LA-UR-00-218

Approved for public release;
distribution is unlimited.

Title: Guidelines for Eliciting Expert Judgment as Probabilities or Fuzzy Logic

Author(s): Mary A. Meyer
Kenneth B. Butterfield
William S. Murray
Ronald E. Smith
Jane M. Booker

Submitted to: *Fuzzy Logic and Probability Applications*
American Statistical Society, SIAM Series

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory report LA -UR-00-218 to appear in Ross, Booker and Parkinson (Eds), *Fuzzy Logic and Probability Applications*, American Statistical Society, SIAM Series. Forthcoming.

DRAFT Chapter 5

Guidelines For Eliciting Expert Judgment As Probabilities or Fuzzy Logic

Mary A. Meyer, Kenneth B. Butterfield, William S. Murray, Ronald E. Smith, Jane M. Booker

1. Introduction

We recommend **formal elicitation of expert judgment** as a method for obtaining probabilities or fuzzy rules from individuals. Formal elicitation of expert judgment draws from the fields of cognitive psychology, decision analysis, statistics, sociology, cultural anthropology, and knowledge acquisition. It entails the use of specific procedures to identify the experts, define the technical problems, and elicit and document the experts' judgment.[1] Expert judgment may be expressed as **probabilities** (either point estimates, 90%, or as probability distribution functions), or **fuzzy terms** (for example, *low*, *medium*, and *high*). The experts' sources of information, considerations, and assumptions are all documented as part of the experts' judgment. Formal elicitation can counter common biases arising from human cognition and behavior. The benefits of formal elicitation are added rigor, defensibility of the judgments, and increased ability to update the judgments as new information becomes available.

We provide the reader with guidelines and examples of formal elicitation in the following areas:

- Determining whether expert judgment can be feasibly elicited.
- Determining whether the expert's judgment can be better elicited in a probabilistic or fuzzy framework.
- Formulating the technical questions.
- Structuring the interview situations for one expert, multiple experts, or teams of experts.
- Eliciting and documenting the expert judgment.

---

[1] Expert judgment can also be elicited informally, even tacitly. The experts are asked for their best guess, which is then applied to the analysis with no questions asked. Hence this informal approach has been called "ask and use" (French, McKay, and Meyer, 1999). However, the trend seems to be to formalize expert judgment procedures in a number of areas, such as expert testimony in the legal system, medicine, environmental analysis, and nuclear safety analysis (Stanbro and Budlong-Slyvestor, 2000).

- Representing the expert judgment for the expert's review and refinement.
- Facilitating the comparison of multiple experts' judgments.

Expert judgment is an expert's informed opinion, based on knowledge and experience, given in response to a technical problem (Ortiz et al, 1991). Expert judgment can be viewed as a representation, a snapshot, of the expert's state of knowledge at the time of his or her response to the technical problem (Keeney and von Winderfeldt, 1989).

Expert judgment is used to:

- predict future events;
- provide estimates on new, rare, complex, or poorly understood phenomena;
- integrate or interpret existing information;
- learn an expert's problem-solving process or a group's decision-making processes; or
- determine what is currently known, how well it is known, or what is worth learning in a field (Meyer and Booker, 1991).

Expert judgment may be expressed in quantitative or qualitative form for fuzzy and probabilistic applications. Examples of judgments given in quantitative form include probabilities, uncertainty estimates, and membership functions, and are often given in reference to other quantities of interest, such as performance, cost, and time. Examples of qualitative form include:

- the experts' natural language statements of physical phenomena of interest (for example, "the system performs well under these conditions");
- "if-then" rules (for example, "if the temperature is high, then the system performs poorly"); and
- textual descriptions of their assumptions in reaching an answer; and reasons for selecting or eliminating certain data or information from consideration.

Whatever its form, expert judgment should include a description of the experts' thinking—their reasoning, algorithms, and assumptions—and the information they considered in arriving at a response. Ideally, the expert judgment provides a complete record allowing decision makers, other experts, or novices to track the experts' problem-solving processes to their answers.

