Bulletin of the Seismological Society of America, Vol. 103, No. 5, pp. 2862–2874, October 2013, doi: 10.1785/0120130026

Ⓔ

An Interactive Tool for the Elicitation of Subjective Probabilities in Probabilistic Seismic-Hazard Analysis

by Antonia K. Runge, Frank Scherbaum, Andrew Curtis, and Carsten Riggelsen

Abstract In probabilistic seismic-hazard analysis, epistemic uncertainties are commonly treated within a logic-tree framework in which the branch weights express the degree of belief of an expert in a set of models. For the calculation of the distribution of hazard curves, these branch weights represent subjective probabilities. A major challenge for experts is to provide logically consistent weight estimates (in the sense of Kolmogorovs axioms), to be aware of the multitude of heuristics, and to minimize the biases which affect human judgment under uncertainty. We introduce a platform-independent, interactive program enabling us to quantify, elicit, and transfer expert knowledge into a set of subjective probabilities by applying experimental design theory, following the approach of Curtis and Wood (2004). Instead of determining the set of probabilities for all models in a single step, the computer-driven elicitation process is performed as a sequence of evaluations of relative weights for small subsets of models. From these, the probabilities for the whole model set are determined as a solution of an optimization problem. The result of this process is a set of logically consistent probabilities together with a measure of confidence determined from the amount of conflicting information which is provided by the expert during the relative weighting process. We experiment with different scenarios simulating likely expert behaviors in the context of knowledge elicitation and show the impact this has on the results. The overall aim is to provide a smart elicitation technique, and our findings serve as a guide for practical applications.

Online Material: Interactive software for the elicitation of expert knowledge.

Introduction

The quantification of uncertainty is a topic of interest in a range of disciplines. Many schemes are used for its classification (e.g., Toro *et al.*, 1997; Knol *et al.*, 2010), but two types of uncertainties are almost always distinguished: (1) uncertainty that is primarily caused by incomplete knowledge (epistemic uncertainty) and (2) uncertainty due to intrinsic variability of a process or a system itself (aleatory uncertainty). This article focuses on epistemic uncertainties regarding the applicability of a set of competing models. In probabilistic seismic-hazard analysis, for example, it became a *de facto* standard approach to spread competing models (e.g., different seismic source-zone models or ground-motion models) on alternative logic-tree branches (Bommer *et al.*, 2005). This implies that the corresponding branch weights are supposed to capture the degree of belief (DOB) of one (or several) expert(s) in the models' applicability to a particular site of interest. However, despite the widespread use of logic trees, there is little guidance on how best to set up a tree and assign weights to its branches (Bommer *et al.*, 2005; Bommer and Scherbaum, 2008). Furthermore, there is ambiguity in the meaning of branch weights, that is, in what they are assumed to represent (Scherbaum and Kuehn, 2011; Bommer, 2012).

Attempts to find "objective" logic-tree branch weights led to an approach that defines weights based on the performance of the models with respect to a set of numerical quality measures (Bommer *et al.*, 2005). This "grading matrix" approach, however, is no longer recommended by the original authors (Bommer, 2012), because with an increasing number of models, normalized weights (which sum up to unity) will lead to an apparent insensitivity of hazard curves to the branch weights (Scherbaum and Kuehn, 2011).

When using information-poor data sets, it is impossible to refrain from the explicit use of expert knowledge. Beyond the implicit subjectivity in forming hypotheses or designing models, it is necessary to apply transparent transforms of expert beliefs into explicit quantities. Kolmogorov axioms offer principled rules of handling expert beliefs represented as probabilities. Additionally, probability theory allows us to adhere to a Bayesian statistical interpretation. However, it is

well known to cognitive psychologists that human limitations of memory often result in subconscious deployment of heuristics (rules of thumb) used to estimate subjective probabilities which may violate Kolmogorov's axioms. Although heuristics are often efficient from certain points of view, they commonly lead to a variety of biases, such as overconfidence, anchoring and adjustment, availability, and motivational bias, which may distort the estimates of the underlying expert beliefs. Details on biases can be found in Tversky and Kahneman (1974), Baddeley *et al.* (2004), O'Hagan *et al.* (2006), and other scientific literature on elicitation theory (a discipline using findings from psychology and statistics).

Previous Work

A variety of elicitation methods have been proposed that aim to minimize people's need to resort to heuristics, which is also this article's purpose. O'Hagan *et al.* (2006) describe several De Finetti game analogies which assist the expert in assessing a single probability. These methods mainly ask the uncertainty of the event in question to be compared with an event upon which the expert might gamble. This may be supported visually by displaying an urn filled with marbles in such a way that the probability in question is represented (e.g., De Finetti's Game written by Scherbaum [2013]; see Data and Resources) or by showing a pie chart with corresponding proportions. However, all gambling methods are designed for binary events (assessing a single probability) and are therefore difficult to apply to the quantification of a whole set of subjective probabilities. In general, experts may easily become overwhelmed if they have to evaluate multiple competing options simultaneously, which most often leads to biased assessments.

We favor the division of such problems into a number of subproblems by requesting option comparisons of only small subsets (≤ 5) of all possible options and estimating only subsequently the whole set of DOB values. Saaty (1980) proposed a technique for pairwise data analysis as a part of his Analytical Hierarchy Process (Saaty, 1980). The latter is a decision analytical tool, for which an expert is required to provide relative comparisons by evaluating only two options at a time. This can be made in a qualitative (e.g., equal, better, or worse) or in a quantitative (using numerical scales) evaluation. In both cases, the expert's comparisons are quantified using a matrix algebraic approach resulting in a set of n numerical DOB values for the n options in question. Specifically, these weights are the entries of the normalized eigenvector for the maximum eigenvalue of the $n \times n$ matrix, which is filled with all comparisons in its upper triangular portion. Diagonals are set to 1s, and the lower portion of the matrix is filled with the reciprocal values of the upper portion. Saaty (1980) also derived an index of consistency of the expert's evaluations using matrix theory (Meyer and Booker, 2001). However, this method requires experts to evaluate all possible relative comparisons, which can be unfeasible in practice. In general, for n options there are $(n-1)+(n-2)+\ldots+1$ possible comparisons.

Curtis and Wood (2004) suggested an alternative, optimized approach. Again, experts have to provide relative comparisons to judge a set of competing options, in particular a set of models, but it is no longer necessary to evaluate all possible comparisons. The expert's DOB values are estimated via the method of least squares (a technique used for regression analysis), which is applied to the set of elicited pairwise comparisons. Experimental design theory is used to optimize the process of elicitation in real time. That is, the total of all previously elicited information determines the design of all future questionnaires. Consistency of the expert's statements is also measured and used to estimate a measure of confidence in the expert assessments.

Presented Method

We implemented an elicitation tool, that builds upon the method of Curtis and Wood (2004), which will be introduced in the Interactive Elicitation Tool section. This platform-independent, interactive program is aimed at assisting experts visually during the process of elicitation. Both the distribution of the expert's DOB in a set of models, or more generally in a set of options, and the consistency of the expert's own statements are shown graphically at the end of each part of an elicitation session. Thus, detected inconsistencies may counteract expert overconfidence. Compared with the method of Curtis and Wood, who collected all necessary model comparisons via e-mail questionnaires, the usage of our interactive program will accelerate the whole elicitation process and even allows experts to self-elicit themselves in order to test their consistency.

This article goes on to test the performance of the approach of Curtis and Wood by simulating four different scenarios of expert judgment, which are applied to their method. This allows us to analyze thousands of simulated elicitation sessions, whereby it becomes possible to statistically identify expected DOB estimations, average consistency values, and the method's rate of convergence, hence the average expected time effort in all four scenarios. The results provide practical guidance for a real application, that is, a real elicitation of a set of DOB values, which for example may be used within a logic-tree framework or as a subjective prior in a Bayesian context.

Interactive Elicitation Tool

This section will introduce our approach, which includes the method proposed by Curtis and Wood (2004) and covers issues that arose during the implementation of their approach.

