Use of Expert Judgment Elicitation to Estimate Seismic Vulnerability of Selected Building Types

K.S. Jaiswal
U.S. Geological Survey/Synergetics Inc., Golden, CO

W.P. Aspinall
Dept. of Earth Sciences, University of Bristol UK

D. Perkins & D. Wald
U.S. Geological Survey Golden, CO

K.A. Porter
University of Colorado at Boulder, USA

SUMMARY:
Pooling engineering input on earthquake building vulnerability through an expert judgment elicitation process requires careful deliberation. This article provides an overview of expert judgment procedures including the Delphi approach and the Cooke performance-based method to estimate the seismic vulnerability of a building category. The objective is to pool professional judgments from an array of experts in order to quantify the seismic vulnerability of building types at different levels of earthquake shaking and, moreover, to quantify objectively the uncertainties associated with their estimates. We describe Cooke's method, which facilitates a weighted combination of individual probability judgments from multiple experts, since it offers the most rational and auditable approach in the context of seismic vulnerability estimation. We also discuss various activities related to an expert elicitation on seismic vulnerability that will be carried out in collaboration with other consortia partners of Global Earthquake Model's (GEM) Global Vulnerability Consortium (GVC) project.

Keywords: Expert Judgment, Vulnerability, Elicitation

1. INTRODUCTION

There is a rich body of literature on seismic vulnerability data and assessment procedures for a range of structure types; however, the challenges remain enormous for systematically assessing seismic vulnerability for all the dominant building types that are found worldwide. Efforts have been made in the past to compile and analyse structural vulnerability data and models for specific building types in different parts of the world. For example, in the U.S. these efforts include ATC-13 (1985), NIBS-FEMA (2009), ATC-58 project (ATC, 2012). Similarly, the World Housing Encyclopedia-Prompt Assessment of Global Earthquake for Response (WHE-PAGER) project resulted in compilation of both empirical and analytical vulnerability data of many of the most common building types around the world (http://pager.world-housing.net/).

Building on WHE-PAGER and other on-going efforts, the GEM Vulnerability Consortium (GVC) project now aims at developing the standards for *a global vulnerability database*, that practitioners and researchers can use and add to. In addition, the database and the guidelines will remain 'open' in the sense that users can access and/or modify them in a transparent, reproducible way (Porter et al., 2012). The GEM risk engine aims to operate three parallel vulnerability estimation approaches: empirical, analytical, and expert judgment, for global earthquake damage and loss computations. The empirical approach will draw on publicly available building damage and loss data for past earthquakes and

eventually develop seismic vulnerability functions for specific structure types defined using GEM's Basic Building Taxonomy (Brzev et al., 2012). Where data are lacking, the analytical approach will suffice to the extent possible. The analytical approach will be based on either nonlinear pseudostatic or stochastic nonlinear dynamic structural analysis, followed by a damage analysis using fragility functions that apply to individual building components (detailed at the level of UNIFORMAT-II or ATC-58 components), and a loss analysis where component repair costs are estimated given the component damage states. The applicability of specific analytical procedures will depend upon availability of various structural capacity and fragility related parameters, which may or may not be available for the building types that are common in different parts of the world.

Because of the limited data on past earthquake damage to derive empirical vulnerability functions, and because of the potential cost to derive analytical vulnerability functions, it will be necessary in some cases to rely on expert judgment to assess building vulnerability of certain building types.

In general, expert elicitation is considered to be useful in order to provide estimates regarding new, rare, complex, or poorly understood problems or phenomena (Lin and Bier, 2008). As applied by the GVC, the expert elicitation process consists of synthesizing a collective estimate of vulnerability by formalized group-pooling the individual judgments of multiple experts. Expert elicitation has been widely used in military intelligence (Dalkey 1969, Cooke, 1991), probabilistic risk analysis (Clemen and Winkler, 1999), global climate change (Titus and Narayanan, 1996), natural hazards, environmental and public health studies (Aspinall and Cooke, 1998; Tyshenko et al. 2011), and many other fields.

The success of any expert elicitation process is highly dependent upon how carefully the experts are chosen, whether the experts provide unbiased judgments, and, finally, how multiple opinions are reconciled or combined. The expert-opinion process necessitates identifying and recruiting qualified experts, training and conditioning them to avoid various heuristic biases, soliciting their opinions in a structured and efficient manner, reconciling divergent opinions, retrieving a best estimate, and – perhaps most importantly - quantifying the uncertainties associated with such an estimate. The next section provides a brief literature review on some of the commonly used elicitation procedures in the context of seismic vulnerability estimation.

