of each item was the same. If we multiply each n_i in (6) by a constant c then the information will be decreased by the amount $\log (c)$.

A more elegant way of dealing with the problem of differing intrinsic ranges is to calculate the information in P relative to the uniform distribution over Rg. Let U denote this uniform distribution and let $n = \sum n_i$, then:

$$I(P, U) = \sum_{i=1} P'(i) \log (nP'(i)/n_i). \tag{7}$$

Whereas (6) has a maximum value of zero and a minimum value of $-\log (n)$, (7) has a minimum value of zero, and no finite maximum*. Moreover, $I(P, U)$ is not affected by a change of scale, as the constant drops out in the logarithm. Hence, the relative information is not sensitive to the intrinsic range of the quantity in question. If we consider two independent uncertain quantities, then it is easy to check that the information of the product of the expert's distributions relative to the product uniform distribution is just the sum of the relative informations given by (7). We conclude that $I(P, U)$ provides a suitable method for scoring information in the fractile tests. The total information score for an expert is defined as the sum of the values $I(P, U)$ for each item.

Measuring calibration with fractile tests is not without its disadvantages. First, it demands a bit of mathematical sophistication on the part of the experimental subjects. Second, it can only be applied for variables whose ranges are continuous, for all practical purposes. Third, there is evidence that the score is influenced by the order in which the fractiles are solicited (see Pickhardt and Wallace, 1974 and Selvidge, 1980). On the other hand it has the very important advantage of producing uncertain quantities for which the subjective distributions are identical.

The discrete method. The fractile method for measuring calibration is not the method most commonly used. A typical calibration test involves giving the subject a set of statements, asking the subject to determine whether the statements are true or false, and asking him to state his probability that his choice is correct (for variations on this method see Lichtenstein *et al.*, 1982). The probabilities are coarse-grained into discrete probability bins, usually 50%, 60%, 70%, 80% and 90%. In this way we generate sets of identically distributed variables, one set for each bin. For each probability bin we can apply the foregoing theory directly. The problem now is to define scores for all the bins simultaneously.

A calibration score could be introduced as follows. If p_i is the probability corresponding to the i-th probability bin, then we can associate the probability distribution $\{p_i, 1-p_i\}$ with the i-th bin. A sample distribution $S_i = \{s_i, 1-s_i\}$ can also be associated with the i-th bin by considering the fraction s_i of items in the i-th bin which are correct.

What does it mean to be well calibrated on this type of test? In accord with the discussion in the previous section, we must identify good calibration with a statistical hypothesis. We shall say that a subject is well calibrated for a discrete test if he is well calibrated for each bin in the sense defined previously, and if the items in distinct bins are independent.

If the subject is well calibrated and if the items are properly chosen, then we should expect the fluctuations in different bins to be independent. A complication arises here as the total number of items, n, is fixed beforehand, and this can introduce a dependency between different bins. Roughly speaking, if $n-1$ items have been placed in the first bin, then the fluctuations in the remaining bins are severely constrained. Under suitable conditions, however, these fluctuations are asymptotically independent in n. Let n_i be the number of items in the i-th bin. We have:

$$\begin{aligned} &P\{2n_i I(S_i, P_i) \leq K\} \\ &\quad = \sum_{L=0}^{\infty} P\{2LI(S_i, P_i) \leq K \mid n_i = L\} P\{n_i = L\}. \end{aligned} \tag{8}$$

For large L the conditional probabilities approach the value given by the chi-square distribution with one degree of freedom. Therefore, if $P\{n_i = L\}$ is sufficiently small for small L, we may approximate the right-hand side with its asymptotic value. The dependencies between the distributions for the relative information statistics for different bins under these conditions disappear as the total number of items n gets large. This entails that:

$$R := \sum_i 2n_i I(S_i, P_i) \tag{9}$$

is the sum of asymptotically independent asymptotic chi-square variables with one degree of freedom. It follows that R itself is asymptotic chi-square with k degrees of freedom, where k is the number of probability bins. Hence, R could be used to define a likelihood function which in turn could be used as a calibration score as in the case of the fractile tests.

Concerning the information score, we note that in a discrete test all variables have the same intrinsic range, namely "correct" and "not correct". The quantity

$$I := \sum_i n_i I(P_i) \tag{10}$$

---

* In saying that $I(P, U)$ has no finite maximum we assume that P can be concentrated on fractional units. If the uncertain item is given in kilograms, for example, P may be concentrated on grams. If we do not admit this possibility, $I(P, U)$ has a maximum at $\log (n)$.