In order to apply the concept of subjective probability, we assume that it is possible to elicit DOBs from experts and that probabilities are their correct representation. Beliefs of experts who are subject to social interactions or to new observations may vary over even very short timescales (Polson

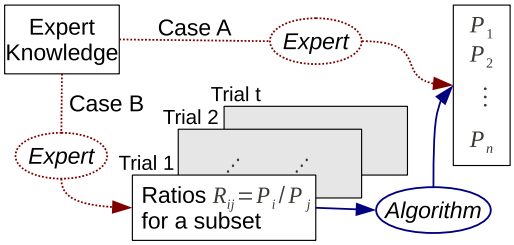

Figure 1. Representation of two different ways to elicit subjective probabilities. Case A (standard approach): the expert directly quantifies her/his DOB P_i in model M_i. Case B (present approach): the expert provides probability ratios for pairs of models, which represent her/his DOB in one model relative to another model. The color version of this figure is available only in the electronic edition.

and Curtis, 2010; Curtis, 2012). Therefore, an elicited quantity can only be interpreted as a single sample in time of a continuously varying belief distribution. A sample of the expert's belief distribution across n competing models can be obtained via two different ways. The expert can simply be asked directly for n numbers between zero and one representing absolute probabilities (case A in Fig. 1). Alternatively, the expert can provide a sufficiently large set of model comparisons in the form of probability ratios or relative probabilities from which the set of n absolute probabilities will be derived (case B in Fig. 1). The first method is straightforward, fast, and appropriate when n is small (≤ 5) but becomes unfeasible when many models have to be judged at one time. The advantage of the second method is that relative judgment is much more intuitive and is thus an easier task. Experts only have to concentrate on a small subset (≤ 5) of models at one time, and the consistency of their statements can be checked easily.

Curtis and Wood (2004) proposed a method using least-squares estimation to approximate experts' DOBs in a set of models. For this, ratios are elicited, which represent the expert's DOB in one model with respect to another model. These model comparisons (ratios) determine the approximate solution, which is in essence a best-fit solution to the system of constraints (set of ratios) similar to a regression curve that fits a set of data points. Experimental design theory is applied to increase the accuracy of the estimate by optimizing the setup of the experiment prior to expert elicitation. Curtis and Wood suggest to condition the termination of elicitation on the standard error between the system of constraints and the best-fit solution, which varies during the process of elicitation. That is, if this error stops changing significantly it is not expected that additional information will be gained from continuing and one may stop the elicitation process.

We describe here the practical implementation of this approach in order to introduce the fundamental ideas. Detailed mathematical argumentations can be found in the Appendix. Ⓔ For the fully platform-independent, interactive elicitation program, see the electronic supplement to this article.

Prior to any belief elicitation, the selection of the models to be judged is an important issue that may influence results (e.g., the results of a hazard analysis) to the same extent as the subsequent judgment of the applicability of the chosen models. The matter of model selection is discussed, amongst others, by Burnham and Anderson (2002), Scherbaum *et al.* (2004), Delavaud *et al.* (2009), and Bond *et al.* (2012). This article focuses on the judgmental part and thus will assume that experts have already defined the set of n competing models $M_1, \ldots, M_n$ to be judged. Additionally, in this sequel we ignore possible correlations between models.

The next step is to elicit expert knowledge in the form of pairwise comparisons which is done in the following way. An elicitation session is composed of several trials. During each trial, a small subset of the models is presented at one time in such a way that the number of models k is at least two but less than or equal to five. If more than two models are presented, the expert still has to judge them relative to each other by moving sliders on the user interface (Fig. 2). That is, she/he provides ratios

$$R_{ij} = \frac{\mathrm{DOB}(\text{model } M_i)}{\mathrm{DOB}(\text{model } M_j)} \tag{1}$$

for pairs of models, in which $i \neq j$, and each ratio R_{ij} expresses her/his DOB in the applicability of model M_i relative to model M_j (e.g., twice as strong, 0.5 times as strong, etc.). As in Curtis and Wood (2004), we elicit this information by first asking the expert which of the models is most applicable

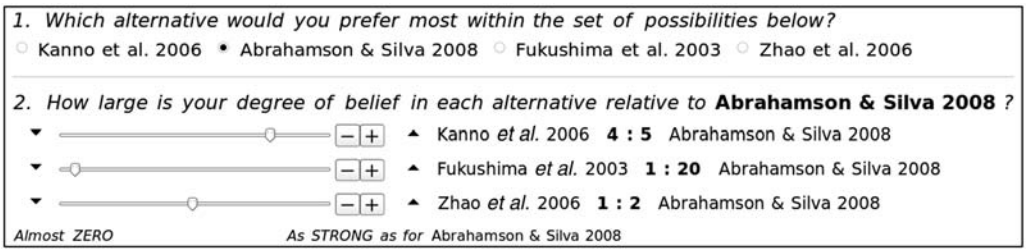

Figure 2. Illustration of a fictional elicitation trial conducted with our interactive program, in which four ground-motion models are judged relative to each other.

Table 1
Number of Ratios n_R Derivable from One Trial

k	2	3	4	5
n_R	1	3	6	10

k is the number of models presented per trial; it holds: $n_R = (k/2)$.

and then asking for the DOB in all other models relative to the most favored one (see Fig. 2). The number of collected ratios per trial depends on the number of models presented (Table 1). Each ratio R_{ij} results in a linear constraint on the expert's DOB in the following sense: if we call vector $(P_1, \ldots, P_n)$ the (unknown) distribution of the expert's DOB in the set of models $M_1, \ldots, M_n$, then the knowledge of ratio R_{ij} is equivalent to

$$R_{ij} \times P_j - P_i = 0 \tag{2}$$

simply by rearranging $R_{ij} = \frac{P_i}{P_j}$. After a sufficient number of trials (explained below), we end up with a system of linear constraints; we would like to approximate the system solution, the vector $\mathbf{p} = (P_1, \ldots, P_n)^T$. A standard approach to this problem is to use the method of least squares. This method approximates a solution for an overdetermined system by minimizing the sum of squared residuals. In our context, a residual is the deviation from equality in any observed constraint (2), when inserting a potential solution for $(P_1, \ldots, P_n)$.

Besides providing a best-fit solution, this method, which is widely used in regression analysis, enables the consistency of all the expert's statements to be checked by analyzing the residuals. For this purpose, the so-called standard error of the regression (SER) will be used, which is the square root of the averaged sum of the squared residuals. The exact form of this SER is defined in equation (A6). A high SER indicates many contradicting statements, that is, a low level of consistency. A low SER means that many statements are consistent with one another. Extending the approach of Curtis and Wood (2004), we will additionally consider a partial standard error; that is, we evaluate the error $\mathrm{SE}(P_i)$ which is contributed by all constraints of the form (2) that contain P_i. This quantifies the consistency of all statements which concern model M_i. The exact formula can be found in equation (A8).

Central questions which arise relate to the optimal order in which model comparisons should be presented to the expert and to the criterion for termination of the elicitation process. In order to solve the former issue, experimental design theory will be applied. Statistical experimental design is a field of statistics that developed with the aim of maximizing the information expected to be obtained from a future experiment prior to observation. This is achieved by selecting a particular (statistically optimal) setup for an experiment, from the set of all possible setups (Chaloner and Verdinelli, 1995). Mathematical details of the method of selecting the optimal set of models to be presented in the next trial can be found in the Appendix and in Curtis and Wood (2004). Essential in maximizing the information obtained through an experiment is that it not only improves the accuracy of estimations, but also reduces the duration of the experiment (the number of trials) required to achieve reliable results. However, optimality relies on expected information gain. This requires a minimum of elicited expert knowledge prior to the design of future trials. That is why the first information is elicited by presenting the expert arbitrarily chosen models within initial trials. This is done in such a way that:

1. every model occurs at least once (condition C1);
2. every trial includes at least one model that also occurs in another trial (condition C2).

Both conditions together ensure that all model comparisons are connected. Both conditions require a minimum number of trials to be conducted, the value of which depends on the number of models presented per trial. This number of minimum trials enables the calculation of expected information, which is followed by an iterated process of optimally designing the next trial, then eliciting new model comparisons. (See Appendix for additional information on these conditions.)

The question of when to stop the elicitation can be addressed by observing the change of the SER during the process of elicitation. After a sufficient number of trials the SER will converge to an equilibrium level. The reason for this is that, if enough model comparisons have been collected, any additional comparison will not add new information to the already elicited set of comparisons. That is, additional comparisons will not increase the expert's consistency and thereby decrease the SER. Of course, if trials continue for too long the expert will tire, her/his consistency will decrease, and the SER will increase. So, once the SER does not change significantly from iteration to iteration, one should stop the elicitation process. Refining this idea of Curtis and Wood (2004), we implemented a statistical significance test, which automatically terminates the conducting of trials. The exact form of this test is given by equations (A16) and (A17). This test requires the definition of a threshold, which introduces some arbitrariness. This threshold (how narrow the equilibrium level of residual variation from iteration to iteration should be) strongly influences the termination condition.