2. DELPHI APPROACH

One of the most commonly used methods of expert elicitation and synthesizing expert judgments is the Delphi method (Dalkey, 1969). The method was developed for the U.S. Air Force in the 1950s by RAND Corporation, though it did not become publicly available for a decade or so (Cooke, 1991). The method was widely used and improvised in both military and non-military governmental agencies and in industrial applications for decision-making (Dalkey and Brown, 1971).

Technological forecasting and policy analysis are the two most common categories of applications for an expert elicitation process. The former category includes engineering-related applications for which the objective is to form a consensus based on opinions sought from selected subject matter experts (Ayyub, 2001). In the Delphi process, each expert provides an assessment, and the method allows each expert to view each other's assessments in an anonymous fashion, and gives an opportunity to revise their opinions or assessments. The process is repeated until a single acceptable assessment (consensus) is produced either agreeing upon a single assessment or through some acceptable combined mathematical aggregation. The anonymous revision is undertaken so that prevailing views, reputation, or personality factors of individuals should not affect the process.

The Delphi process has evolved over the last several decades as a structured method for eliciting expert judgment. Dalkey et al. (1970) offer a procedure for expert self-rating that tends to improve group estimates of uncertain quantities by removing experts who rate their confidence low. Tversky and Kahneman (1974) classified the types of heuristic biases that can pollute judgments of probability.

Spetzler and von Holstein (1972) addressed several of these biases and proposed procedures for eliciting expert judgment to avoid or minimize the biases.

3. DELPHI FOR SEISMIC VULNERABILITY ESTIMATION

Kustu et al. (1983) applied the Delphi process to develop damage probability matrices for 57 building classes. The Applied Technology Council's ATC-13 (1985) project applied a modified-Delphi process to obtain judgments from engineering professionals and to consolidate these judgments in order to understand and evaluate earthquake damageability aspects of the Californian building stock.

The ATC-13 was aimed at obtaining probabilities of damage defined in terms of a damage factor (the ratio of earthquake dollar damage divided by the facility replacement value) at various levels of ground motion intensity for selected facility classes found in California. The process was conducted through multiple rounds in which experts were asked to review their own answers from previous rounds relative to other experts, and if necessary, they were allowed to revisit/change their answers in subsequent rounds. Figure 1 shows typical responses obtained for low (5^{th} percentile), best (50^{th} percentile) and high (95^{th} percentile) estimate of damage factor after the Round Three Damage Factor Questionnaire for low-rise ductile reinforced concrete frame structures.

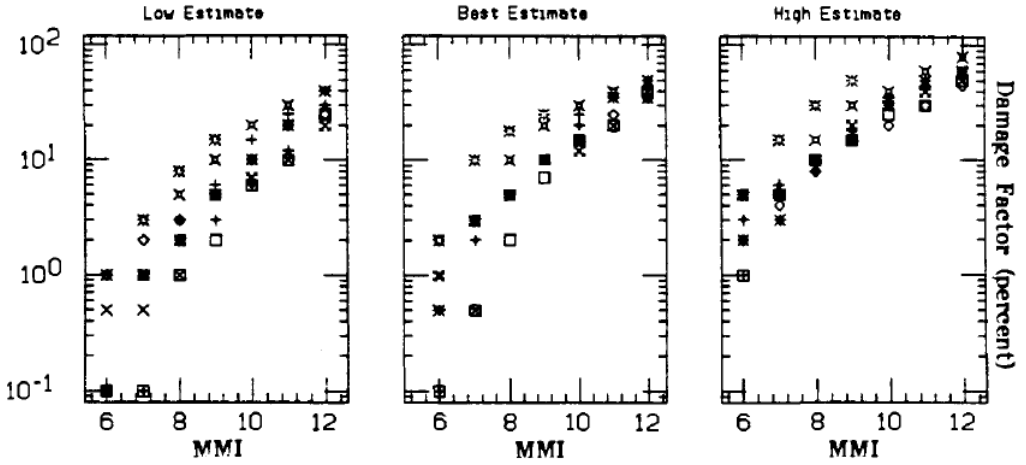

Figure 1. Low, best and high value of damage factor estimates by multiple experts (symbols shown correspond to answers from different experts) for low-rise moment resisting ductile concrete frame building at different Modified-Mercalli (MMI) shaking intensity levels (resketched from ATC-13, 1985).

