*Earthquake Engineering, Tenth World Conference © 1992 Balkema, Rotterdam. ISBN 90 5410 060 5*

# Calibrated expert judgement in seismic hazard analysis

G.Woo
*BEQE Limited, Cambridge, UK*

ABSTRACT: Modern probabilistic seismic hazard assessment involves the specification for input seismological parameters of weighted distributions, which reflect both the available observational data and subjective degrees of belief. The latter requires the formal elicitation of expert judgement, which is prone to distortion because of the poor calibration and uninformativeness of some individuals. Within the context of a review of elicitation methods, attention is drawn to a method for calibrating expert judgement relevant to seismic hazard analysis.

## 1 INTRODUCTION

In engineering seismic hazard analysis, seismic loading criteria depend on the development of simplified numerical models of regional seismic sources and ground motion attenuation. With extreme events and ground motion levels being of primary concern, these models typically lack direct observational support in the parameter regimes most relevant to the design of critical facilities. Inevitably, there must be some measure of data extrapolation and appropriation from other regions, for these models to be fully functional. Clearly, this measure is greater in areas of occasional sporadic seismic activity, where the earthquake catalogues and strong-motion databases are sparse. To some extent, geophysical theory can substitute for missing data, but seismological predictions of earthquake recurrence and seismic wave propagation are not yet sufficiently well advanced for theoretical simulations to be accepted as a surrogate for local earthquake observations.

Given the finite limitations of the scientific data available for a seismic hazard assessment, and the long return periods often specified in seismic design regulations, expert judgement has to be exercised in the parameterisation of a seismic hazard model. In order for this judgement to be well informed, and be kept to a practical minimum, it is essential that as much scientific research as possible be undertaken into the regional seismicity and seismotectonics; judgement should not be excused as a replacement for scientific investigation.

However, when observational and budgetary limits are reached, judgement must be elicited. The manner in which this is done is as important as any procedure in seismic hazard analysis. It is unfortunate that the term 'judgement' has a connotation which may suggest a lack of structure and informality. This interpretation finds no support in the risk analysis community, nor can it be sustained by the null argument that details of the procedure make no difference. Indeed, it is well appreciated by professional seismic hazard analysts (Reiter 1990) that significant diversity in expert opinion is primarily responsible for the disparity in seismic hazard curves produced for a site, as exemplified in Figure 1. This diversity highlights the aggregation of expert judgements as the most critical outstanding problem in seismic hazard analysis.

At the outset, it is important to distinguish between carefully elicited judgement on seismic model parameters, and what is often termed and understood as 'engineering judgement'. The latter is associated with professional decisions made in engineering practice. By contrast, expert judgement on seismological and geological parameters calls for the methodical interpretation and synthesis of scientific data, so that the uncertainty in parameter values can be represented statistically in terms of probability distributions. Lack of distinction between the two forms of judgement has led to a mongrel breed of seismic hazard analysis, notable for the feature that parameters are assigned single so-called 'best-estimate' values, arrived at through informal use of judgement.