Simulation Experiment

In order to analyze the behavior of the elicitation algorithm presented in the Interactive Elicitation Tool section, we chose four scenarios of possible expert judgment. The purpose of this simulation experiment was to identify the adequate threshold mentioned in that section and to better understand the interpretation of particular elicitation results.

Each scenario describes one particular type of expert judgment. To simulate each elicitation experiment we draw the inputs, which in a real application would have been given by an expert, from a previously defined probability distribution. Thus the probability distribution defines the performance of a symbolic or virtual expert. The advantages of this are twofold: (1) we know precisely how this virtual expert performs and hence can evaluate how the elicitation algorithm would help that expert, and (2) we can perform many more trials with a virtual expert than with a real expert due to human fatigue (in our virtual expert we do not simulate fatigue). For each case of expert judgment we simulate 10,000 elicitation experiments, each with as many iterations as are required, to analyze the average performance of the algorithm.

Noninformative Judgment

The first case deals with experts who have no preference for any of the considered models; that is, they judge all models as being equally applicable. Possible reasons for indifferent judgment may be of cognitive origin (paralysis of choice [Scheibehenne *et al.*, 2010], decision fatigue [Baumeister, 2002]) or of motivational origin (e.g., social pressure), or it may simply be due to lack of knowledge. Indifferent model evaluation can result in two contrary effects: the expert provides either ratios, which are all equal to one, or the expert randomly varies ratios. This section provides the effects of equal judgment on the algorithm's results, and then the effects of arbitrary judgment will be demonstrated.

Equal Model Judgment. When being faced with too many models, people often resort to heuristics such as the following, which was described by Fox and Clemen (2005): "… a judge begins with equal probabilities for all events to be evaluated and then adjusts this uniform distribution based on his or her beliefs about how the likelihoods of the events differ."

Subjective estimates will be biased because this adjustment, which strongly depends on the number of considered models, shifts estimates toward the simple average. We will simulate the extreme case for which no beliefs about differences are expressed and models are judged as being equally applicable. Hence, the expert simply sets $R_{ij} = 1$ for all models M_i and M_j presented. Because all values are equal, the only difference between individual simulated sessions is which models are presented during each trial and, therefore, which models are chosen to be presented in the next trial (by the design algorithm explained in the Appendix). We fix the number of fictional models considered in this simulation to $n = 9$ and present two models per trial.

Simulation Results. The experiments show that the estimation of each DOB P_i equals $\frac{1}{n}$:

$$\hat{P}_i = \frac{1}{9} = 0.111\ldots i = 1, \ldots, 9, \tag{3}$$

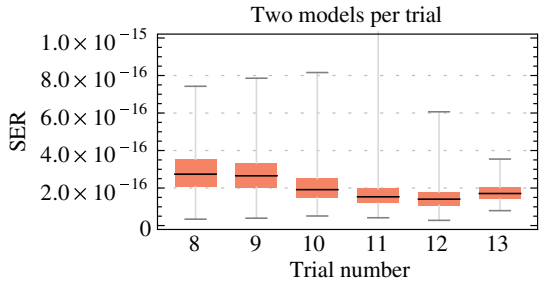

Figure 3. Boxplots of SER distributions with increasing trial number: black line, median; shaded box, half of all SERs in the 10,000 elicitation sessions; lower (or upper) whisker, smallest (or largest) SER. The color version of this figure is available only in the electronic edition.

in which $\hat{P}_i$ is the estimate of P_i. This result was to be expected because all models were judged as being equally applicable and all P_i's have to sum up to one. Convergence of P_i to $\frac{1}{n}$ is very rapid, due to the constantly small SER (with respect to the magnitude) over all trials (Fig. 3): Median(SER) $\approx 1 \times 10^{-16}$ (machine precision). The reason for this is that if all ratios equal one there will be no inconsistent constraints, and thus the vector $(\hat{P}_1, \ldots, \hat{P}_9)$ fulfills all constraints.

An elicitation session is designed such that it stops automatically if the SER does not change significantly anymore (explained in the Appendix). The absence of inconsistencies leads to a median termination reached after 12 trials. This should therefore represent a minimum number of trials in order to ensure a reasonable estimation of $(P_1, \ldots, P_9)$ in practice. Because only two models are compared within every trial, the sessions stop after 12 model comparisons in the case of such small SERs.

Interpretation. The use of the heuristic to judge models as being equally applicable rapidly leads to very consistent statements and also to the same probabilities for all n models. If the results fulfill

$$\hat{P}_1 \approx \ldots \approx \hat{P}_n \approx \frac{1}{n} \quad \text{and} \quad \text{SER} \approx 0, \tag{4}$$

an expert could have used this heuristic. However, it is not possible to distinguish between equal model judgment due to cognitive or motivational reasons or due to conscious judgment based on (lack of) knowledge or on equal applicability of the models in question.

Arbitrary Model Judgment. Arbitrariness poses the opposite extreme case of indifferent model judgment; we assume here that an expert offers an arbitrary judgment for each comparison. In addition, we want to discover the magnitude of the SER in the case that the algorithm's inputs (ratios) take arbitrary values. This should provide an average upper bound for the possible SER.

Arbitrary ratios R_{ij} will be simulated by distributing the corresponding slider positions S_i, S_j uniformly on their full range. We then set

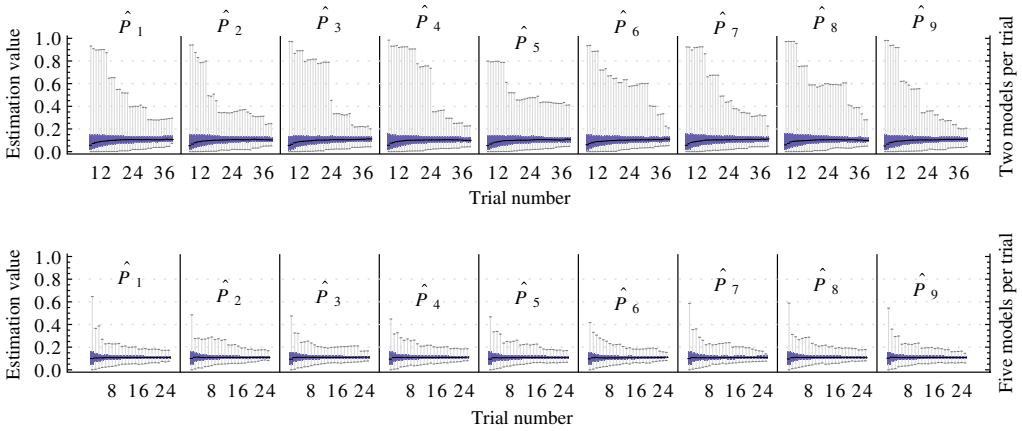

Figure 4 Boxplots of the distributions of estimates $\hat{P}_1, \ldots, \hat{P}_9$ with increasing trial number (horizontal axes). Black line, median; shaded box, half of all estimates; lower (or upper) whisker, smallest (or largest) estimate. The color version of this figure is available only in the electronic edition.

$$R_{ij} = \frac{S_i}{S_j}, \qquad S_i, S_j \sim \text{Uniform(range)}. \tag{5}$$

A complete session terminates when the SER converges to an equilibrium level. This happens sooner the more ratios are derivable of a single trial, so the more models that are presented (Table 1). Below four cases will be compared: the convergence if 2, 3, 4, or 5 models are presented per trial. Again, we assume that the total number of models is nine.

Simulation Results. Regardless of the number of models presented per trial, the estimations $\hat{P}_1, \ldots, \hat{P}_9$ all converge to values close to $1/9$ (Fig. 4). This is consistent with the idea that on average, arbitrary judgment should not put more weight on one model than on any other. However, if presenting two models, the spread of the estimated $\hat{P}_i$ will be much larger than in all other cases: there is a tendency to underestimate the majority of the probabilities P_i while at the same time overestimating a single probability. This effect especially crops up if sessions terminate too early (Table 2). Hence, the fewer the models that are presented, the more trials are required to get a reasonable estimate.

