Use of expert opinion elicitation to quantify the internal erosion process in dams

A. J. BROWN, KBR, Leatherhead, UK
W. P. ASPINALL, Aspinall & Associates, Beaconsfield, UK.

SYNOPSIS. Expert Opinion Elicitation is a generic term for a number of similar techniques that have been developed to provide quantitative estimates of parameters which cannot readily be quantified through direct measurement or other sampling techniques. The initial motivation for their development was the 1986 Challenger Shuttle disaster in the space industry, and subsequent applications have spread into many other areas: the techniques have been widely used in the nuclear industry, for instance. One particular procedure consists of obtaining responses to a set of quantitative questions from a number of experts, including the range of uncertainty in each response, and then combining these through a weighting procedure to obtain a pooled best estimate of the parameters of interest.

This paper describes an application of that procedure as part of a research contract to improve methods of early detection of progressive internal erosion in UK embankment dams. For some of the parameters, information is also available from a questionnaire circulated to British dam professionals, and the paper compares the outcomes produced by the two approaches. The paper concludes with comments on the future role that expert opinion elicitation could play in providing a better understanding of dam safety issues, in particular in the determination of relevant uncertainties.

INTRODUCTION

KBR are currently undertaking a research contract for the UK government (Department of environment, food and rural affairs, Defra) to "identify a cost effective means of early detection of progressive internal erosion in embankment dams". The terms of reference entail major emphasis on embankment dams which pre-date modern geotechnical engineering (no filters or instrumentation), and that the hazards posed by unprotected pipes and culverts require particular attention. The final output from the project is to be Technical Guidance on the management of internal erosion.

Long-term benefits and performance of dams. Thomas Telford, London, 2004

The approach adopted to respond to the terms of reference comprised a questionnaire to dam owners and panel engineers to identify recent case histories of internal erosion, a literature review and expert opinion elicitation. This paper describes the latter from the parameters selected for quantification, through the results it gave, lessons learned and where the technique could be of value in other areas relating to the management of high hazard industries.

EXPERT JUDGMENT AND ELICITATION OF EXPERT OPINIONS

General

In recent years, important changes have occurred in engineering which affect the way in which many safety-related decisions are made. These changes have resulted mainly from the development of risk-based methods for the design and appraisal of engineered systems. One feature of these methods is the objective of quantifying the level of safety in order to estimate the likelihood of engineering failure. The introduction of probabilistic concepts for treating uncertainty requires an engineer to exercise a form of judgment which differs from the conventional professional judgment that he (or she) may have developed during his career through training and practical experience. This alternative form of judgment, which surfaces in all attempts at estimating probabilities, in whatever domain, is generically termed 'expert judgment', and involves enumerating subjective probabilities that reflect an expert's degrees of belief. Hitherto, this subjective element in assigning probabilities has often been treated informally, or ignored altogether, but methodological advances, such as that reported here, are bringing this form of judgment increasingly to the fore.

Various approaches for combining expert opinions are possible (see, e.g., Cooke,1991; Meyer & Booker, 2001), including: *simple averaging, decision conferencing (the committee), the Delphi method, expert 'self-weighting', and the mathematical theory of scoring rules*. It is the latter that has been most refined by Cooke (1991), with his "Classical model" for expert judgment pooling (designated 'classical' because there is a close relationship with hypothesis testing in classical statistics). Cooke's scheme has been extensively tested and used in many areas of science and engineering, including the aerospace industry, nuclear industry, meteorology, hydrology (in the Netherlands), earthquake engineering and volcanology.

Examples of the use of expert elicitations in UK include:

a) O'Hagan (1998), where a consensus approach was used to address future capital investment needs of a major water company, and also in assessing the rock mass permeability at a possible nuclear waste repository at Sellafield
b) Aspinall & Cooke (1998), who describe the use of the structured elicitation methodology and decision-support procedure based on the "classical model" during the Montserrat volcanic eruption crisis, and
c) unpublished work on flight operations safety for British Airways (W.P. Aspinall, pers. comm.).

Classical method

The basis of Cooke's method is that the experts are posed a number of "seed" questions for which the answer is known (or knowable). Their responses are then assessed to obtain scores and individual weights, as defined in Table 1 and illustrated in Table 2; full mathematical details can be found in Cooke (1991). The procedure can be used to greatest benefit when the opinions of several experts (say, five or more) have to be elicited efficiently and promptly - for smaller groups, it may not be justified.