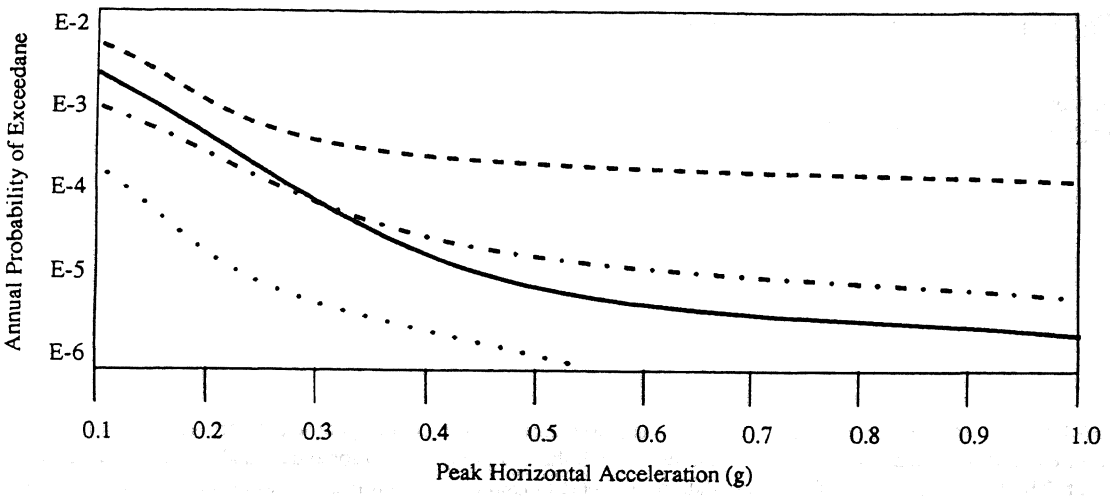


Figure 1: Inter-expert variation in expected seismic hazard curves for a site with sparse regional data

## 2 ELICITATION METHODS

Apart from ad hoc approaches, a broad spectrum of auditable methods now exist for the elicitation of expert judgement, and a number of options have been adopted in practice. Amongst the choices are the number and selection of experts; the encoding of judgements; the treatment of potential bias; the accountability of experts, etc... For each procedural choice, there may be mathematical, logistical, financial or even psychometric reasons to favour one alternative over another.

A principal discriminant of elicitation methods is the provision for group interaction. The main point in favour of experts interacting with each other, either directly at meetings or indirectly through access to the other responses, is that differences of opinion can be identified, and some form of consensus may be reached. On the other hand, technical experts tend to be familiar with other opinions, and contrary views may be suppressed in the quest for a consensus.

In one of the pioneering group interaction methods, named Delphi, experts are each sent a questionnaire with boxes to be filled with parameter estimates. The returned questionnaires are processed to allow the mean, and upper and lower quartile values to be determined. The respondents, whose anonymity is preserved, are then sent back the results, and offered the opportunity for revision. The revised answers are returned for further evaluation, and the process is iterated several times.

A practical problem with the Delphi method is the drop-out rate from one iteration to the next, and the burden of written explanation placed on those whose estimates fall outside the interquartile band. Two alternative ways exist of dealing with contrasting responses: the simpler is just to aggregate the disparate opinions without participant interaction; the more sophisticated route is to provide the opportunity for participants to confront each other directly, after their initial assessments have been given. Of course, if experts are permitted to meet after initial expression of their opinions, a framework could be found for them to meet and make decisions from the start.

Different ways of achieving effective group decision making have been explored within the framework of decision analysis. One of the most successful methods is decision conferencing. This approach was originally developed in the U.S., and later adopted in Britain and elsewhere. The standard format of a decision conference involves a selected group of experts meeting over a period of two days to discuss their judgements, exchange information, hidden agendas and motives, prejudices and opinions, and seek ultimately to reconcile differences. The current state-of-the-art procedures (Bonano 1990) involve the selected experts in a strict regime of methodology induction, training, and elicitation sessions. Apart from the experts, a conference facilitator has to be in attendance. He is the lead analyst, knowledgeable about decision analysis, and experienced in the organization of group meetings.

## 3 SEISMIC HAZARD ELICITATION

Decision conferencing methods have been adopted in both formal and informal guises to arrive at consensus judgements on model parameters required for seismic hazard computations. Where a team of engineering seismologists and geologists are engaged in a seismic hazard investigation, the joint discussions may perhaps be facilitated and documented according to the precepts of decision conferencing. More likely, the meetings will be loosely structured, without adherence to formal elicitation procedures.

Whatever consensus is reached in a decision conference, the question may be raised as to the outcome if another group of experts had been chosen. A collective bias may distort the judgement of a particular group of experts attending a decision conference. The diversity of outcomes which may arise from alternative convocations of experts is shown in the assessment of U.S. seismic hazard (EPRI 1986). In this study, no less than six separate Earth Science teams of four to eight members were fielded to address the scientific issues. Even with access to a common information base, and the opportunity for open discussions in workshops, essential systematic differences of data interpretation and theoretical opinion remained to influence group judgements. No consensus between the six groups could have been reached without coercion and the suppression of opinion, and none was attempted. Instead, classical statistical estimation techniques were introduced to combine the hazard estimates of the various groups of experts.