The ATC survey resulted in a compilation of expert judgment data on damageability for a range of facility types at different levels of ground shaking intensity, in terms of Modified-Mercalli Intensity (MMI). In addition to the damage factor estimates, the ATC-13 questionnaire process also required experts to rate their experience and confidence levels in each of their assessments. The experience levels were ranked on a scale of 0 to 10, where 0 represented 'no experience', and 10 represented extensive experience in design and/or in post-earthquake investigation of a given facility class. Similarly, the experts' confidence levels in individual judgments they provided were also scaled between 0 and 10 where 0 represented 'no confidence' and 10 represented a statement of absolute certainty. The confidence levels basically represent each expert's own degree of belief in providing specific answers or giving specific responses to particular questions in the questionnaire. Dalkey et al. (1970) showed that such self-rating helps to identify the best qualified experts for each type of assessment.

Damage Responses from Eight Experts
Facility Class No. 1, MMI = IX

Experience q_i	Confidence C_{ijk}	Damage Factor Y_{ijk}	Exp.^{4} x Conf. $q_i^4\ C_{ijk}$	Weight w_{ijk}
9	9	5	59,049	.22
9	7	2	45,927	.17
8	7	8	28,672	.11
8	8	6	32,768	.12
7	8	10	19,208	.07
7	8	7	19,208	.07
9	7	1	45,927	.17
7	8	2	19,208	.07
			Σ = 269,967	

Figure 2. ATC-13 procedure showing typical process of estimating weights for each expert. (resketched from ATC-13 1985).

The ATC-13 team modified the self-rating aspect of the Delphi process to include all the experts (i.e., no censoring as in Dalkey et al., 1970) but weighted low-confidence judgments very low, considering both the experts' familiarity with the facility class of interest and the particular value being judged. In general, experts who self-rated very high in terms of confidence and subject-matter experience received higher weight as shown in Fig 2. The expert-specific weight was used to combine multiple judgments and eventually develop Damage Probability Matrices (DPMs) for 78 building classes and other assets common to California (ATC-13, 1985).

The Delphi process can often be slow and tedious and does not necessarily provide a rational basis for workable solutions. The Delphi process often aims to get 'position statements' from individual experts, iterating the responses one or several times. However, this sometimes engenders individual experts revising their inputs in the direction of supposed 'leading experts', rather than in the direction of strongest arguments (Aspinall, 2010). Moreover, in this context Burgman et al (2011) have shown recently that *a priori* perceptions of 'leading experts' often are not reliable.

4. COOKE'S APPROACH

Rather than the self-assessment of the Delphi process, Cooke's classical method is based on the principle of objective calibration scoring and hypothesis testing in classical statistics (Cooke, 1991). The process does exhibit certain key attributes such as empirical control (weights are determined based on performance on seed questions), reproducibility (scientific peers can review and reproduce any or all steps of the calculations), fairness (prior to performance measurements, all experts are treated equally), accountability (all assessments are recorded and can be checked by a reviewer), and neutrality (uses proper scoring rules for pooling probabilities, thus encourages experts to state their true belief/opinions). Details of these attributes and their significance in the elicitation process are discussed further in Cooke (1991).

Cooke's approach consists of estimating two separate scores and then multiplying them together to get the overall weight for each expert. The first score is a 'statistical accuracy' measure, called the *calibration score*, which is derived from the logarithmic scoring rule obtained in terms of distributional likelihood, and the second score is an *information score* that is based on a measure of the sharpness, that is, on the concentration of personal probability distributions in comparison to the uniform (or log-uniform) background distributions. In order to estimate these two types of scores, the experts are given a set of seed questions as a part of expert elicitation process. The seed questionnaire, or quiz, is generally conducted as a controlled exercise (without access to books, reference material, internet or group discussion) with the specific purpose of ascertaining an individual's ability to make judgments about uncertain values or parameters. Each seed question has a distinct and unambiguous answer which is not immediately available to the experts. and thus each expert is evaluated solely on his/her performance in answering these questions, which can be used to gauge expert's ability to enumerate his/her best estimate and ascribe uncertainty assignments to these estimates. The responses

on a set of seed questions are compiled and controlled by the examiner, and handled separately from the rest of the elicitation process in which responses are obtained for 'target questions' (i.e. those for which answers are unknown but desired by the problem owner). Responses to these 'seed' items are used to compute both calibration and information scores for each expert as discussed in the next subsection.