There are some important explanatory remarks in relation to Table 2. Firstly with only two seed questions, the number of degrees of freedom in the Chi-square test for the calibration statistic are too few to obtain results reflecting the accuracy of individual experts – hence Experts 1 & 2 have the same calibration score even though, in this example, one was more 'accurate' in his predictions than the other. Expert 3 falls between Experts 1 and 2 for informativeness, but falls below the threshold level for calibration (with Expert 4) when, as here, the DM's performance is optimised. Expert 4 is highly opinionated, and always fails to make his confidence limits wide enough to score any hits, but there is still a non-zero probability (0.007) that he is actually well-calibrated.

The fully-optimised DM has the highest calibration score, (when it is added to the group, as a virtual expert) but its Informativeness score appears poor because it amalgamates the spreads of all (positively weighted) experts. The DM's overall normalised weight is, therefore, slightly less good than the best real expert in this example, but then the DM's range reflects the collective spread of opinions. When optimised, the DM's 50%ile estimates for both seed questions are very close to the actual realizations, notwithstanding the scatter in the four experts' opinions.

In a real exercise, more seed questions are used for scoring the experts, and different combinations of statistical test power and significance level can be set to constrain relative performance scores across the group and DM.

Table 1. Basis of 'classical model' for combining experts' opinions – terms, scores, weights and factors

Term	Explanation / basis
Item	A 'seed' variable (for calibration purposes) or a question of interest for which an evaluation is sought from a group of experts
Calibration score	Test the hypothesis "This expert is well calibrated" with respect to his peers, on the basis of his estimates for a set of 'seed' variables. The score is the significance level in a chi-square test at which the hypothesis would be just rejected
Informativeness (Inverse is Entropy score)	a) Quantify the individual's 'informativeness' by indexing his cumulative information distribution function for all seed items relative to a uniform 'background' distribution (strictly, an inverse of a chi-square test statistic for closeness of correspondence); b) this 'background' distribution is either uniform linear (suitably truncated) or log normally distributed between quantiles; the latter is typically used when the range of possible values can vary over two orders of magnitude or more
Synthetic decision-maker (DM)	a) constructed from a weighted sum of the experts' responses to the items of interest, item-by-item. b) extracting the DM's distributions for each seed variable, the DM can be treated as a 'virtual expert' and scored against the seed items at different significance levels; the opinion of this virtual expert then can be iteratively re-combined with the real experts.
Expert weights	a) For each expert, the product of his calibration multiplied by informativeness scores across all seed items, normalized so that the sum of all expert weights, including that of the DM, is unity b) The 'classical model' software allows adjustment to the power of the chi-square test and the related significance level setting, which determines the threshold calibration score at which experts are given a non-zero weighting.

Table 2 : Illustration of scores and weights for 4 experts answering (only) two seed questions.

Expert	Experts' opinion ranges	Calibration Score	Inform. score	Normalized wt., incl opt. DM
1	*10, 35, 90* *15, 35, 80*	0.36	0.12	0.05
2	*40, 50, 60* *45, 52, 58*	0.36	1.27	0.52
3	*10, 25, 45* *15, 30, 55*	0.18	0.60	0
4	*80, 90, 95* *75, 80, 85*	0.007	1.60	0
DM		0.94	0.41	0.43
Results	Actual Seed values	5%ile	50%ile	95%ile
DM soln 1	50	*22.8*	*49.7*	*72.3*
DM soln 2	50	*26.4*	*51.8*	*66.8*

The rational mathematical basis for the 'synthetic decision-maker' is one feature of the method which makes it superior to other schemes for pooling judgments, making use of expertise weighted according to the quality of response to the whole set of seed variables. Usually, but not invariably, the DM ends up with a heavier weight than most, if not all, of the 'real' experts. Thus, the concept of the DM can also be described as the creation of a 'rational consensus', for the problem of combining a range of opinions (as opposed to reaching a simple average, democratic compromise or some other variant of egalitarian consensus). That said, in some applications, where suitable seed data are sparse or repeated tests are not possible, the scoring power of the calibration scheme may be weak, and its impact on individual weightings may have to be constrained.

Nonetheless, Cooke's method has at its heart a basis which replicates the formal scientific method, and one of its most valuable attributes is the scope it provides for quantifying realistically the spread of scientific or engineering uncertainty in relation to any parameter of interest. Thus, the procedure is usually framed to elicit suitable lower and upper percentile confidence estimates from the experts (in the present case 5%ile and 95%ile values), as well as a central or 'best' estimate value (which can be the mode, mean or median, depending on the distributional properties being sought). This aspect of the structured elicitation procedure is especially important for those variables for which adequate data do not exist for conventional statistical analysis - where the need for precise differentiation between engineering judgment and expert judgment comes into play.