It may be that the units in which the uncertain items are given influence the subject's distributions. Asking for the number of eggs produced in the U.S. in 1965 may produce distributions with higher information than asking for the number of "mega-eggs" produced in the U.S. in 1965. It is our impression that this type of influence may be present to some degree on the general knowledge items, and that the information scores for these items are "more noisy" than for the technical items.

We therefore analysed the information with a different method as well. In the second analysis we computed the information for each uncertain item and ranked the subjects with respect to information for each item. A total information rank was determined by adding the ranks for each individual item. The results reported as significant were also significant under this alternative scoring method, with one exception. For the experienced group on the technical test the (negative) correlation between calibration and information ranks was not quite significant.

is the information of the joint distribution for all items when the items are independent and the probability for each item is the probability corresponding to the bin in which it has been placed. For a discrete test, I defines a suitable information score.

An example will help illustrate this quantity. Suppose for one year two experts are asked each day to give their probability of rain on the next day. The probability bins run from 10% to 90%. Both experts know that the yearly "base rate" of rainy days is 20%. The first expert simply predicts rain with 20% probability each day and may expect that he will be well calibrated. The second expert tries to distinguish between days in which rain is more or less likely. As he is also aware of the base rate, he will assign days to probability bins such that:

$$\sum_j (n_j/n)p_j = 20\%.$$

Let $I(E_i)$ denote the information score of expert i. Considering $I(p_j)$ as a function of p_j, we note that $I(.)$ is convex, so by Jensen's inequality:

$$I(E_1) = nI(20\%) \leq \sum_j n_j I(p_j) = I(E_2).$$

Under the hypothesis of independence, the information contained in the second expert's responses is greater or equal to that of the first. The second expert need not be well calibrated. However, he will be well calibrated if in addition to the above, his sample distribution satisfies:

$$s_i = p_i; \quad \text{for all } i.$$

Concerning the calibration score, the discrete method is inferior to the fractile method in two respects. First, the speed of convergence in (9) is determined by the numbers n_i which the experimentor does not control. We should also recall that according to standard statistical procedure, the chi-square approximation in (9) is appropriate for the 90% bin only if about 50 items are placed in this bin. Second, no account is taken of the fact that the subject's probabilities are coarse-grained by placing them into discrete probability bins. On the other hand, the information score (10) represents a distinct improvement over the score used in the fractile tests.

Methodological difficulties. The score R introduced above is not found in the literature, so far as we know. In the more recent literature the *weighted Euclidean distance score* is used. Using the above notation, this score is defined as:

$$\sum (p_i - s_i)^2 n_i/n.$$

Murphy (1973) extracted this score from the Brier score (1950). As pointed out by Lichtenstein *et al.* (1982), the sampling properties of this score are not known. Indeed, this score depends not only on the sample distribution in the various bins, but also on the number of items which the subject puts in the various bins. The experimentor does not have this number under control and does not know its statistical behavior. When the results of an experiment are reported in terms of this scoring variable, we have no way of distinguishing miscalibration from mere statistical fluctuations.

In practice, people seem to rely on gathering a large amount of data. How much data is necessary? Of course, without performing a statistical analysis it is impossible to say. To get a rough idea of the size of the statistical fluctuations, suppose a subject is given 15 items and puts them all in the 90% bin. Suppose 73% of these items (i.e. 11) are in fact true. Assuming the items are independent we can represent the subject with a binomial model and easily check that the hypothesis "well calibrated" could not be rejected at the 5% significance level. If the subject distributes 15 items over five probability bins more or less evenly, then it may be very difficult to reject the hypothesis that he is well calibrated.

One way of obtaining more responses, of course, is to give the subject more items. As this method places demands on both the subject and the experimentor, the method often used in practice is to give the same items to more subjects. This method is employed when one is interested in a set of subjects which have been selected according to some common characteristics, or have been "treated" in some particular way.

However, this way of obtaining more data is problematic. If the subjects place the same items in the same bins, then the number of *independent* items in the bins need not increase, or need not increase very rapidly. In general, we should expect that different subjects should have a tendency to place the same items in the same bins, especially when the subjects have been pre-selected or treated in some way. In any event, the number of independent items in each bin cannot exceed the number of items given to each subject. Experiments involving a small number of items per subject (on the order of 10) and a large number of subjects (on the order of 50–100) are not uncommon in the literature (see for example Lichtenstein and Fischhoff, 1977). Experiments of this type always neglect to report the multiplicities of the items in the various bins. It remains to be demonstrated whether such experiments yield statistically significant results.