4.1. Principles of Weighting Scheme

The software implementation of Cooke's classical model offers 'equal', 'global' and 'item'-based weighting scheme for aggregating the experts' probability distributions (Cooke, 1991). In the 'equal' weighting approach, each expert is assigned an equal weight, which leads to estimating the mean response from the individual probability distribution of each expert for a given variable. The other two schemes are reward or performance-based schemes in which the experts who perform better in answering the seed questions get rewarded with higher weights, unlike the 'equal' weighting approach. Cooke's approach is the only approach in which real data (seed questions) are the basis for evaluating the experts and calculating weights (Clemen, 2008). Henceforth, Cooke's classical approach with its 'global' weighting scheme is referred to here as *Cooke's approach*.

4.2. Scoring System

As mentioned earlier, Cooke's approach uses a *calibration score* and an *information score* in order to evaluate the weight given to each expert's judgments. The mathematical background for estimation of the two scores is discussed below.

4.2.1. Calibration Score

The information score aims at providing a quantitative performance assessment value that can measure the expert's discrepancies against the realizations for the number of seed questions. Let us assume that for each seed question, an expert provides his/her best, low and high bound estimate to the test quantity. The low, best and high bound estimates generally refer to the three quantiles, namely 5^{th} percentile, 50^{th} percentile, and 95^{th} percentile estimate of the quantity. These percentiles help to split up the judged location of the variable into four intervals with 5%, 45%, 45% and 5% confidence levels in each interval respectively. It is considered that the expert is well calibrated if, over a set of seed items, the above intervals contain the expert's realizations in a statistically-balanced way (Cooke, 1991). In order to find the calibration score, the process estimates the proportion of times the variables' true values lie in each of the four intervals – ideally, these proportions should reflect the 5%, 45%, 45% and 5% confidence splits, mentioned above. Deviations from this support spread are penalized according to the degree of mismatch, and hence the corresponding calibration score is down-graded.

The 5^{th} and 95^{th} percentile quantities are not necessarily symmetric around the 50^{th} percentile estimate value. By introducing a small identical overshoot, say 10% on both sides (Cooke, 1991), it is possible to develop a background distribution so that the nominal 100% range for the quantity in question encloses all the experts' quantiles and the relevant realization. This provides approximate representations of the tails of the distributions beyond 5^{th} and 95^{th} percentile quantities. Because the added intrinsic range depends on the assessment of all experts, the information score of a given expert can change slightly when experts are added or removed from the process. The background distribution obtained in this way is associated with the expert's assessments for each target variable (i.e., the expert's density for that variable), which is ultimately referred to here as expert's informativeness.

Cooke (1991) defines the term $I(s, p)$ to indicate the relative information of s with respect to p measured for a certain number of chosen intervals of s and p, where p_i represents background reference density function and s_i represents the sample distribution obtained from the expert for each seed question at four intervals respectively.

$$I(s, p) = \sum_{i=1}^{4} s_i \ln\left(\frac{s_i}{p_i}\right) \tag{4.1}$$

The information term $I(s, p)$ is always nonnegative, and a continuous version of information $I(s(x), p(x))$ for certain low (l) and high (h) bounds of the distribution is given as

$$I(s, p) = \int_{l}^{h} \ln\left(\frac{s(x)}{p(x)}\right) s(x)\, dx \tag{4.2}$$

For total n seed variables, following Cooke (1991) and Tyshenko et al. (2011), the quantity $2 \times n \times I(s,p)$ is chi-squared distributed with 3 degrees of freedom. The calibration score $\Theta(e)$ of expert e is the p value of the statistical test of the hypothesis H_e that the expert is well calibrated and it is given as

$$\Theta(e) = P_r\left[\chi_3^2 > \chi^2(e) \mid H_e\right] \tag{4.3}$$

The quantity $\chi^2(e)$ is estimated as

$$\chi^2(e) = 2 \times n \times \sum_{i=1}^{4} s_i(e) \times \ln\left(\frac{s_i(e)}{p_i}\right) \tag{4.4}$$

4.2.2. Information Score

The information score or entropy represents the degree with which the elicited distribution is concentrated with respect to some background distribution. For each seed variable j one can use a uniform (or log-uniform) distribution as a background distribution [l, h]. The quantities l and h are estimated using the same small intrinsic overshoot, as discussed earlier.