APPROACH USED ON THIS PROJECT

The approach used on this project was based on that formulated by Cooke (1991), with the best estimate and 5% and 95% uncertainty distribution quantified for each item. To avoid peer pressure biases, the responses of the individual experts are provided independently by each directly to the facilitator, everyone remaining anonymous when the results are reported back to the group of experts. In the present project, the full set of questions had to be completed during the workshop, to avoid compromising the calibration seed questions used to evaluate the 'accuracy' and 'informativeness' of the experts' judgments (given time and opportunity, the experts could have looked up the relevant answers from published papers).

On certain questions of interest for the Defra study, some significant or systematic differences emerged amongst the experts, and the elicitation process was repeated a second time, partly in order that it could be preceded by more extended discussion of the technical issues, but also to further widen the base of experts to include two academics. Eleven experts took part in the second workshop, comprising two owner's representatives (who are both Supervising Engineers), two academics, and seven consultants' staff (six Panel AR and one Supervising Engineer); conduct of the workshop was overseen by the independent facilitator.

PARAMETERS SELECTED FOR QUANTIFICATION

The primary objective was to obtain a separate view from that in the questionnaire on the rate of deterioration of embankment dams due to internal erosion, and thus inform the output from the research project in terms of recommendations of the frequency of surveillance.

One of the key issues was devising a model of internal erosion that could be quantified using both the elicitation and questionnaire. Such a model should ideally include the effect of time, the indicators that internal erosion is occurring (indicators), those factors that determine both the predisposition to internal erosion (intrinsic condition) and how events may progress at a particular dam (event trees). It proved impossible to devise one model that satisfied all these requirements, so three models were constructed, as presented in Brown & Gosden (2004). The questions were devised to quantify elements in each of these models, with the variables of most concern being summarized in Table 3, and issues to be addressed in devising the detailed text of the questions included in Table 4.

Table 3. Groups of variables selected for expert opinion elicitation

		No. of questions
1	Seed questions	11
2	Prevalence of leakage and internal erosion	16
3	Average leakage and erosion rates	4
4	Minimum detectable leakage rate, dam critical flow	5
5	Rate of deterioration i.e. how long from detection to failure	10
6	Contributory factors to rate of progression	14
7	Chance nodes in event tree; i.e. what are the likely proportions of possible types of behaviour?	14
	Total	74

Table 4. Issues in devising questions for expert opinion elicitation

Issue	Adopted
For which dam type(s) the question should be posed	The UK populations of puddle clay core, and homogenous dams. This was on the basis that the data in the BRE database shows that these are the most common types; together comprising 84 % of the UK embankment dam population.
To which dam(s) does the question apply?	Questions were generally posed to apply to the whole UK population of that type of dam.
Clarity of question	The question should be unambiguous. The draft questions were subject to external review by (non-dam) experts familiar with expert elicitation.
How many questions can be included	The first workshop had 11 calibration and 63 elicitation questions, as shown in Table 3. Although this is towards the upper limit of a number for one session, it was achieved, partly, by including a break in the elicitation session.
Content of seed questions	A minimum of 11 questions were required to calibrate the experts. There was some difficulty in finding suitable questions, i.e. those which covered the relevant subject area and for which the majority of experts would not know the answer.

In retrospect it has been realized that the term "vertical puddle clay core" actually describes three separate facets of a dam core, for example a dam which is homogenous in terms of material can have a puddle core (i.e. a core zone where the fill is placed by puddling). Although this issue was raised in discussion during the elicitation, the wording of the questions was not formally updated to reflect this need for precision.

RESULTS OF ELICITATION

Weighting of experts

Although in the results of the first workshop every expert had a non-zero weighting (i.e. contributed to the synthetic DM), it was decided for the second workshop that the weight of the synthetic DM should be allowed to increase towards a maximum, subject to the constraint that a majority of the group (*i.e.* for no less than six of the experts) must retain non-zero weights (see Figure 2 below for an example). This point was reached for a calibration power of 0.5, and a chi-squared significance level of 1%. The net effect of excluding the five lowest scoring experts is to raise the normalized relative weight of the synthetic DM to 0.44, from 0.15 for the first workshop (no non-zero weights). The six surviving (non-zero weighting) experts have weights ranging from 0.19 down to 0.02 (equivalent to a highest-to-lowest weight ratio of 9*x*). The synthetic DM would now have more than twice the weight of the best positively weighted individual expert, and 22*x* the weight of the lowest, positively weighted expert.