If models are judged arbitrarily, the SER is large because there are many inconsistent constraints and a best-fit solution $(\hat{P}_1, \ldots, \hat{P}_n)$ cannot fulfill all of them at the same time. In this case, the SER converged to a value around 0.06 irrespective of the number of models presented per trial (Fig. 5).

Table 2
Values of Median $\hat{P}_i$ in the 10,000 Elicitation Sessions after t Trials

	k			
t	2	3	4	5
8	0.06	0.10	0.11	0.11
11	0.09	0.11	0.11	0.11
14	0.10	0.11	0.11	0.11
29	0.11	0.11	0.11	$0.11 \approx \frac{1}{9}$

k is the number of models presented per trial.

As would be expected, the median trial number at which the session is terminated depends on the number of models presented per trial (Table 3). Simulated sessions stop earlier if more models are presented because more ratios are derivable after each trial (compare Table 1).

Interpretation. Judging models randomly by providing arbitrary DOB ratios lead to a large number of inconsistencies between the elicited constraints from each trial indicated by a large SER which on average is around 0.06. The fact that the median P_i is generally underestimated is explained by the fact that large residuals tend to dominate least-squares misfit functions and thus solutions. DOBs are assigned randomly and, hence, so are R_{ij}. One R_{ij} value will be the minimum value obtained in each session. The constraint in equation (2) will therefore require that P_j is large for that particular j, and ensuring that this is the case will be critical to minimizing the least-squares residuals. Because $\sum_{i=1}^{n} P_i = 1$ and P_j is large (and random), other P_i s (for $i \neq j$) must be smaller than their true value to compensate. This effect is reduced by adding more trials, hence there are more of the less-extreme constraints, which eventually outweigh the effect of the single constraint from that minimum R_{ij}.

Estimations of P_i represent a compromise between the elicited statements. Generally, when large SERs occur, one should distrust the statements provided by the expert and question the applicability of estimated model weights.

Informative Judgment

Assume an expert believes that the models in question are of different quality. That is, she/he can identify a preference for some models in comparison to the others. Here we assume that we can simulate this situation through two different distributions supposed to capture the uncertainties of experts on the applicability of a set of models. Different distributions make different assumptions about the potential behavior of experts.

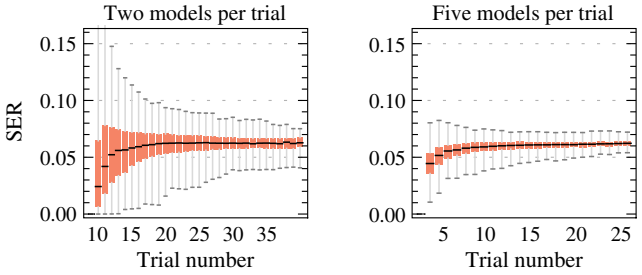

Figure 5 Boxplots of SER distributions with increasing trial number. Black line, median; shaded box, half of all SERs; lower (or upper) whisker, smallest (or largest) SER. The color version of this figure is available only in the electronic edition.

We will first consider the Dirichlet and then the normal distribution. The former distribution generates sets of weights following Kolmogorov's axioms, thus its samples are sets of probabilities. Herewith an expert can be modeled who is aware of, and who consequently applies Kolmogorov's axioms. This is often required if subjective probabilities are used to define a (prior) probability distribution on the set of options in question, which could be achieved for example through the stick-breaking analogy (Paisley *et al.*, 2010; Scherbaum and Kuehn, 2011). The stick-breaking analogy only accepts a weight for a particular model, which is taken from the interval between 0 and 1 minus the sum of the weights which have already been assigned. As a second distribution we use a truncated normal distribution. This models experts, who may not be aware of Kolmogorov's axioms.

Furthermore, the following scenarios simulate informative judgment with a particular degree of confidence in order to capture the expert's uncertainty on her/his judgments. We consider two cases: (1) the expert's confidence in her/his judgments is the same for all models or (2) the confidence in the judgment can be low for some models, yet high for the remaining models.

Same Level of Confidence in the Expert's Judgment. Suppose there is an expert, who has a high degree of confidence in her/his judgment. If this expert would express her/his confidence in the form of DOB ratios, these ratios should correspond closely to the hypothetical ratios, which could be calculated if one knew her/his *exact* probabilities. That is, a repeated request of such ratios will not introduce inconsistencies. DOB ratios from each trial of an expert who has only a low degree of confidence would most probably differ strongly from the hypothetical ratios of her/his exact or underlying probabilities (assuming that low confidence would be translated into imprecise numbers and that such exact or underlying probabilities exist). That is, a repeated request of such ratios will introduce inconsistencies. How can we model this effect? To approach this question, the next paragraph provides a short introduction to the Dirichlet distribution and then describes its application in the present context.

Table 3
Median Trial Number t at Which Session Terminated

k	2	3	4	5
t	31	25	20	20

k is the number of models presented per trial.

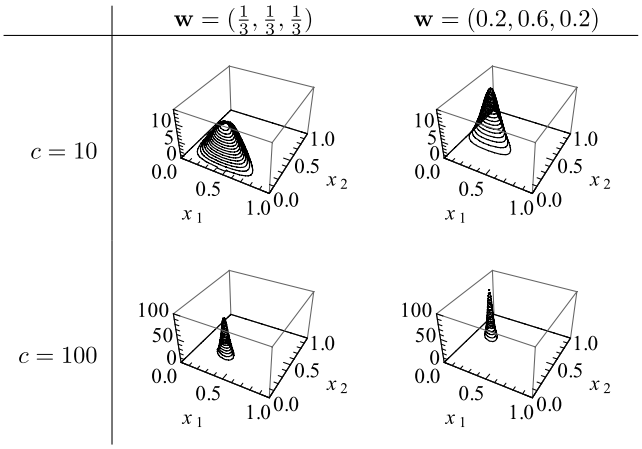

Figure 6. Probability density functions of $\mathrm{Dir}(\mathbf{w}, c)$ for weights vector $\mathbf{w}$ and concentration parameter c.

The Dirichlet distribution, denoted $\mathrm{Dir}(\mathbf{w}, c)$, is a multivariate probability distribution, parametrized by a vector $\mathbf{w} = (w_1, \ldots, w_n)$ of positive real numbers summing to unity and a positive concentration parameter c. Its probability density function returns a random vector of weights $(x_1, \ldots, x_n)$, in which $x_n = 1 - \sum_{i=1}^{n-1} x_i$. The larger the concentration parameter c, the more does the random vector $(x_1, \ldots, x_n)$ concentrate around vector $(w_1, \ldots, w_n)$. Hence, c captures a notion of spread around $(w_1, \ldots, w_n)$. Some examples are given in Figure 6.

For the simulation, the total number of models is set to $n = 9$, and the two cases of presenting two or five models per trial will be compared. In order to model an example of informative expert judgment following Kolmogorov's axioms, distribution $\mathcal{D}$ will be $\mathrm{Dir}(\mathbf{w}, c)$ with

$$\begin{aligned}\mathbf{w} &= (w_1, \ldots, w_9)\\ &= (0.02, 0.5, 0.08, 0.02, 0.02, 0.02, 0.02, 0.3, 0.02)\end{aligned} \quad (6)$$

and $c = 100$. The judgment for each trial t is modeled by drawing one vector of weights $\mathbf{x}_t = (x_{t1}, \ldots, x_{t9})$ from distribution $\mathcal{D}$ and setting the required ratios equal to the ratios of sampled weights that correspond to the models to be judged:

$$R_{ij}^t = \frac{x_{ti}}{x_{tj}}. \quad (7)$$

For each further trial another vector of weights from the same distribution $\mathcal{D}$ will be drawn and the corresponding ratios derived. Thereby, the concentration parameter c defines different levels of consistency within the statements of separate trials.