The information score $\Lambda(e)$ for expert e can be estimated by averaging the relative information $I(s(x), p(x))$ from Eq 4.2 for *all* the seed variables and is given as

$$\Lambda(e) = \frac{1}{n} \sum_{j=1}^{n} I(s_{j,e}, p_j) \tag{4.5}$$

The total weight awarded to each expert thus represents a combined measure of an individual's calibration score $\Theta(e)$ and his/her information score $\Lambda(e)$ and it is given as

$$W_\alpha(e) := \chi_{(\alpha,\infty)}(\Theta(e)) \times \Theta(e) \times \Lambda(e) \tag{4.6}$$

The term α is to be determined by maximizing the two scores. The characteristic function $\chi_{(\alpha,\infty)}(\Theta(e))$ gives zero weights to the judgment which have a p value less than α. Cooke's method rewards experts with high calibration scores based on their statistical accuracy coupled with good informativeness; however, it penalizes heavily when these judgments are outside the α thresholds. The scoring rules can also be used for training purposes or adjusting the assessments as discussed in O'Hagan et al (2006).

4.3. Combining Multiple Judgments

Combining calibration and information scores results in an estimate of weight for each expert. These weights are then used to combine the judgments on target items from different experts, according to their weights. In Fig. 3, Aspinall (2008) demonstrates a schematic of Cooke's process to calibrate experts' responses to the seed questions at a given quantile to produce performance-based weights. These weights are then used to pool the experts' judgments for the corresponding quantiles of the target question. The total weight estimated for each expert using Cooke's approach is *global* in the sense that it is obtained from the same seed question for all the experts and that the same weights apply to all "unobserved" (or target) variables. The weights are dynamic since with each new observation or addition of new experts in the process, the weights can be recomputed.

4.4. Applications

Cooke's approach has been applied in different study areas in the past (e.g., a volcanic eruption crisis

by Aspinall and Cooke 1998, the internal erosion process in dams by Brown and Aspinall, 2004). It has been shown that the performance-based combination of expert judgments often yields more informative and statistically accurate results than the 'equal weight' or 'consensus' or 'best expert' approach for a combination of expert distribution (Cooke and Goossens, 2000; Cooke, 2004; Goosens et al., 1998). The method gives a positive incentive to experts to report their probability beliefs impartially and honestly. Fig. 4 shows a typical application of Cooke's approach to estimate uncertainties related to the risk of disease transmission through different causative factors. Responses obtained for specific target questions indicated that the mean incubation period for a person infected with secondary variant Creutzfeldt-Jakob disease (vCJD) by blood transfusion ranged from 2 to 37 years with best estimate to be 4.6 years based on performance-based weights (reported values rounded to remove spurious precision). However, when all the 10 experts were weighted equally, the best estimate value of mean incubation period increased to 9 years with low and high estimate range increased to 2 to 55 years, as shown in Fig 4. Data on this infection pathway are very sparse because, according to the authors, only a small number of vCJD cases are known to have been transmitted through blood transfusion, hence the recourse to expert elicitation.

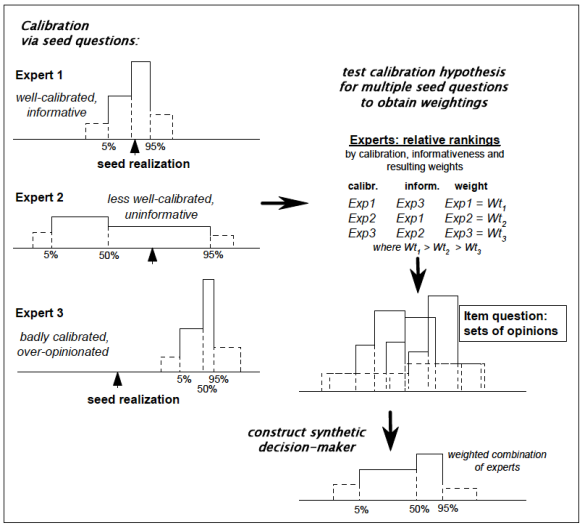

Figure 3. Schematic illustration of typical expert judgment traits, and how these feed into Cooke's procedure to evaluate individual expert weights for pooling responses on seed questions (Aspinall 2008).