As a comparison with the weighting from the elicitation, based on performance with the known seed questions, a mutual weighting of colleagues in the group was carried out in the first workshop. There are some significant changes in ranking between the two, for example some experts scoring significantly less well on the performance-based measure than their colleagues might anticipate, while others do much better. This is not an uncommon pattern of ranking in groups of specialists of any discipline: some experts are well-regarded but tend to be strongly opinionated, while other more reflective individuals, who may be considered indecisive or diffident are, in fact, better estimators of uncertainty. In the present case, where the quantification of model parameter uncertainties is one of the main objectives, it is appropriate that the latter experts gain credit for their ability to judge these things well.

Output from process

The 5%, 50% and 95% estimates provided by each of the eleven experts were combined numerically in a computer code version of the classical model to provide a pooled uncertainty assessment for each query variable, using each individual's weight as derived from his calibration and informativeness on the known seed questions.

A typical result is shown in Figure 1 in the form of the experts' range graphs. Figure 2 illustrates both a question with significant variations between experts and also the effect on the synthetic DM results when its weight is allowed to increase by raising the significance level of the calibration test. Figure 3 shows a sequence of how the combined results of the elicitation for one item changed:

- between the first and second workshops,
- after the second workshop, when one outlying expert reconsidered his responses,
- when a change was made to the way in which the synthetic decision maker's effective score was constrained.

Features of note are the significant differences in widths of ranges between experts, and also the commonly wide ranges spanning the pooled 5% and 95% responses, reflecting the significant uncertainty in some of the parameters of interest. For some questions there is a failure of some experts' confidence limits to overlap with others, suggesting significant discrepancies of opinion. This is as illustrated on Figure 2, where the maximum number of experts who overlap at any one value is only four out of the total of eleven experts; additionally there are two groups of opinion about what the appropriate scaling of the value should be. One of the reasons for repeating the elicitation was that the results of the first workshop had produced some items where responses clustered in two disjoint groups in this way, representing 'high' and 'low' schools of thought. This effect had generally disappeared in the results of the second workshop, leaving only marginal instances, as shown on Figure 2.

In Figures 1 and 2, the 5% to 95% confidence spread of the synthetic decision maker spans the whole range of 50% estimates that are provided by the experts when each has a non-zero weight in the analysis. As a result, the DM encompasses the full extent of opinion but then, inevitably, exhibits a much wider confidence range than that of any one expert.

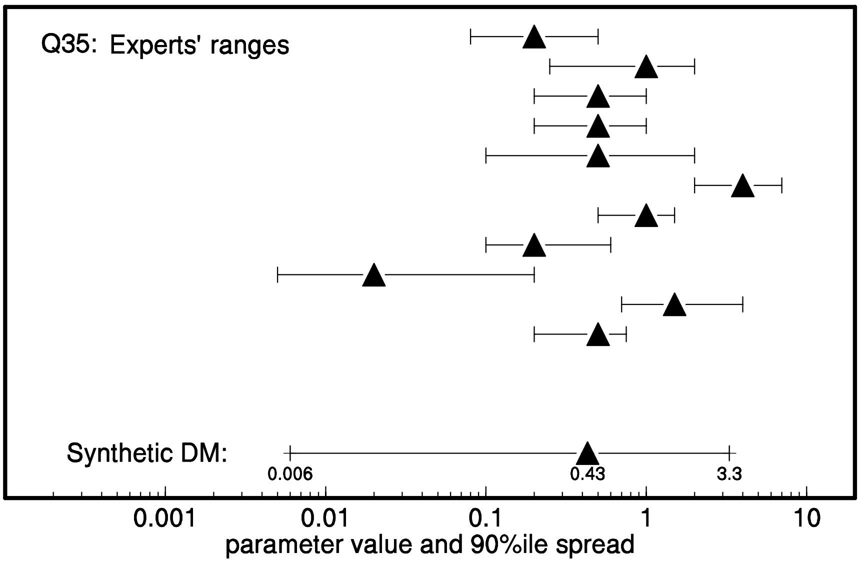

Figure 1. Typical range graph (Q35, median value for population of all UK embankment dams of dam critical flow i.e. uncontrolled erosion flow at which control of the reservoir has been lost)