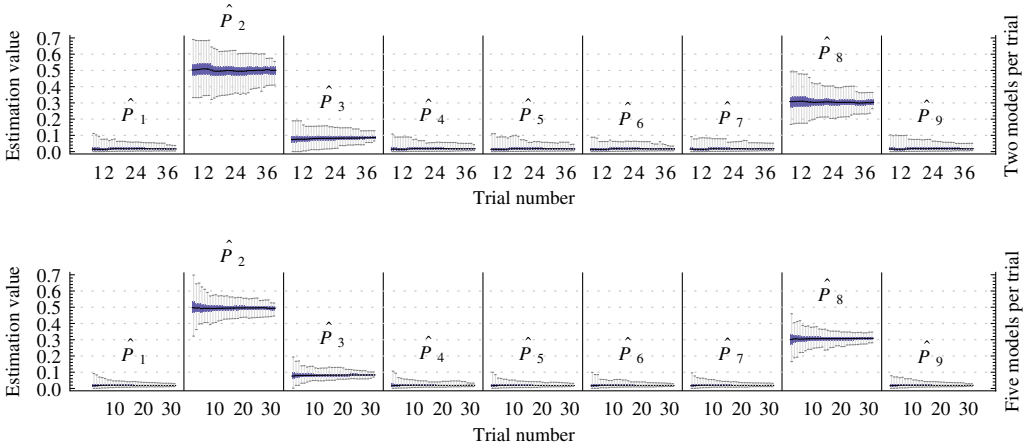

Figure 7 Boxplots of the distributions of estimates $\hat{P}_1, \ldots, \hat{P}_9$ with increasing trial number (horizontal axes). Black line, median; shaded box, half of all values; lower (or upper) whisker, smallest (or largest) value. The color version of this figure is available only in the electronic edition.

If $c \times w_i \approx 1$ (for most $i = 1, \ldots, n$), random vectors drawn from that distribution will differ strongly from each other, and therefore the number of inconsistencies between different trials will be large. If $c \gg 1$, all sampled random vectors will concentrate around $(w_1, \ldots, w_n)$: different trials will comprise only few inconsistencies.

Simulation Results. Distribution $\mathcal{D}$ with concentration parameter $c = 100$ allows its samples to vary around the initial set of weights modeling reasonable consistent judgments. The simulations show that on average estimations $\hat{P}_1, \ldots, \hat{P}_9$ fit the initial weights $(w_1, \ldots, w_9)$, but the sample spread indicates possible over- or underestimation of particular P_i's (Fig. 7). Again, a larger number of models presented per trial contributes positively to the estimation's accuracy, and the sample spread decreases more rapidly.

Initially little inconsistencies increase with more trials up to an SER of around 0.02 (Fig. 8). The median session termination trial number is depicted in Table 4.

Interpretation. An SER of 0.02 on average still enables the estimation of the initial weights $(w_1, \ldots, w_9)$, which would be unknown DOB values in a real application. That is, the preference for particular models (in our case M_2 and M_8), which was implicit in the exact or underlying ratios, has been detected. However, the Dirichlet distribution provides samples with spread dependent on the magnitude of a single weight w_i and a fixed concentration parameter c; human judgment will probably be more complex. For example, it is known that humans are biased, when they attempt to estimate the probability of occurrence of extreme (low probability) events—those on the tails of distribution functions (O'Hagan *et al.*, 2006). This might imply that one would typically encounter greater uncertainty for very low values of w_i than is modeled by the Dirichlet distribution. This is modeled in the next section.

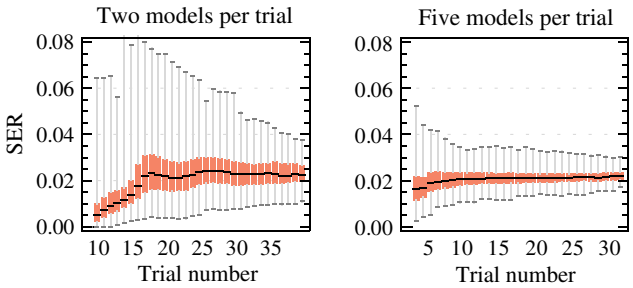

Figure 8. Boxplots of SER distributions with increasing trial number. Black line, median; shaded box, half of all SERs; lower (or upper) whisker, smallest (or largest) SER. The color version of this figure is available only in the electronic edition.

Two Different Levels of Confidence in the Expert's Judgment. Suppose the expert who is judging a set of models, consciously or subconsciously ignores Kolmogorov's axioms. We model this by assuming that the expert thinks in terms of marginal distributions of uncertainties that are independent for each of the models. In this case it is straightforward to accommodate high and low confidence through the variances of the marginal distributions. Let the expert's confidence in their judgment of some of the models be low (which we call LC models), and let that confidence in the judgment of the remaining models be high (HC models). We use the truncated normal distribution, which is restricted to values between zero and one, to model marginal distributions.

Table 4
Median Trial Number t at Which Session Terminated: Half of All Sessions Stopped after $t(k)$ Trials

k	2	3	4	5
t	31	27	25	25

k is the number of models presented per trial.

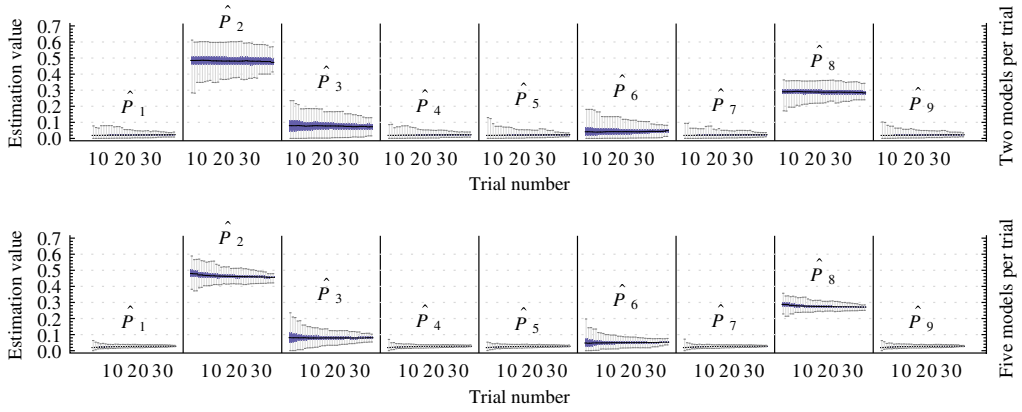

Figure 9 Boxplots of the distributions of estimates $\hat{P}_1, \ldots, \hat{P}_9$ with increasing trial number (horizontal axes). Black line, median; shaded box, half of all values; lower (or upper) whisker, smallest (or largest) value. The color version of this figure is available only in the electronic edition.

Subsequently, a ratio required in trial t is set equal to the ratio of normal random variables

$$R_{ij}^t = \frac{x_{ti}}{x_{tj}}, \tag{8}$$

in which x_{tk} is drawn from the truncated normal distribution $\mathcal{N}(w_k, \sigma_{\mathrm{LC}})$ if model M_k is an LC model and from $\mathcal{N}(w_k, \sigma_{\mathrm{HC}})$ if M_k is an HC model. Two different variances are applied, $\sigma_{\mathrm{LC}} = 0.25^2$ and $\sigma_{\mathrm{HC}} = 0.05^2$. For reasons of comparability, mean values are given by the same vector of 9 (nonuniform, or informative) weights as in the previous scenario:

$$\mathbf{w} = (0.02, 0.5, 0.08, 0.02, 0.02, 0.02, 0.02, 0.3, 0.02). \tag{9}$$

Models 3 and 6 are treated as LC models and all remaining are treated as HC models. Model 6 is a lowest-probability model (probability 0.02), there we consider the effect of the extreme-event bias mentioned above. Finally, the comparison of considering two or five models per trial is illustrated.

Simulation Results. Compared with estimates $\hat{P}_1, \ldots, \hat{P}_9$ in the previous scenario (Fig. 7), the results in this scenario show an increased uncertainty in the estimations for P_3 and P_6, indicated by a larger sample dispersion around the final estimate (Fig. 9). These are exactly the parameters that correspond to LC models. This strengthens the assertion that it is possible to detect different levels of consistency. Furthermore, Figure 9 depicts a slight shift of all estimates toward the average weight of $1/9$, when five models are presented per trial. This may be explained by the fact that more presented models provide more ratios per trial, which means that also more inconsistencies caused by the uncertainty in models M_3 and M_6 are added. That is, the more the inconsistencies that occur during an elicitation session, the more the estimates will shift toward the simple average, which is similar to the effect described by Scherbaum and Kuehn (2011).