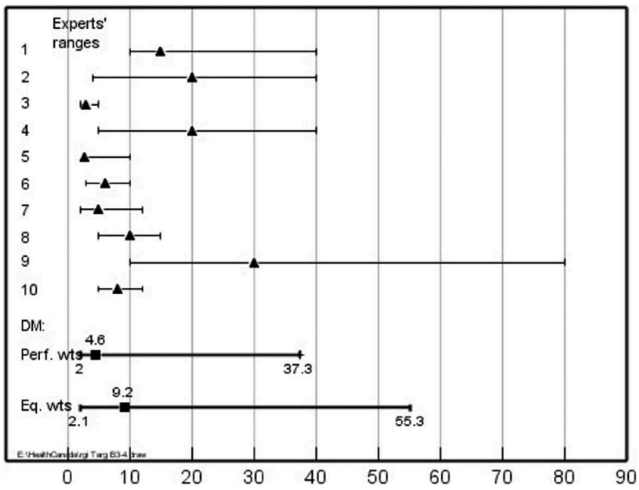

Figure 4. Top portion of the plot shows the responses obtained from 10 experts for a specific target question "What is the mean incubation period in years for a secondary variant Creutzfeldt-Jakob disease (vCJD) infected human by transfusion?". Bottom portion 'DM' shows the application of Cooke's procedure to estimate performance-based weights and equal weights solutions (after Tyshenko et al., 2011).

4.5. Overconfidence

In most expert elicitation studies, the 90% probability interval estimates provided by experts are expected to contain the true values of the corresponding quantities. A study conducted by Russo and Schoemaker (1992) found that only 40-60% of the total interval estimates provided by experts contained the true value. When the confidence interval estimated by experts is too narrow, it could be associated with the tendency of each expert to be overconfident in specific assessments. Lin and Bier (2006) provide a detailed review on the topic of overconfidence. Their study highlighted two things. (1) The calibration score varies among between different fields of research. For instance, option trading, radioactive deposition, and building temperature studies showed experts receiving notably higher calibration scores compared to some other study areas, like space debris, soil transfer, and movable barriers. (2) The calibration score may also vary among the types of seed questions chosen within a specific field, for example, performance differences were manifest for certain questions within a dike-ring research study (Lin and Bier 2006). Thus care is needed in the selection of seed questions, and they should be tested for coherency as performance-based metrics.

5. SEISMIC VULNERABILITY AND FRAGILITY ESTIMATION

The GVC project aims to develop seismic vulnerability estimation guidelines and functions for application within the Global Earthquake Model engine (Porter et al. 2012). In this project, we want to explore the application of Cooke's method in combining multiple expert judgments on seismic vulnerability. Before discussing the proposed activities, we provide a brief background on the topic of seismic vulnerability.

As used here, the seismic fragility function is represented by a curve in *x*-*y* space where damage or loss is measured on the *y*-axis (quantified in terms of exceedance probability of a specified damage state, such as collapse) and the *x*-axis measures the input *intensity measure* (IM) (quantified in terms of a measure of ground motion such as peak ground acceleration [PGA], peak ground velocity [PGV], 5%-damped elastic spectral acceleration response at some fixed period of vibration T, or macroseismic shaking intensity measure such as MMI or EMS-98. The curve can be fit via regression analysis of empirical or analytical observations, resulting in a mean or median seismic vulnerability function, along with a fixed or variable residual uncertainty. Sometimes the fragility function takes the form of a parametric distribution that the analyst imposes on the data or experts' judgment. A common choice is the lognormal cumulative distribution function (CDF, see, e.g., NIBS-FEMA 2009). The CDF can be interpreted as the distribution of the uncertain capacity of the specified component, such as the structural part of the building, to resist a specified damage state such as collapse, where the capacity is quantified in terms of the estimated structural response. Details of seismic vulnerability estimation procedures in the context of the GVC project are discussed in Porter et al. (2012).

6. PROPOSED ACTIVITIES RELATED TO GVC EXPERT ELICITATION

We plan to conduct an expert elicitation workshop in late 2012 to convene selected domain experts from different parts of the world and perform expert elicitations on seismic vulnerability measures of pre-selected building types. Prior to the actual elicitation process, experts will be provided with a set of guideline documents and a pre-survey questionnaire in which they will be asked to state their topic of interest, professional and research experience with specific building types, post-earthquake reconnaissance experience, and others. In addition, experts will be given a choice of specific building types and of ground motion intensity measures with which they might be comfortable in providing their seismic vulnerability assessments. Based on the pre-survey questionnaire, the technical manager of this project (the first author, who is also a Co-PI of GVC project) working closely with the Project Steering Committee, will select the domain experts by specific building types, and then choose specific ground motion intensity measures for that building type. The pre-survey questionnaire process will thus help getting all the topic experts on board, and bring clarity to the scope and objective of the actual elicitation. It will also allow them to express their feelings and concerns, and empower them via the choice of which specific ground motion or damage/loss measure parameter they are most

comfortable assessing.