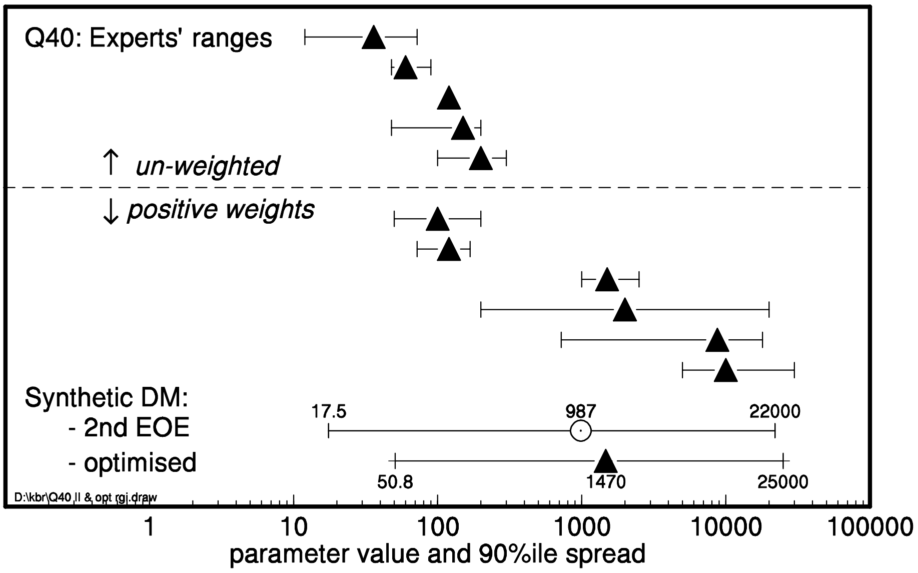

Figure 2. Range graph for Q40 (the time from detection to failure of puddle clay dams due to concentrated leak, in hours; for which only 10% of incidents are slower than this). Note for optimised DM the five lowest scoring experts, above the dashed line, are discounted – note their relatively high opinionation.

Steps can be taken to moderate this effect. If the synthetic decision maker is treated as a virtual expert, and included in the analysis, the calibration test significance level can be chosen so as to optimise the DM's distribution. While reducing the significance level enables all experts to receive positive weight, it does so at the expense of degrading the DM's calibration and entropy scores. Thus, an uncritical combination of expert assessments generally results in very large confidence bounds for the DM, as evinced in Figure 1. In the present case, the significance level was adjusted to the point at which there was still, overall, a majority of real experts with positive scores, as described earlier, thereby reducing the 'noise' of diverging opinions and improving the DM's calibration at the same time. Figure 2 illustrates how the DM's range is reduced slightly, and its 50% value more closely reflects the views the better-weighted experts; however, while some experts are discounted by this decision, similar views survive amongst those with positive weights, so such opinions remain represented in the elicitation.

It can be argued that, even though the DM's 5% - 95% range is typically larger than that of any individual, the spread is more representative of the proper scientific or engineering uncertainty for the variable in question. This is not implausible as some of the experts also present spreads in belief of similar magnitude.

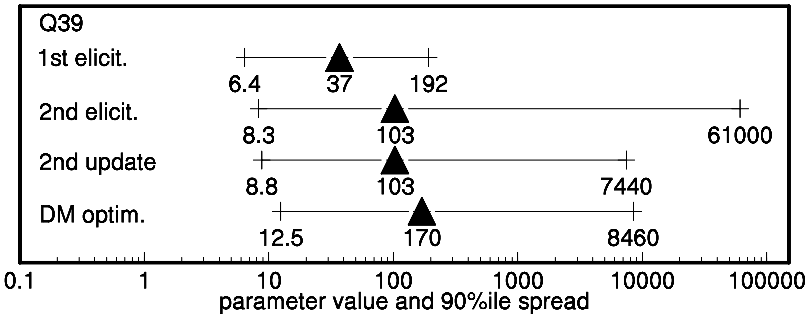

Figure 3. Example of changes between first and second elicitation

The way in which in which the synthetic decision maker's results changed through the various stages of the present elicitation process is illustrated in Figure 3. In this instance, the most marked change arose at the time of the second workshop, when technical issues were re-visited in detail and additional experts added to the panel. A few participants, who gave extreme or discordant values, were then given the opportunity to review their responses, resulting in the revised '2nd update' results. These outcomes were not greatly modified when the DM's weight was allowed to increase at the expense of a minority of the group ('DM optim.'), as just described, above.

Lessons learned

The elicitation process itself was new to all those who took part, and the key aspect that could be improved in future exercises of this kind is to increase ownership of the questions and issues by those taking part. This could be achieved by a longer workshop where the experts themselves assisted in setting the questions to be evaluated. Additionally, discussion could be stimulated by appointing protagonists to argue the case for extremes of possible responses (in some cases, it has been found effective to ask people holding opposing views to play 'devil's advocate', to argue the case for a particular position they themselves don't adhere to - this often reduces strongly-held dichotomies of opinion!).

ASSESSMENT OF RESULTS: ACCURACY AND PREDICTION

This section compares the elicitation responses with data available from elsewhere, and comments on the predictions made by the experts.