3. *The experiment*

A series of Bayesian tests was recently conducted at a training facility for operators of large technological systems. The purpose of these tests was to investigate whether subjects with more practical experience exhibit more conformity with the axioms of Bayesian decision theory in areas related to their specific competence, and if so, whether this generalizes to other areas. Overall results of this study will be described in another publication (Thijs, 1987). We focus on the calibration tests.

The subjects. The experimental subjects fell into two groups. One group, the inexperienced operators, was in the last year of a 5-year training program, roughly equivalent to a bachelor of science program at an American university. Their field of study is mechanical engineering. All these subjects were between 20 and 23 years of age, and had completed a course in statistics.

The second group, the experienced operators, had all completed the training program. Their average age was 36 years and they had on the average 15 years of practical experience. Some of them were teachers at the training facility. Twenty two inexperienced and 12 experienced subjects took both general knowledge and expertise-specific calibration tests. Three additional experienced operators took only the general knowledge test. All subjects were male.

The tests. The tests were modelled on the fractile calibration tests of Alpert and Raiffa (1982). Some of the general knowledge items were taken literally from this test, and others were adapted to the situation in Holland. The following are examples of uncertain quantities from the technical test:

the maximal efficiency of the Tyne RM1A gas turbine

the maximum admissible intake temperature for gas in the Olympus power turbine.

Each test contained 10 uncertain quantities, and for each quantity the 1%, 25%, 50%, 75% and 99% fractiles were solicited. The format of the test was such that the subjects themselves determined the order in which the fractiles would be chosen. The tests were explained in detail, and a question was worked out and discussed beforehand as an example.

Scoring and results. Calibration was scored in the manner set forth in the second section for fractile tests. For each subject the relative information of his sample distribution with respect to the distribution P' in (5) was calculated for both the general knowledge test and the technical test. As the number of uncertain quantities was the same for all subjects, $I(S, P')$ was used as an index for calibration. We then rank-order the entire set of subjects for the two tests, with rank 1 corresponding to the lowest calibration index. The results are presented in Table 1.

Using the Wilcoxon two-sample test we determine whether

TABLE 1. CALIBRATION AND INFORMATION SCORES ON THE GENERAL KNOWLEDGE AND TECHNICAL TESTS

	Calibration rank		Information rank	
Subject number	General knowledge	Technical	General knowledge	Technical
1	8	2	13	23
2	27	31	1	7
3	3	25	31	4
4	6.5	6	24	10
5	17	1	22	24
6	25.5	4	18	33
7	11	7	36	22
8	22	15	15	5
9	12	11	11	16
10	29	3	6	1
11	19.5	22	27	3
12	25.5	20	28	2
13	31	—	5	—
14	23	—	21	—
15	18	—	26	—
16	4	12	19	30
17	24	26	10	27
18	9	13	17	26
19	1	5	25	28
20	32	30	29	8
21	5	23	32	9
22	19.5	8	34	34
23	2	19	33	19
24	10	24	35	15
25	21	17.5	8	20
26	33	16	37	32
27	14.5	21	30	13
28	34	17.5	14	18
29	30	32	23	12
30	35	29	12	17
31	16	27	2	25
32	14.5	14	20	29
33	36	33	3	6
34	6.5	9	9	31
35	37	10	7	14
36	28	34	16	11
37	13	28	4	21

Subjects 1–15 were experienced operators.
Subjects 16–37 were inexperienced operators.

the experienced operators are significantly higher ranked with respect to calibration in the general knowledge and the technical tests. This was indeed the case for the technical test (significance level 0.012) but not for the general knowledge test.

For each uncertain quantity, we also score the entire group of subjects with respect to the information relative to the uniform distribution for that quantity. We do this in the manner set forth in the second section by distributing the probability mass evenly between the solicited fractiles and applying equation (7). The values for the information relative to the uniform distribution for each item are added to determine a joint information score, and these scores are then ranked (highest information corresponding to rank 1). The results are presented in Table 1. The Wilcoxon two-sample test for determining whether experienced operators had significantly more information in their distributions yielded results almost identical to the results for calibration. On the technical test the experienced operators had significantly more information (significance level 0.022), but not on the general knowledge test.

The Spearman rank correlation coefficient test was used to test the null hypothesis "good calibration is not correlated with low information". For the experienced group on both tests and for the whole group on the general knowledge test there was significant correlation at the 5% level. For the inexperienced group on both tests and for the whole group on the technical test there was significant correlation at the 1% level.