Other than team elicitation techniques, a number of seismic hazard case studies have been undertaken, in which the opinions of experts have been individually elicited. The most direct approach (Okrent 1975) involved seven experts in providing a direct estimate of the likelihood of ground motion levels being exceeded at eleven U.S. nuclear power plant sites. At one site, at remote exceedance levels, the estimates ranged over five orders of magnitude. This remarkable spread of values motivates the search for rational methods for aggregating expert judgements. A number of possibilities exist for such an aggregation procedure. One approach (TERA 1980) invited experts to self-weight their own opinions on each question. In principle, this seems a fair way of discerning truly informed opinions, but this approach is open to distortion through experts modestly understating their knowledge, or immodestly dissembling their ignorance.

Rather than self-weighting, experts might be invited to assign weights to each other's opinions on each question. This would be somewhat invidious, even if the weights were assigned anonymously. The simplest procedure which avoids subjective weighting is to accord equal weight to each opinion, thereby treating all experts equitably. This latter option has been followed within the context of a seismic hazard study for the Cascadia zone in the western U.S. (Coppersmith and Youngs 1990). Expert opinion was elicited through a two-stage process of interviews at the offices of the experts, with several elicitators present. Similar to the Delphi method, the second stage of interviews took place with the experts provided with summaries of all the first stage responses. With a delay of about a year between the two stages, and access to the judgements of the other experts, changes of opinion could be freely made at the second stage.

A notable outcome of this study is the significant disparity between the hazard curves generated separately from the source models devised by the fourteen experts. Even without contemplating differences of opinion on attenuation, the disparity found between the median hazard curves of two experts is as much as an order of magnitude in exceedance probabilities for a given peak acceleration. Such a discrepancy could have important engineering implications for seismic loading criteria.

The large variation in the profiles of individual hazard curves begs the question of their consistency with the evidence of Intensity recurrence at the site. This qualitative check is available wherever the duration of the historical record of earthquake documentation is more than a hundred years old, and allows an estimate to be made of the return period of low levels of ground shaking. If a particular expert happens to judge seismic input parameters very conservatively, his resulting hazard curve may well suggest comparatively high levels of ground motion at short return periods: these may conflict with inferences drawn from the observational record of earthquake felt effects at the site.

The practice of simply averaging the judgements of the individual experts fails to discriminate between them on any grounds, even though reasons may exist to suspect the value of some of them. Furthermore, the maintenance of expert anonymity is itself a practice which discourages such discrimination. Concern over simplistic methods of aggregating expert judgements leads on to the consideration of methods which involve the calibration of experts.

## 4 CALIBRATION OF EXPERT JUDGEMENTS

A salutary lesson learned from the elicitation of expert judgement in risk analysis, is that, however well qualified they may be, experts may differ widely in their ability to estimate parameter values and associated confidence bands. Some experts may be poorly calibrated, in the sense that their median estimates may be consistently biased and discordant with reality. Other experts may be over-confident, in the sense that they assign far too narrow uncertainty bands on their median estimates; still other experts may be uninformative, in the sense that they assign such broad uncertainty bands that little useful information is conveyed. The particular category into which an expert may fall cannot be determined a priori from credentials of academic qualifications or professional experience, but must be decided by some calibration test procedure.

Recently, an elegant mathematical approach has been devised (Cooke 1991) to calibrate expert judgements. This method overcomes many of the technical difficulties which have beset previous attempts, while being convenient and efficient for regular practical use. In Cooke's approach, each expert is asked to assess the values of a range of quantiles for a given set of calibration parameters, as well as for the parameters of actual interest. An optimal decision maker is then constructed on the basis of the performance of the experts on the calibration parameters. This construction is based on the mathematical theory of scoring rules to gauge the quality of experts. Although first conceived as a way of eliciting probabilities, scoring is a systematic way of rewarding the positive aspects of expert judgement, and is implicit in any performance-based human assessment.

The mathematical formalism is outlined as follows: Let $q_e$ (e=1,2,..E) be the distributions of the E experts whose judgements are elicited for an uncertain parameter. Then if $p_A$ denotes the analyst's distribution:

$$p_A = \sum w_e q_e; \text{ where } \sum w_e = 1; \; w_e \geq 0 \qquad [1]$$

For the discrete case, where a quantity can take n values, the entropy H(p) is defined to indicate the lack of information in the distribution:

$$H(p) = - p_i \ln (p_i) \qquad [2]$$

If s is a probability distribution over the same n values, the relative information of s with respect to p can be defined as:

$$I(s,p) = s_i \ln(s_i/p_i) \qquad [3]$$

I(s,p) is always non-negative, and I(s,p) = 0, if and only if s is identically equal to p. I(s,p) is commonly taken as an index of the information learned if one initially believes p , and later learns s is correct.