Questionnaire to UK dam industry

In parallel with the elicitation, a separate questionnaire was sent to 117 respondents, comprising all owners of more than 15 dams (20 number), a sample of 15 owners of one or two dams, all Panel AR Engineers (56 number), 10% of Supervising Engineers (24 number) and two research bodies. As well as questions relating to personal experience of internal erosion and opinion of the effectiveness of surveillance, requests were made for specific case histories of serious near miss incidents relating to internal erosion. This produced a total of 34 incidents from 19 respondents, and the data obtained are used here for comparison with the results of the elicitation exercise. It should be noted that these data were not available at the time of the first workshop, but a preliminary assessment was available by the time of the second.

Prevalence of leakage

The best estimate, from the elicitation, was that about 10% of puddle clay dams had ongoing steady leakage at each of the body of the dam, along an interface with appurtenant works and through the foundation, with 7% have leakage from the body of the dam into the foundation. Where leakage was occurring it was considered that ongoing internal erosion was occurring at about 10 to 17% of these locations. For homogenous dams steady leakage was judged as less likely (3 to 11% of dams, depending on location), with 7 to 17% of the leakages having ongoing internal erosion.

The questionnaire only provides data on serious progressive (deteriorating) internal erosion, which is likely to be less prevalent than steady ongoing erosion. This reported on average, for the period 1992-2002 three emergency drawdowns and ten precautionary drawdowns a year due to concern about internal erosion. This represents 0.2% and 0.5% of the stock of British embankment dams per year. These confirm that internal erosion is a serious threat.

Erosion and leakage rates

Figure 4 shows the results from three elicitation questions superimposed on a sensitivity study of how concentrated leakage might be expected to vary with crack width for a given crack height and length. The three points for each question represent 5, 50 and 95% uncertainty values. Flow in the crack is laminar up to 0.6mm, then turbulent. The experts' responses appear reasonable when compared with the sensitivity study.

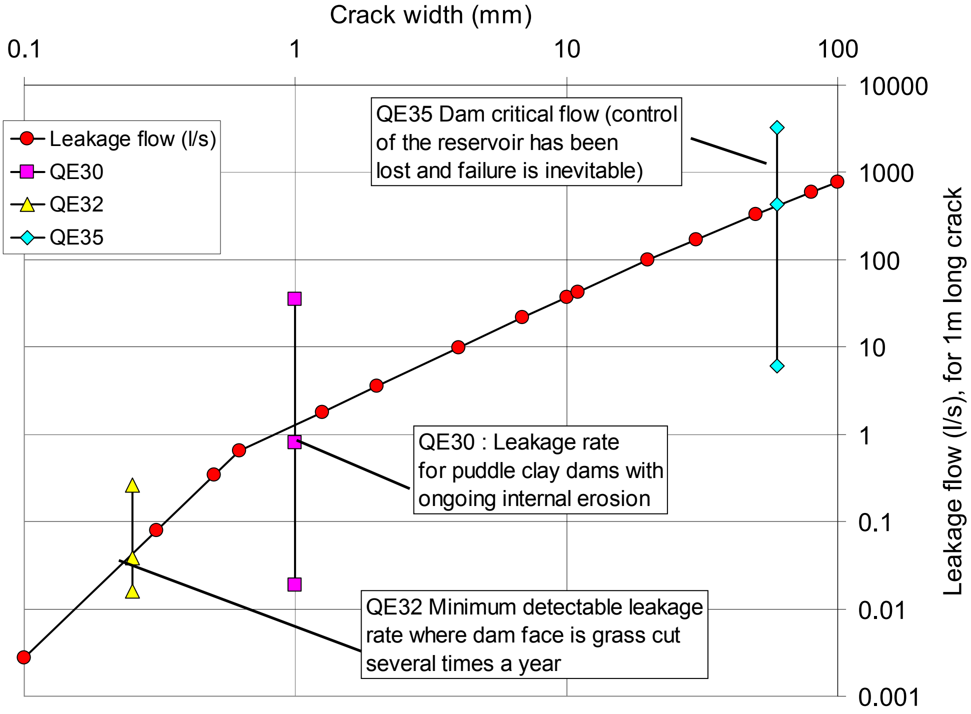

Figure 4. Sensitivity study of concentrated leakage flow to crack width (for flow through a 1m high 3m long crack under 10m head)

Rate of deterioration

Figure 5 shows the experts' opinion of the distribution of the time-to-failure for the whole population of UK puddle clay dams, if progressive internal erosion commenced at every dam, the time-to-failure being defined as that from the moment internal erosion was detected at a level of concern sufficient to call in an Inspecting Engineer to the time at which the dam critical flow rate was reached. Also shown on the figure is the distribution of the questionnaire respondent's opinions on how long before the dam would have failed in that incident, if there had been no intervention.