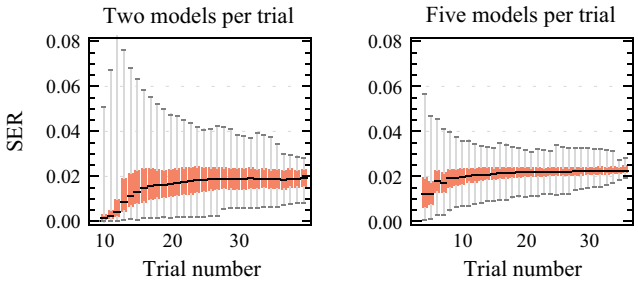

Figure 10. Boxplots of SER distributions with increasing trial number. Black line, median; shaded box, half of all SERs; lower (or upper) whisker, smallest (or largest) SER. The color version of this figure is available only in the electronic edition.

The two different variances $\sigma_{\mathrm{LC}} = 0.25^2$ and $\sigma_{\mathrm{HC}} = 0.05^2$ result in an SER value of around 0.02 and, similarly to before, the sample dispersion of the SERs decreases with a larger number of trials and with more models presented per trial (Fig. 10). Sessions terminate similarly to the previous scenario (Table 5).

In order to analyze the influence of an individual model M_i on the overall SER, we consider only those constraints of the form $R_{ij}P_j - P_i = 0$, which include P_i and ignore all constraints without P_i measured by the partial error $\mathrm{SE}(P_i)$ mentioned in the Interactive Elicitation Tool section. However, $\mathrm{SE}(P_i)$ is observed to depend on the magnitude of $\hat{P}_i$ (Fig. 11a). A higher $\hat{P}_i$ leads to a higher $\mathrm{SE}(P_i)$. In order to obtain a measure which indicates the partial SE relative to

Table 5
Median Trial Number t at Which Session Terminated: Half of All Sessions Stopped after $t(k)$ Trials

k	2	3	4	5
t	31	27	26	29

k is the number of models presented per trial.

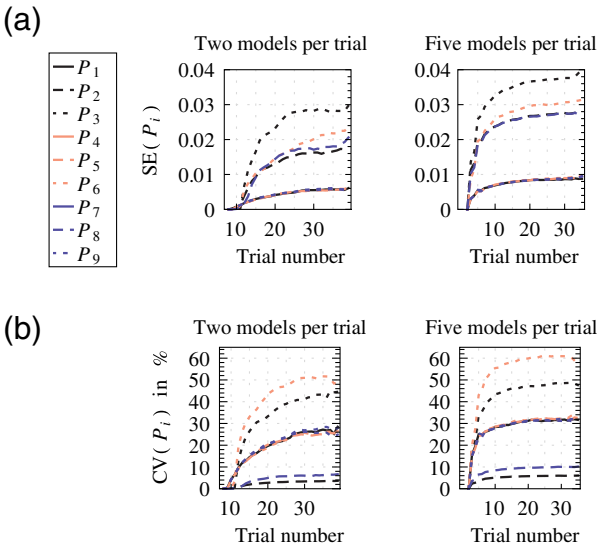

Figure 11. (a) Median SE($\hat{P}_i$) and (b) median CV($\hat{P}_i$) with increasing trial number, for $i = 1, \ldots, 9$. The color version of this figure is available only in the electronic edition.

$\hat{P}_i$'s magnitude, the so-called coefficient of variation (CV) is applied. It is expressed as a percentage and given by

$$\mathrm{CV}(P_i) = \frac{\mathrm{SE}(P_i)}{P_i} \times 100. \qquad (10)$$

From the perspective of CV, a significant difference exists between the inconsistency introduced by the LC models M_3 or M_6 and the inconsistency introduced by all the HC models (Fig. 11b), which is much higher in the former case. The CV of P_3 and P_6 is at least twice the size of all remaining CVs.

Interpretation. Analyzing the SE($\hat{P}_i$) and the CV($\hat{P}_i$) for each estimation $\hat{P}_i$ enables the different degrees of consistency concerning individual models to be distinguished. However, their interpretation remains ambiguous: the presence of many inconsistent constraints indicates that the expert's confidence in her/his judgment of the relevant model is low. A high level of consistency, on the one hand, may be a consequence of high confidence in the judgment. On the other hand, the use of heuristics like equal judgment (possibly to hide a lack of knowledge or low confidence) typically also leads to more consistent statements. Nevertheless, the example illustrates that it is possible to detect different degrees of consistency concerning individual models. This offers a basis for further investigations.

Discussion

In the Interactive Elicitation Tool and Simulation Experiment sections, we presented a technique for eliciting an individual expert's belief. Understanding the belief of a single expert is useful, but making decisions in the presence of uncertainty may require beliefs of several experts. In this context, ambiguity arises concerning the beliefs stated by different experts, who each judge the applicability of a set of models in terms of a subjective preference (question 1 in Fig. 2). The interpretation of the term "preference" may differ from expert to expert, which can lead to a diverging applicability judgment on the same set of models.

At this point, it has to be clarified which aspects were considered during the evaluation of individual judgments in order to eliminate misunderstandings. First, in the context of multiple experts being available, we do not attempt to elicit or estimate biases due to group interactions such as those found in earth sciences applications by Polson *et al.* (2009), Polson and Curtis (2010), and Curtis (2012). The methods presented here instead attempt to quantify each expert's beliefs in the context of such biases. In future it may be possible to combine our methods with those of Polson and Curtis (2010) in order to analyze the dynamism in belief that group interaction can cause. Second, the Simulation Experiment section illustrates how particular inputs influence the results and the termination of the elicitation technique. Equal and arbitrary judgment, both representations of noninformative judgment, lead to equal (average) DOB estimations, but corresponding SERs differ strongly and thus the algorithm termination varies for each case. If one is interested in using the elicitation sessions to try to identify the expert's own definition of a subjective prior in a Bayesian setting, then based on such noninformative expert statements the value of the SER may determine the prior's variance. Similarly, informative expert judgment leads to an SER which may serve as a basis for estimating the variance of an informative prior.

Other computerized elicitation methods have been developed (e.g., Probability Elicitation System [PROBES] by Lau and Leong [1999] or Elicitator by James *et al.* [2010]), which offer a user-friendly interface and apply different techniques to extract subjective beliefs. In this context, our interactive program complements their work and represents an expansion of their pool of probability assessment techniques.

Returning to our design algorithm itself, we repeated the experiments in the case of informative judgment and used randomly designed trials instead of the design algorithm. In 10,000 simulated sessions, random design was observed in median to lead to an increased spread of the SER (difference between the upper and lower quartiles) by around 17% of the spread obtained with optimal design. This indicates that the application of an optimal design improves the accuracy of the DOB estimates compared to a random design of trials.

Furthermore, the algorithm termination is dependent on the number of models presented per trial in addition to the SERs fluctuation (compare Tables 3, 4, and 5). A larger number of presented models provides more ratios per trial (Table 1) and thus results in a convergence of the standard error within fewer trials. On the other hand, it will be easier for an expert to concentrate on only two models at a time. Therefore, one has to trade off the feasible number of trials to be conducted by the expert against the advantages of presenting only two or three models at one time.

During the discussion of the limitations and benefits of pairwise comparison elicitation through the presented program we have not included any real trials with live experts. Instead we focus on characterizing the performance of the interactive program and method for different types of virtual experts, the advantage being that we know (define) their behaviors. Curtis and Wood (2004) tested the algorithm on a single expert. However, real trials introduce a level of psychological interpretation that is complex and which requires dedicated analysis. Clearly future work will require that the new interactive program is tested on a cohort of real experts for its practical utility to be verified.

Future experience of practitioners may certainly lead to an improvement of the tool's utility by implementing additional features; for example, the possibility to group models and ask for relative weights within and between subsets might reduce the number of trials and help to focus effort on possible substantial differences between clusters of models. Furthermore, future usage of the tool will provide tacit experience to help to evaluate which is the most useful measure of confidence to be applied in the relative weighting. Currently the SER quantifies the consistency (compare equation A6), but it is also possible to think of perturbing some default value, that is, decreasing it when confident (or opinionated) or increasing it when not (or exhausted). If different confidence measures will be incorporated it might be necessary to modify the algorithm's termination rule (equation A17), which depends on the applied measure of consistency/confidence.

Conclusions

In the light of considerable uncertainties, many applications in science and engineering require experts to judge several competing options. In order to support this task, this article presented:

1. a platform-independent, interactive program for intuitive and rapid (self-)elicitation of expert beliefs, which at the same time indicates the strength of their beliefs;
2. the simulation of expert judgment in order to analyze end-member cases of indifferent judgment and two simple cases of informative judgment; and
3. the identification of substantial key data, such as the rough number of trials to be conducted, typical DOB estimations for particular scenarios, and corresponding residual errors.