Let $X_1, X_2, \ldots X_M$ be the set of calibration variables; let $X_{M+1}, X_{M+2}, \ldots X_N$ be the variables of real interest; let $f_1, f_2, \ldots f_R$ be the quantile probabilities elicited; and let $G_{me}$ be the minimum information cumulative distribution for $X_m$, defined as the distribution for which the entropy is as large as possible and which satisfies the constraint that the $f_i$ quantile values agree with the assessments of expert e.

Indices of calibration and informativeness of expert e's assessments, respectively C(e) and H(e), can then be defined (Cooke 1991) in terms of I(s,p) and $G_{me}$, and a corresponding weight $w_e$ can be attributed to expert e according to the formula:

$$w_e \propto I_\alpha [ C(e)] . \{C(e)/H(e)\}; \qquad [4]$$

$$\text{where } I_\alpha(c) = 0 \text{ if } c < \alpha$$
$$= 1 \text{ if } c \geq \alpha \qquad [5]$$

The parameter $\alpha$ may be chosen to optimize the analyst's distribution $p_A$ in equation [1]. With this definition, the weight assigned to an expert e increases with the calibration index C(e), and decreases with H(e).

The availability of the free parameter $\alpha$ to optimize the analyst's distribution in a manifestly impartial maner, avoids critical judgement of the quality of individual experts, who might otherwise be self-conscious over personal accountability. In practical applications, the calibration term C(e) can vary over three orders of magnitude for a reasonably large group of experts, whereas the informativeness index rarely varies by a factor greater than about three. There is a natural trade-off between calibration and informativeness: an expert can achieve good calibration by being uninformative. However, high informativeness is not a substitute for poor calibration, and in Cooke's model, informativeness modulates the calibration term, providing a quantitative means of distinguishing between experts of similar calibration.

## 5 CALIBRATION AND INFORMATIVENESS

The calibration and informativeness of experts contributing to seismic hazard assessments are not matters which are generally made explicit. There is universal familiarity with the display of seismological data presented with uncertainty bounds and error bars, and with the notion of observational data presented as a histogram or as a distribution on a probability plot. However, the implicit uncertainty bounds, error bars, histograms and probability plots associated with subjective judgements are rarely mentioned, let alone shown.

Within the modern probabilistic paradigm, in which probabilities reflect degrees of belief rather than equate with frequencies of rare events, the methodical use of expert judgement is essential. Probabilities may change with the state of knowledge, whereas frequencies do not. It follows that the probability assigned to an event can only equal its frequency for certain states of knowledge, which pertain when the only information available about an event is its frequency. Inevitably, the smaller the exceedance probability of seismic hazard concern, the greater the demand on expert judgement to quantify degrees of belief, and thereby parameterize a hazard model.

Where an individual's degree of belief is vague and hazy, his probability distributions will tend to be flat and hence uninformative; where his degree of belief is full of conviction, his probability distributions will tend to be peaked and thereby informative. Poor calibration will be more serious in the latter case than in the former.

Within the logic-tree framework of seismic hazard computation, which requires input variables to be prescribed as weighted parameter distributions, the distortion in input associated with poor expert calibration and uninformativeness constitutes a major implicit source of variability undermining confidence in the results.

To illustrate the kinds of distortion which are possible in the specification of weighted distributions for seismological parameters, an exercise has been carried out in which ten experts were asked to specify lower 5%, 50% (median) and upper 95% values for the regional parameters relevant to the seismic hazard for a specific site in Iran. These parameters included maximum instrumentally measured magnitude; largest historical felt area; highest historical Intensity; highest recorded acceleration; the deepest recorded focal depth.