The significant range for the best estimate is noted, ranging from quicker than a day for 2% of dams to about 4 months for the slowest 2%. However, the response to the questionnaire suggests that the time to failure would have been much slower, with 75% of dams taking longer than 4 months. The significant uncertainty bands for the expert's opinion are also noted.

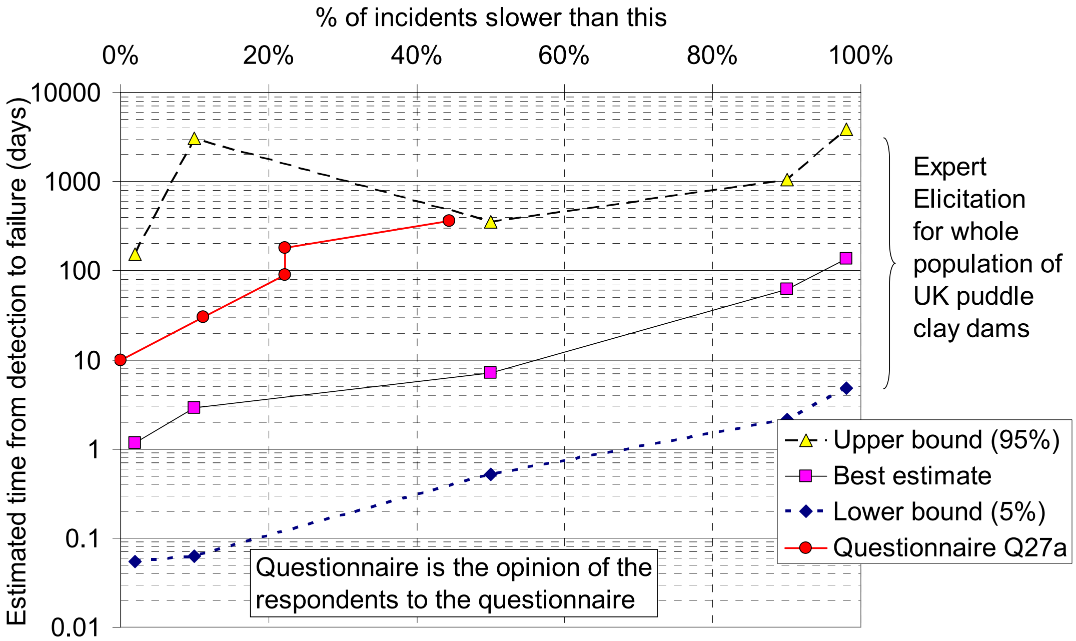

Figure 5. Distribution of time to failure for puddle clay dams

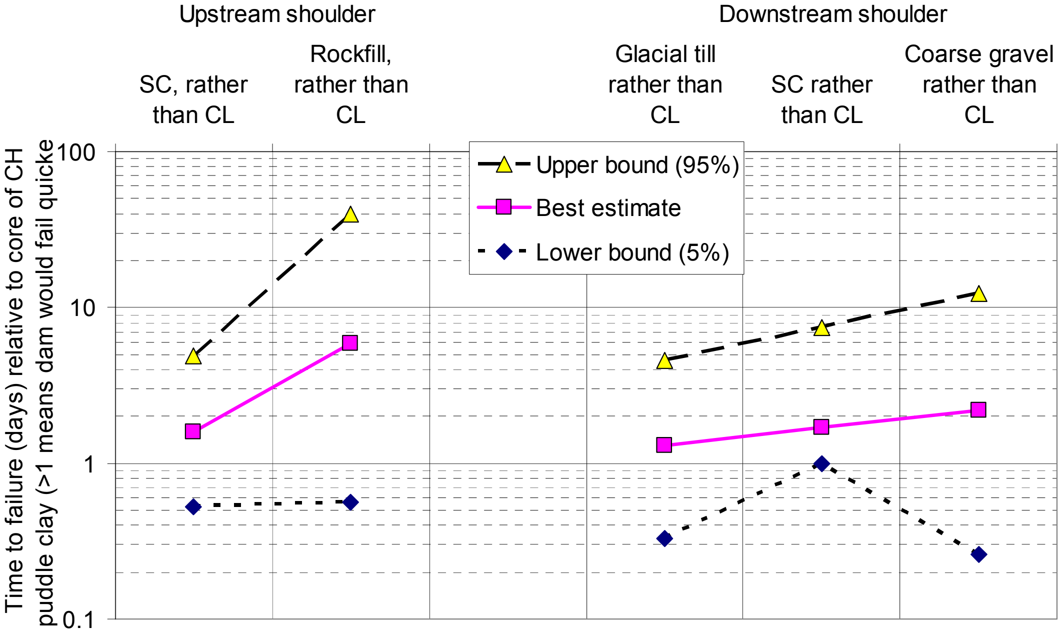

Figure 6. Effect of characteristics of dam shoulders on time to failure

Contributory factors to rate of erosion

The elicitation questions included the effect of factors such as the hydraulic gradient, the plasticity and degree of compaction of core material and properties of the shoulder materials on the time to failure. Typical output is shown in Figure 6. The expert opinion typically gave changes in rate of deterioration of up to 10; this may be low when compared to the ranges in rate of deterioration of several orders of magnitude.

DEBATABLE ISSUES

The understanding of internal erosion processes is still immature, with quantitative methods only available for limited elements. Tools that can help in either quantitatively capturing existing knowledge and experience, or in probing unexplored areas are therefore of value. The elicitation process set out by Cooke is of value in providing rational consensus, in that the opinions of the quiet reflective expert are considered, with appropriate weightings, just as much as those of more dominant personalities.

Elicitation has proved of value in making the wide spread of uncertainty explicit, and in capturing knowledge. The process adopted for this research contract did not fully explore the reasons for the wide discrepancy of results, but this could be pursued in future exercises. Debatable issues raised include:

a) most of the dam experts appear to give uncertainty bounds which are narrower than the true uncertainty, particularly where the uncertainty covers orders of magnitude - however, this trait has been found to be true of technical experts of all kinds;
b) the validity of questions which ask for the spread of a variable over the whole population of a particular dam type. It could be argued that for some of the dams the question is irrelevant, or inappropriate; however, to advance the knowledge of internal erosion further work is required at both a detailed level on specific dams and in understanding of the behaviour of groups of dams;
c) the validity of questions which simplify a complex problem down to focus on only one aspect of the problem, assuming “all other things being equal”. For issues governed by two (or more) important interdependent variables this may be an over-simplification.