On the technical test one (experienced) subject was extremely well calibrated (rank 3) and extremely informative (rank 1). This subject emerges as a very good expert. Interestingly, there was no "good expert" for the general knowledge items. The above-mentioned individual was ranked 29th and 6th for calibration and information respectively on the general knowledge test (see Table 2).

A chi-square table may be used to determine the calibration score (3) for these subjects. Since there are 10 items in each test and five degrees of freedom in the "theoretical distributions", a calibration index greater than 0.6 would be significant at the 5% level. If we regard the uncertain quantities as independent, we could not reasonably believe that scores higher than 0.6 were produced by statistical fluctuations. Of the 34 subjects participating in the technical calibration test, 31 would be regarded as miscalibrated at the 5% level. For the general knowledge test, these figures are 37 and 33, respectively. For the experienced group nine of the 12 would be rejected at the 5% level on the technical test, and 14 of the 15 on the general knowledge test.

Since the expected number of observations in the tails of

TABLE 2. GRAPHICAL REPRESENTATION OF THE CORRELATIONS BETWEEN CALIBRATION AND INFORMATION RANKS FOR THE TOP THREE RANKED SUBJECTS ON THE GENERAL KNOWLEDGE AND TECHNICAL KNOWLEDGE TESTS

	General knowledge		Technical knowledge	
Rank	Calibration rank	Information rank	Calibration rank	Information rank
1	i	e	e	e
2	i	i	e	e
3	e	i	e	c
4	i	i	e	e
5	i	e	i	e
6	i	e	e	i
7	e	i	e	e
8	e	i	i	i
9	i	i	i	i
10	i	i	i	e
11	e	e	e	i
12	e	i	i	i
13	i	e	i	i
14	i	i	i	i
15	i	e	e	i
16	i	i	i	c
17	e	i	i	i
18	e	e	i	i
19	i	i	i	i
20	e	i	e	i
21	i	e	i	i
22	e	e	e	i
23	e	i	i	e
24	i	e	i	c
25	e	i	e	i
26	e	e	i	i
27	e	e	i	i
28	i	e	i	i
29	e	i	i	i
30	i	i	i	i
31	e	e	e	i
32	i	i	i	i
33	i	i	i	e
34	i	i	i	i
35	i	i		
36	i	e		
37	i	i		

i denotes inexperienced operator; e denotes experienced operator.

the theoretical distributions is quite small (0.1 for each tail), the chi-square approximation is not very reliable. For example, the theoretical probability of finding more than one event in the tails is 0.017, but two observations in either of the tails contribute only 0.599 to the calibration scores.

As with other test reported in the literature, these calibration results can be described as poor. One way of judging this is to compare the number of rejected assessors with the number which would be rejected if in (3) the uniform distribution were used instead of P' in calculating the experts' calibration score. On the technical test, only two experienced operators and three inexperienced operators would be rejected at the 5% significance level in this case. This means, roughly speaking, that most of the subjects behave more as if they were asked for the fractiles 17%, 33%, 50%, 67% and 83%.

The "surprise index" is the percentage of true values falling in the tails of the solicited distributions. On the general knowledge test the surprise indices for experienced and inexperienced operators were 43% and 45%, respectively. On the technical test these were 30% and 43%.

In spite of the overall poor showing in calibration, we think it important to emphasize that the experienced group was significantly better in their area of expertise. Apparently, calibration and information do measure something objective about being an expert.

Several attempts have been made in the past to relate calibration to "knowledge". Adams and Adams (1961) found no correlation between knowledge and calibration for subjects taking a final examination. Sieber (1974) found similar results. Attempts to calibrate experts in the area of their expertise have produced sharply divergent results. Some groups of experts show excellent calibration, other groups show very poor calibration. Lichtenstein and Fischhoff (1977) found a negative correlation between calibration and difficulty on general knowledge items and found that calibration first improves, then declines with increasing knowledge. Another study (Lichtenstein *et al.*, 1982) found that calibration improved as the number of true statements among the test items was increased. The more recent research suggests that there may be some relation between calibration and knowledge. We hope the present results help clarify this relation.

4. *Conclusions*

It is interesting to compare the classical and the Bayesian approaches to expert resolution with respect to the foregoing experiment. Very roughly, a Bayesian would use prior information to recalibrate each expert, and would use each recalibrated expert to update his distribution on the decision variable of interest. The classical approach would say: on items similar to those on the technical test, take the advice of the expert who was ranked high on the both calibration and information.