Each building type specific expert will be asked to answer 10 to 12 seed questions, for which the technical manager knows the values. Individual experts are not expected to know these values precisely but should be able to make good judgment calls in terms of an encompassing confidence range. The seed questions need to be chosen in such a way that there is no ambiguity in the phrasing of a question or in the interpretation of their answers. The GVC partners will help define the specific topic areas within the subject matter and also construct some of the seed questions. Even so, only the examiner of the project will know the exact phrasing and answers. In addition, a separate set of target questions (10 to 12) will be prepared for which answers are not knowable and for which elicitation is to be carried out. The answers obtained for the target questions ultimately will be combined by the expert pooling, and used to generate seismic vulnerability functions. As a check, these performance-based outcomes will be compared with results obtained by allocating equal weights to all experts. Overall, 15 to 20 experts will be engaged in the elicitation process. The target questions will be allocated according to specific building types and thus only the experts who specialize in those specific structure types will provide their judgments. However, the seed (or calibration) questions will remain the same for all experts who participate in the elicitation process.

7. SUMMARY

The central objective of the GVC project is to systematically characterize the global building stock in terms of structural systems, and to develop structural vulnerability models and guidelines for empirical, analytical and expert judgment-based approaches. The expert judgment-based approach can provide a useful interim working option to generate seismic vulnerability functions for certain building types for which both empirical data and analytical models are missing. The process should help rationalize differing views within the engineering community and retrieve an optimal estimate and a credible interval enclosing the true seismic vulnerability of each given building type. Most important, the goal should be to rationally quantify the uncertainty rather than removing it from the decision process (Aspinall 2010).

The consensus-based Delphi method has been used to generate seismic vulnerability functions in the past. Our goal here is to use the performance measure-based Cooke's approach in order to elicit expert judgments for global building types for the first time. In Cooke's approach, expert's weights are determined from responses to a number of seed questions. Experts' calibration scores indicate the probability that any divergence between their personal distribution estimates and the corresponding distribution of observed values of seed variables might have arisen by chance. In a statistical sense, a low calibration score (near zero) indicates that an expert's evaluations could be inconsistent with actuality, and his or her technical opinions should be discounted to some extent. But, at the same time, and more constructively, the most capable experts are identified and their judgments given appropriate positive weights to provide an optimal decision – only the Cooke procedure empirically determines which experts warrant heavier weights; reputation and publication records are, sadly, quite poor indicators. Thus the weighted pooling process generally produces uncertainty spreads that are narrower than the 'democratic/traditional' (equal weights) pooling approaches, but wider than those provided by single, often over-confident, experts.

ACKNOWLEDGEMENT

This research study was funded by the GEM Foundation, Pavia, Italy as a part of GEMGVC project. WPA is supported in part by a European Research Council grant to Prof. RSJ Sparks FRS: VOLDIES Dynamics of Volcanoes and their Impact on the Environment and Society

REFERENCES

Aspinall, W. (2008), Expert Judgment Elicitation using the Classical Model and EXCALIBUR, Briefing notes for Seventh Session of the Statistics and Risk Assessment Section's International Expert Advisory Group on Risk Modeling.

Aspinall, W. (2010). A route to more tractable expert advice *Nature* **463**: 294-295.

Aspinall, W. and Cooke, R.M. (1998) Expert judgment and the montserrat Volcano eruption. In: Mosleh, Ali & Bari, Robert A. (eds.) Proc. Of the 4th International Conference on Probabilistic Safety Assessment and Management PSAM4, September 13-18, 1998, New York, USA.

(ATC) Applied Technology Council (1985). ATC-13, Earthquake Damage Evaluation Data for California, Redwood City, CA, 492 pp.

(ATC) Applied Technology Council (2012). ATC-58, Seismic Performance Assessment of Buildings, Volume 1-Methodology, 100% draft, Redwood City, CA.

Ayyub, B. (2001). A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities. *IWR Report 01-R-01 prepared for U.S. Army Corps of Engineers, Institute of Water Resources, Alexandria, VA*.