Possible applications of the technique include research into parameters which cannot readily be quantified, for example floods with an annual probability of less than 10^{-4}/ annum. Additionally in increasingly litigious times the underlying structured basis of the method can provide a valuable record of the way a decision was reached, the impartiality of which could offer both a significant shield against personal liability to individual experts providing critical advice and a transparent decision process for major organisations.

CONCLUSIONS

Expert Opinion Elicitation, a technique first developed for the space industry, was one of the techniques used in an ongoing research contract for Defra to explore current knowledge of internal erosion. It provided a useful set of judgments and insights, including explicit confidence limits, broadly consistent with the findings from the questionnaire to the wider UK dam industry. Significant advantages of the technique are the encouragement which the procedure gives to all participants to express their true engineering beliefs (unbiased by peer pressure). In addition, the combined output from the procedure (the synthetic decision maker) generally provides values for the complete set of questions that are, overall, more coherent and closer to reality than those that would be obtained from any one individual expert, however good.

It is concluded that expert elicitation provides a valuable technique for quantifying those variables that cannot be determined by direct measurement, and for evaluating realistic likely spreads of scientific or engineering uncertainty on engineering parameters.

ACKNOWLEDGEMENTS

The work described in this paper was carried out as a research contract for Defra, who have given permission to publish this paper. However, the opinions expressed are solely those of the authors and do not necessarily reflect those of Defra.

REFERENCES

ASPINALL W. & COOKE R.M. 1998, Expert judgment and the Montserrat Volcano eruption. In: *Proc. 4th Intl. Conf. Prob. Safety Assessment and Management PSAM4*, New York, (eds. Ali Mosleh and Robert A. Bari), Vol.3, 2113-2118.

BROWN & GOSDEN. 2004, Outline Strategy for the management of internal erosion in embankment dams. *Dams & Reservoirs*. Vol 14 no 1

COOKE R.M. 1991, *Experts in uncertainty: Opinion and subjective probability in science*. Oxford. Oxford Univ Press 321pp

MEYER M.A. & BOOKER J.M. 1991, *Eliciting and Analysing Expert Judgement: A Practical Guide*. Rep NUREG/CR-5424. Nucl. Reg. Comm., Washington DC (republished 2001: ASA-SIAM Series in Statistics and Applied Probability, Philadelphia; 459pp)

O'HAGAN, A. 1998, Eliciting expert beliefs in substantial practical applications. *The Statistician*, 47 Part 1, pp21-35, Discussion 55-68.