The results show that traits of conservative bias, poor calibration, over-confidence, and uninformativeness, which would distort seismic hazard judgement, whether the experts were elicited for their opinions singly or in a group. Figure 2 illustrates the cumulative findings, expressed for brevity of presentation in terms of a typical seismological parameter, normalized so as to have a possible range from 0 to 1, with 0.5 being the actual empirical value. For each of the ten experts, the 5%, 50% and 95% values are indicated. Expert 1 is seen to be biased and over-confident; expert 3 is well calibrated, but uninformative; expert 10 is conservatively biased, and also over-confident. The desirability of some form of calibration for experts is apparent.

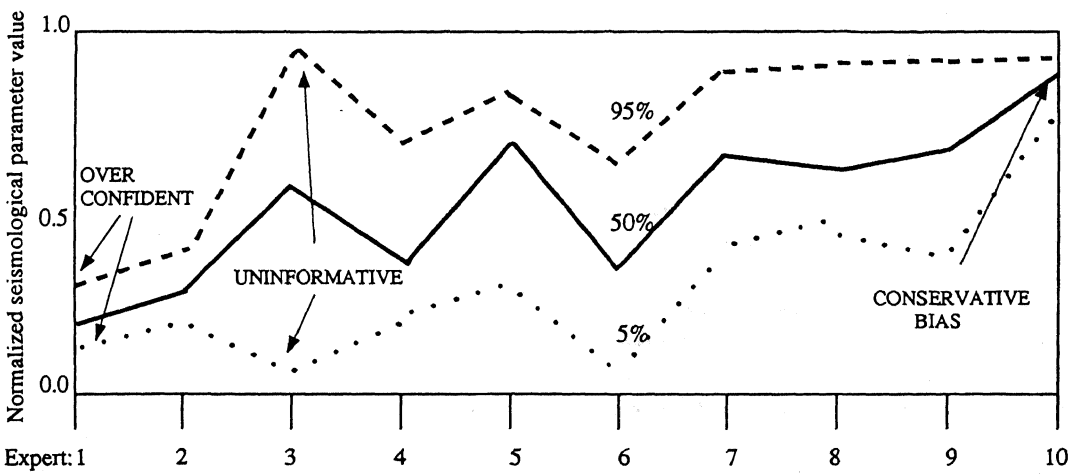


Figure 2. Inter-expert variation in 5%, 50% and 95% estimates of seismological parameters

## 6 CONCLUSIONS

The elicitation of expert judgement is an integral component of any seismic hazard assessment. The great diversity in the range of hazard curves obtained by different experts, or even teams of experts, highlights the crucial problem of aggregating expert judgements. Because the elimination of bias and the treatment of uncertainty are both critical to the reliability of a seismic hazard investigation, attention must be paid to the inherent tendencies individuals have to exaggerate or under-estimate values, or to be over-confident or self-effacing over their responses.

Calibration exercises allow these personal traits to be identified and quantified, and perhaps rectified through training on test cases. Whereas calibration is accorded little attention in the common elicitation schemes adopted in seismic hazard assessment, methods now exist for improving decision making through the optimal combination of expert judgements, taking account of variability in the calibration and informativeness of the selected experts.

The introduction of these optimal decision analysis techniques should systematically rationalize the important element of expert judgement present in all seismic hazard assessments, and thus improve their accuracy and reliability.

## REFERENCES


Bonano E.J, S.C. Hora, R.L. Keeney & D. von Winterfeld (1990). Elicitation and use of expert judgement in performance assessment for high-level waste repositories. *NUREG/CR-5411*, Washington.

Coppersmith K.J., R.R. Youngs (1990) Probabilistic seismic hazard analysis using expert opinion: an example from the Pacific North-West. *Geol. Soc. Amer. , Reviews of Eng. Geol.* VIII:29-45.

Cooke R.M. (1990) Experts in Uncertainty: expert opinion and subjective probability in science. Oxford: Oxford University Press.

EPRI (1986) Seismic hazard methodology for the Central and Eastern United States. *Report NP-4726.*

Okrent D. (1975) A survey of expert opinion on low probability earthquakes. *Annals of Nuclear Energy*, 2: 601-614.

Reiter L. (1990) Earthquake hazard analysis. New York: Columbia University Press.

TERA (1980) Seismic hazard analysis: solicitation of expert opinion, *NUREG/CR-1582*, 3, Washington.