All findings from the simulation studies may serve as a guide for practical applications. In a Bayesian context, they support our suggestion of a possible connection between the expert's belief and an informative prior distribution, for which the location and spread are determined on the basis of the DOB estimates and the SER.

Data and Resources

No data were used in this paper. Some plots were made using simulated data. The interactive program De Finetti's Game written by F. Scherbaum can be found at http://demonstrations.wolfram.com/DeFinettisGame (last accessed 29 April 2013).

Acknowledgments

We thank anonymous reviewers for helpful comments. Antonia K. Runge is funded by BMBF project PROGRESS and Carsten Riggelsen is funded by DFG project RI 2037/2-1.

References

Baddeley, M. C., A. Curtis, and R. Wood (2004). An introduction to prior information derived from probabilistic judgments: Elicitation of knowledge, cognitive bias and herding, *Geol. Soc. Lond. Spec. Publ.* **239,** 15–27.

Baumeister, R. F. (2002). The psychology of irrationality, in *The Psychology of Economic Decisions*, J. D. Carrillo and I. Brocasand (Editors), Vol. 1**,** Oxford University Press Inc., New York, 1–15.

Bommer, J. J. (2012). Challenges of building logic trees for probabilistic seismic hazard analysis, *Earthq. Spectra* **28,** 1723–1735.

Bommer, J. J., and F. Scherbaum (2008). The use and misuse of logic trees in probabilistic seismic hazard analysis, *Earthq. Spectra* **24,** 997–1009, doi: 10.1193/1.2977755.

Bommer, J. J., F. Scherbaum, H. Bungum, F. Cotton, F. Sabetta, and N. A. Abrahamson (2005). On the use of logic trees for ground-motion prediction equations in seismic-hazard analysis, *Bull. Seismol. Soc. Am.* **95,** 377–389.

Bond, M. A., R. J. Lunn, Z. K. Shipton, and A. D. Lunn (2012). What makes an expert effective at interpreting seismic images? *Geology* 40,75–78, doi: 10.1130/G32375.1.

Burnham, K. P., and D. R. Anderson (2002). Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach, Second Ed., Springer, New York, 488 pp.

Chaloner, K., and I. Verdinelli (1995). Bayesian experimental design: A review, *Stat. Sci.* **10,** 273–304, doi: 10.1214/ss/1177009939.

Curtis, A. (2012). The science of subjectivity, *Geology* **40,** 95–96.

Curtis, A., and R. Wood (2004). Optimal elicitation of probabilistic information from experts, *Geol. Soc. Lond. Spec. Publ.* **239,** 127–145.

Curtis, A., A. Michelini, D. Leslie, and A. Lomax (2004). Deterministic design of geophysical surveys by linear-dependence reduction, *Geophys. J. Int.* **157,** 595–606.

Delavaud, E., F. Scherbaum, N. M. Kuehn, and C. Riggelsen (2009). Information-theoretic selection of ground-motion prediction equations for seismic hazard analysis: An applicability study using Californian data, *Bull. Seismol. Soc. Am.* **99,** 3248–3263.

Fox, C. R., and R. T. Clemen (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior, *Manag. Sci.* **51,**1417–1432.

James, A., S. L. Choy, and K. Mengersen (2010). Elicitator: An expert elicitation tool for regression in ecology, *Environ. Model. Software* **25,** 129–145.

Knol, A. B., P. Slottje, J. P. Van Der Sluijs, and E. Lebret (2010). The use of expert elicitation in environmental health impact assessment: A seven step procedure, *Environ. Health* **9,** 1–16.

Lau, A.-H., and T.-Y. Leong (1999). PROBES: A framework for probability elicitation from experts, in *AMIA Annual Symposium, American Medical Informatics Association*, 301–305.

Läuter, H., and R. Pincus (1989). *Mathematisch-Statistische Datenanalyse*, Akademie-Verlag, Berlin, 392 pp.

Meyer, M. A., and J. M. Booker (2001). *Eliciting and Analyzing Expert Judgment: A Practical Guide*, in ASA-SIAM Series on Statistics and Applied Probability, R. F. Gunst, S. Keller-McNulty, R. L. Mason, M. D. Morris, J. P. Buckingham, J. F. Pendergast, J. A. Calvin, R. N. Rodriguez, and G. C. McDonald (Editors), American Statistical

Association and the Society for Industrial and Applied Mathematics, Alexandria, Virginia/Philadelphia, Pennsylvania.

O'Hagan, A., C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow (2006). Uncertain Judgments: Eliciting Experts' Probabilities, in *Statistics in Practice Series*, S. Senn, M. Scott, and P. Bloomfield (Editors), John Wiley & Sons, England, 338 pp.

Paisley, J., A. Zaas, C. W. Woods, G. G. Ginsburg, and L. Carin (2010). A stick-breaking construction of the beta process, presented at the *27th International Conference on Machine Learning*, Haifa, Israel, 2010.

Polson, D., and A. Curtis (2010). Dynamics of uncertainty in geological interpretation, *Geol. Soc. Lond.* **167,** 5–10, doi: 10.1144/0016-76492009-055.

Polson, D., A. Curtis, C. Vivalda, and S. Saunier (2009). Process for tracking the evolving perception of risk during CO2 storage projects, Publication SPE 124703, in *Proc. of the 2009 SPE Offshore Europe Oil & Gas Conference & Exhibition*, Aberdeen, United Kingdom, 8–11.

Saaty, T. L. (1980). *The Analytic Hierarchy Process: Planning, Priority Setting, and Resource Allocation*, McGraw–Hill, New York, 287 pp.

Scheibehenne, B., R. Greifeneder, and P. M. Todd (2010). Can there ever be too many options? A meta-analytic review of choice overload, *J. Consumer Res.* **37,** 409–425.

Scherbaum, F., and N. M. Kuehn (2011). Logic tree branch weights and probabilities: Summing up to one is not enough, *Earthq. Spectra* **27**.

Scherbaum, F., F. Cotton, and P. Smit (2004). On the use of response spectral-reference data for the selection and ranking of ground-motion models for seismic-hazard analysis in regions of moderate seismicity: The case of rock motion, *Bull. Seismol. Soc. Am.* **94,** 2164–2185, doi: 10.1785/0120030147.

Toro, G. R., N. A. Abrahamson, and J. F. Schneider (1997). Model of strong ground motions from earthquakes in central and eastern North America: Best estimates and uncertainties, *Seismol. Res. Lett.* **68,** 41–57, doi: 10.1785/gssrl.68.1.41.

Tversky, A., and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases, *Science* **185,** 1124–1131, doi: 10.1126/science.185.4157.1124.

Appendix

The next two sections give a brief overview of the method of Curtis and Wood (2004), while the final section describes the stopping criteria for the termination of elicitation sessions.

Measures of Consistency

Let us assume that trials have been conducted up to any trial number. Rearranging all elicited constraints of the form

$$R_{ij} = \frac{P_i}{P_j} \tag{A1}$$

leads to a system of linear equations of the form

$$R_{ij}P_j - P_i = 0 \tag{A2}$$

for $i, j \in \{1, \ldots, n\}, i \neq j$. Additionally, the constraint

$$P_1 + P_2 + \ldots + P_n = 1 \tag{A3}$$

(that all DOB values should sum up to one) is appended to this system. The equations can be summarized by combining all of constraints (A2) and (A3) into matrix form

$$\mathbf{X}\mathbf{p} = \mathbf{d}. \tag{A4}$$

Here each row of matrix $\mathbf{X}$ corresponds to a single constraint, $\mathbf{p} = (P_1, \ldots, P_n)^T$ is the vector of unknown DOB values to be estimated, and $\mathbf{d}$ is the vector $(0, \ldots, 0, 1)$, assuming that constraint (A3) comes last. The least-squares estimation of $\mathbf{p}$ is given by

$$\hat{\mathbf{p}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{d}. \tag{A5}$$

Conditions (C1) and (C2) in the Interactive Elicitation Tool section guarantee that $\mathbf{X}^T\mathbf{X}$ is invertible.