Brown, A. and Aspinall, W. (2004). Use of expert opinion elicitation to quantify the internal erosion process in dams, Proc. Of 13th British Dams Society (BDS) Conference, University of Kent, Canterbury, June 22-26, 2004.

Brzev, S., Scawthorn, C., Charleson, A.W., and Jaiswal, K. (2012) GEM Basic Building Taxonomy:V1.0. GEM Technical Report, GEM Foundation, Pavia, Italy.

Burgman MA, McBride M, Ashton R, Speirs-Bridge A, Flander L, et al. (2011) Expert status and performance. PLoS ONE 6(7): e22998. doi:10.1371/journal.pone.0022998

Clemen, R. T. (2008). Comments on Cooke's classical method. *Reliability Engineering and System Safety* **93**, 760-765.

Clemen, R. T. and Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis* **19**(2):187-203.

Cooke, R. M. (1991). Experts in Uncertainty- Opinion and Subjective Probability in Science. Environmental Ethics and Science Policy Series. Oxford University Press, New York 10016. ISBN 0-19-506465-8.

Cooke, R.M. and Goossens, L.H.J. (2000). Procedures Guide for Structured Expert Judgement. Report EUR 18820, Brussels-Luxembourg.

Cooke, R.M. (2004). The anatomy of the Squizzle-the role operational definitions in science, Reliability Engineering and System Safety, **85**: 313-319.

Dalkey, N. (1969) The Delphi Method: An experimental study of group opinion A report prepared for United States Air Force Project Rand RM-5888-PR, 79 p.

Dalkey, N., Brown, B., and Cochran, S. (1970). Use of self-ratings to improve group estimates. *Technological Forecasting*, **1**:283-291.

Dalkey, N. and Brown, B. (1971). Comparison of group judgment techniques with short-range predictions and almanac questions, Report prepared for Advanced Research Projects Agency, RAND Corp. R-678, pp. 34.

Goossens, L., Cooke, R., and Kraan, B. (1998). Evaluation of weighting schemes for expert judgment studies. In: Mosleh, Ali & Bari, Robert A. (eds.) Proc. Of the 4th International Conference on Probabilistic Safety Assessment and Management PSAM4, September 13-18, 1998, New York, USA.

Kustu, O., Miller, D. D., and Scholl, R. E. (1983). A computerized method for predicting earthquake losses in urban areas. Report by URS/John A. Blume & Associates JAB-99-111, San Francisco, California.

Lin, Shi-Woei and Vicki M. *Bier* (2008). A Study of Expert *Overconfidence*. *Reliability Engineering and System Safety* **93:5**, 711–721.

(NIBS and FEMA) National Institute of Building Sciences and Federal Emergency Management Agency, (2009). Multi-hazard Loss Estimation Methodology, Earthquake Model, HAZUS®MH MR4 Technical Manual, Federal Emergency Management Agency, Washington, DC

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T., (2006). Uncertain Judgements: Eliciting Experts' Probabilities, John Wiley Ltd, England.

Porter K.A.et al. (2012). Global vulnerability estimation methods for the Global Earthquake Model. *Proc. 15th World Conf. on Earthq. Engineering*, 24-28 Sept 2012, Lisbon, Portugal.

Russo, J. E. and Schoemaker, P. J. H., (1992). Managing overconfidence. *Sloan Management Review* 33:7–17.

Spetzler, C.S. and C.S.S. von Holstein, (1972). Probability encoding in decision analysis. *Proc. ORSA-TIMS-AIEE 1972 Joint National Meeting*, Atlantic City, NJ, 8-10 Nov 1972, reproduced in Howard, R.A., and Matheson, J.E., 1989, Readings on the Principles and Applications of Decision Analysis, Strategic Decisions Group, Menlo Park, CA

Titus, J.G., and Narayanan, V. (1996). The risk of sea level rise: A Delphic Monte Carlo Analysis in which twenty researchers specify subjective probability distributions for model coefficients within their respective areas of expertise. *Climatic Change* 33: 151-212.

Tversky, A. and Kahneman, D., (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185, pp. 1124-1131

Tyshenko, M.G., ElSaadany, S., Oraby, T., Darshan, S., Aspinall, W., Cooke, R., Catford, A., and Krewski, D. (2011). Expert elicitation for the judgment of Prion disease risk uncertainties. *Journal of Toxicology and Environmental Health*, Part A, **74**: 261-285.