Much theoretical and empirical work remains to be done. On the theoretical side, it would be desirable to have more rigorous methods for chosing a "best expert", as the choice will not always be as clear as in the foregoing test. It would also be desirable to investigate other possible parameters for expert resolution. It was noted that the chi-square approximation is not very good under normal test situations when very small or very large fractiles are solicited. It would be useful to have a better approximation.

On the empirical side, it is important to get a consistent picture of the relation between calibration and knowledge, and a coherent view of the reliability of expert opinion. It is unlikely that this can be achieved without first clearing up the methodological difficulties surrounding calibration measurements. More empirical work needs to be done on the factors influencing calibration and overconfidence, and on the possibility of training people to be good probability assessors. The results of the experiment discussed in the third section leave no doubt that the experts in this study have ample room for improvement. That which "nature" teaches inefficiently is what training programs should teach efficiently.

Acknowledgements—The authors gratefully acknowledge several helpful suggestions by two anonymous referees of *Automatica*.

References

Adams, J. K. and P. A. Adams (1961). Realism of confidence judgments. *Psychol. Rev.*, **68,** 33–45.

Agnew, C. E. (1985). Multiple probability assessments by dependent experts. *J. Am. Statist. Ass.*, **80,** 343–347.

Alpert, M. and H. Raiffa (1982). A progress report on the training of probability assessors. In Kahneman, D., P. Slovic and A. Tversky (Eds), *Judgment under Uncertainty: Heuristics and Biases*, pp. 294–306. Cambridge University Press, U.K.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weath. Rev.*, **75,** 1–3.

Cooke, R. M. (1987). A theory of weights for combining expert opinion. Delft University of Technology, Department of Mathematics, Report no. 87–25.

De Groot, M. and S. E. Fienberg (1983). The comparison and evaluation of forecasters. *The Statistician*, **32,** 12–22.

De Groot, M. and S. Fienberg (1986). Comparing probability forecasters: basic binary concepts and multivariate extensions. In Goel, P. and A. Zellner (Eds), *Bayesian Inference and Decision Techniques*. Elsevier, Amsterdam.

Genest, C. and M. Schervish (1985). Modelling expert judgement for Bayesian updating. *Ann. Statist*, **13,** 1198–1212.

Harrison, J. M. (1977). Independence and calibration in decision analysis. *Mgt Sci.*, **24,** 302–328.

Hoel, P. (1971). *Introduction to Mathematical Statistics*. Wiley, New York.

Kempthorne, P. and M. Mendel (1987). *Adjusting Experts' Probabilistic Forecasts*. MIT Press, Cambridge, MA.

Lichenstein, S. and B. Fischhoff (1977). Do those who know more also know more about how much they know? *Orgl. Behaviour Human Perform.*, **20,** 159–183.

Lichenstein, S., B. Fischhoff and D. Phillips (1982). Calibration of probabilities: the state of the art to 1980. In Kahneman, D., P. Slovic and A. Tversky (Eds), *Judgment under Uncertainty: Heuristics and Biases*, pp. 306–335. Cambridge University Press, U.K.

Lindley, D. V. (1982). The improvement of probability judgments. *Jl. R. Statist. Soc. A.*, **145,** 117–126.

Mendel, M. and W. Thijs (1983). Fault management and subjective probability. Proc. 3rd European Conf. on Manual Control and Human Decision Making. Roskilde, Denmark.

Morgan, M., M. Henrion and S. Morris (1979). Expert Judgments for Policy Analysis Report. Brookhaven National Laboratory, New York.

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

Morris, P. (1977). Combining expert judgments: a Bayesian approach. *Mgmt. Sci.*, **23,** 679–693.

Murphy, A. (1973). A new vector partition of the probability score. *J. Appl. Met.*, **12,** 595–600.

Pickhardt, R. and J. Wallace (1974). A study of the performance of subjective probability assessors. *Decis. Sci.*, **5,** 347–363.

Roberts, H. V. (1965). Probabilistic prediction. *J. Am. Statist. Ass.*, **60,** 50–62.

Selvidge, J. (1980). Assessing the extremes of probability distributions by the fractile method. *Decis. Sci.*, **11,** 493–502.

Sieber, J. (1974). Effects of decision importance on ability to generate warranted subjective uncertainty. *J. Personality Social Psychol.*, **30,** 688–694.

Siegel, S. (1956). *Nonparametric Statistics*. McGraw-Hill, New York.

Thijs, W. (1987). Fault Management. Ph.D. dissertation, Delft University of Technology, The Netherlands.

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

Winkler, R. (1986). On good probability appraisers, in Goel, P. and A. Zellner (Eds), *Bayesian Inference and Decision Techniques*. Elsevier, Amsterdam.