The standard error which we discussed in the Interactive Elicitation Tool section and is applied in the Simulation Experiment section, is given by

$$\mathrm{SER} = \sqrt{\frac{1}{k-n}(\mathbf{d} - \mathbf{X}\hat{\mathbf{p}})^T(\mathbf{d} - \mathbf{X}\hat{\mathbf{p}})}, \tag{A6}$$

in which k is the number of rows in $\mathbf{X}$ and n is the total number of models. The SER quantifies the deviation of the best-fit solution from all elicited statements. In essence it measures the scatter of all statements and thus indicates how inconsistent they are from one another.

We refine Curtis and Wood's method in order to estimate the error contribution from a single model M_k. For this purpose we define matrix $\mathbf{X}^{P_i}$ and vector $\mathbf{d}^{P_i}$ such that only equations of the form (A2) that contain P_i are included and expressed as

$$\mathbf{X}^{P_i}\mathbf{p} = \mathbf{d}^{P_i}. \tag{A7}$$

A partial standard error will be defined by

$$\mathrm{SE}(P_i) = \sqrt{\frac{1}{k-1}(\mathbf{d}^{P_i} - \mathbf{X}^{P_i}\hat{\mathbf{p}})^T(\mathbf{d}^{P_i} - \mathbf{X}^{P_i}\hat{\mathbf{p}})}, \tag{A8}$$

in which k is the number of rows in $\mathbf{X}^{P_i}$. A large $\mathrm{SE}(P_i)$ indicates the occurrence of many inconsistencies within elicited statements, that include model M_i.

Optimal Design of Elicitation Trials

Let us assume that the minimum number of trials (such that conditions C1 and C2 described in the Interactive Elicitation Tool section are fulfilled) has been conducted, and let matrix $\mathbf{X}$ and vector $\mathbf{d}$ be defined as explained for equation (A4). Matrix $\mathbf{X}$ contains information on all values of the elicited expert knowledge and on the order of the requested ratios (assuming that the order in which the constraints had been elicited corresponds to the order of the rows of $\mathbf{X}$). Therefore, $\mathbf{X}$ is commonly called the "design matrix" because it defines the design of the experiments conducted so far. A different ordering of requested model comparisons will lead to a different matrix $\mathbf{X}$. The covariance matrix of $\hat{\mathbf{p}}$ in (A5) is estimated by

$$\mathrm{Cov}(\hat{\mathbf{p}}) = \mathrm{SER}^2(\mathbf{X}^T\mathbf{X})^{-1}. \quad \text{(A9)}$$

This means, that the variance of the estimation will be small ($\hat{\mathbf{p}}$ is more informative) if $\mathbf{X}^T\mathbf{X}$ is "large" in some sense. The magnitude of a matrix is commonly characterized by its determinant, its trace, or other functionals of its eigenvalues. The determinant is known to produce balanced designs (Läuter and Pincus, 1989); in our context this means that a high determinant can be associated with experiments that result in elicited information which is evenly spread among all models. A low determinant indicates that most information is concentrated on one (or several) particular model(s), but not all. Because the elicitation of expert knowledge about all models to a similar extent is desirable, our aim is to design future trials that maximize $|\mathbf{X}^T\mathbf{X}|$.

In order to design future trials based on previous trial information we use the method of Curtis *et al.* (2004). Already elicited knowledge will be captured by matrix $\mathbf{X}_1$ and vector $\mathbf{d}_1$, and expert statements $R_{kl}^{\exp} \times P_l - P_k = 0$ which are expected to be derived from the next trial, will be captured by matrix $\mathbf{X}_2$ and vector $\mathbf{d}_2$. Already elicited trials and the next (future) trial are linked together via block matrix $\mathbf{X}$ and vector $\mathbf{d}$,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{d} = \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix}, \quad \text{(A10)}$$

and we expect that $\mathbf{Xp} = \mathbf{d}$ for some unknown future $\mathbf{d}_2$. The design of our experiment is optimal if the DOB ratios derived from the models presented in the next trial add the most possible new information to the set of already elicited ratios. Hence, in terms of determinants, the design of $\mathbf{X}$ will be optimal if we choose matrix $\mathbf{X}_2^*$ from the set of all possible $\mathbf{X}_2$s (which could be constructed after the next trial), such that

$$\left|\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2^* \end{bmatrix}^T \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2^* \end{bmatrix}\right| = \max_{\mathbf{X}_2} \left|\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}^T \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}\right|. \quad \text{(A11)}$$

A fast but potentially suboptimal algorithm to maximize equation (A11) over $\mathbf{X}_2$ was described by Curtis *et al.* (2004) and used by Curtis and Wood (2004). However, in the light of current fast computers, the global maximum in equation (A11) can be determined directly by comparing the determinants of all possible $\mathbf{X}^T\mathbf{X}$. Once $\mathbf{X}_2^*$ had been found, the optimal selection of models to be presented in the next trial is given by the positions of the nonzero entries of the rows of $\mathbf{X}_2^*$.

Automatic Elicitation Termination

Let us assume that t trials had been conducted where $t \geq t_{\min}$ and $t_{\min}$ is the minimum number of required trials to fulfill conditions (C1) and (C2) of the Interactive Elicitation Tool section. Each subsequent trial adds more statements to the set of those previously elicited, which in turn changes the SER. Let matrix $\mathbf{X}_{1:t_{\min}}$ and vector $\mathbf{d}_{1:t_{\min}}$ be defined such that all equations (in the linear form of equation A2) derived from the trials 1 until $t_{\min}$ may be expressed as

$$\mathbf{X}_{1:t_{\min}} \times \mathbf{p} = \mathbf{d}_{1:t_{\min}}. \quad \text{(A12)}$$

If matrix $\mathbf{X}_t$ and vector $\mathbf{d}_t$ describe all new equations derived from the tth trial, then the following holds:

$$\mathbf{X}_{1:t} = \begin{bmatrix} \mathbf{X}_{1:t-1} \\ \mathbf{X}_t \end{bmatrix}, \qquad \mathbf{d}_{1:t} = \begin{bmatrix} \mathbf{d}_{1:t-1} \\ \mathbf{d}_t \end{bmatrix}, \quad \text{(A13)}$$

for all $t > t_{\min}$. Then, the estimate of $\mathbf{p}$ after t trials is given by

$$\hat{\mathbf{p}}_t = (\mathbf{X}_{1:t}^T\mathbf{X}_{1:t})^{-1}\mathbf{X}_{1:t}^T\mathbf{d}_{1:t} \quad \text{(A14)}$$

and the SER after t trials by

$$\mathrm{SER}_t = \sqrt{\frac{1}{k_t - n}(\mathbf{d}_{1:t} - \mathbf{X}_{1:t}\hat{\mathbf{p}}_t)^T(\mathbf{d}_{1:t} - \mathbf{X}_{1:t}\hat{\mathbf{p}}_t)}, \quad \text{(A15)}$$

in which k_t is the number of rows of $\mathbf{X}_{1:t}$, n is the number of models, and $t \geq t_{\min}$. A significance test is constructed such that hypothesis $\mathcal{H}_0$ is tested against hypothesis $\mathcal{H}_a$, in which

$$\mathcal{H}_0 : \delta \geq \varepsilon \quad \text{and} \quad \mathcal{H}_a : \delta < \varepsilon, \quad \text{(A16)}$$

for

$$\delta = \frac{|\mathrm{SER}_\tau - \mathrm{SER}_{\tau-1}|}{\max(0.001, \mathrm{SER}_{t_{\min}}, \mathrm{SER}_{t_{\min}+1}, \ldots, \mathrm{SER}_\tau)} \quad \text{(A17)}$$

and $\tau \in \{t_{\min} + 1, \ldots, t\}$ and threshold $\varepsilon = 0.05$ in our study. If $\mathcal{H}_0$ is rejected (false), then the difference between two consecutive SERs became significantly small (significance level $\alpha = 0.05$). This means that the SER did not change significantly and the elicitation is stopped.

Institute of Earth and Environmental Science
University of Potsdam
Karl-Liebknecht-Str. 24-25
14476 Potsdam, Germany
Antonia.Runge@geo.unipotsdam.de
Frank.Scherbaum@geo.uni-potsdam.de
Carsten.Riggelsen@geo.uni-potsdam.de
(A.K.R., F.S., C.R.)

School of GeoSciences
University of Edinburgh
Grant Institute
The King's Buildings
West Mains Road
Edinburgh EH9 3JW, UK
Andrew.Curtis@ed.ac.uk
(A.C.)

Manuscript received 29 January 2013