2. Method

2.1 Illustration

We illustrate the phases and steps of expert elicitation with four probability and fuzzy examples. Backgrounds on these four examples are given below:

2.1.1 Probability Example: Predicting Automotive Reliability

This example of probability elicitation is an automotive application whose goal is to characterize the reliability of new products during their developmental programs. (Kerscher et al., 2000). Characterizing the reliability in the early developmental phases poses problems because traditional reliability methods require test data to characterize the reliability. Test data are typically not available while the product is in the prototype stage or later in the development. During these early stages, however, another source of reliability information is available—the knowledge of the product experts. We used an award-winning process (PREDICT, 1999) to elicit initial reliability judgments from the product experts. As the product was developed, information from other sources, such as test data, the supplier, and customer, was folded into the reliability characterization using a Bayesian updating approach.

The automotive application has involved lengthy and formal elicitations of teams of experts from the automotive company's national and international sites over several years for several different systems. It has also entailed working closely with automotive personnel to develop a core group with expertise in the elicitation process.

2.1.2 Probability Example: Comparing Expert and Trainee Performance Predictions

The focus of this example of probabilistic elicitation is the performance of an aging defense technology. The goal of this study was to elicit both expert and trainees' predictions on how the technology would perform as a function of its potential condition. It's performance was defined in terms of a metric and it's condition, in terms of its closeness to the original design specifications (for example, ranging from meeting the design specs to being potentially magnitudes outside of them). In this study, the experts themselves were interested in whether/how their predictions would differ from those of the trainees, particularly because at some point the trainees would become the reigning experts. While the experts had mentored the trainees, the trainees lacked the field training of the experts and the experts expected this to lead to differences in their predictions.

Performance data was sparse or nonexistent, especially for potential conditions outside the original design specifications. Thus, the participants' estimates ranged from being based on limited test data, calculations and simulations, for conditions approximating the design specs, to being entirely based on their subjective judgment, for conditions greatly diverging from the specs. The participants gave their judgments as subjective probability estimates with uncertainty ranges during short interviews of about an hour. The participants also described their sources of information, their assumptions in reaching their estimates, the names of their mentors, and their years of experience. The number of participants was small, about seven, because this was the number of knowledgeable persons.

2.1.3 Fuzzy Example: Identifying Radioisotopes

This fuzzy elicitation involves creating an instrument to correctly identify radioisotopes from their gamma ray spectrum. Gamma-ray spectrum are detected indirectly by the ionization they produce in materials. Measurements of the ionization are recorded as a pulse-height distribution. Because gamma-ray spectra can be measured only indirectly, experts must try to identify imprecise features of the pulse-height distribution and match these to precise features of radioisotope spectra. (Please note that we will be using *spectrum* and *pulse-height distribution* interchangeably throughout this paper.)

Identifying radioisotopes is useful to customs agents or law enforcement officers who must deal with suspicious packages. For example, customs agents must verify that packages contain the radioisotopes they are purported to contain and not some other radioisotope that is being shipped illegally. Radioisotopes such as technetium-99 and iodine-131 are routinely shipped to medical institutions.

In contrast to the automotive application, the gamma ray-application has involved only informal elicitations of a few experts. These experts are largely eliciting their own fuzzy rules using texts on fuzzy logic to understand how rules should be formulated and texts on spectroscopy, as well as their own experience, to create the content of the rules. They are selecting the best rules by testing them against a large database of radioisotope spectra.

2.1.4 Fuzzy Example: Comparing Experts on Performance Predictions

The focus of this fuzzy example is similar to the one described above on the expert-trainee study, that is, the reliability of an aging defense technology. However, in this example, two experts were asked to supply their fuzzy rules for predicting how hypothesized conditions of the technology could affect its performance. The goal of this pilot project was to quantify performance, largely based on expert judgment, in the absence of test and other data. In addition, this project was to address a question posed by project sponsors: how they should interpret and handle cases where the only information available was subjective expert judgment and the experts differed.

The fuzzy elicitations, in contrast to those of the expert-trainee study, were intensive, taking about 20 hours per expert and over the course of two years. The two experts were interviewed separately and then were brought together in a structured interview situation to review their sources of information, fuzzy rules, assumptions, and uncertainty ranges. They were allowed to amend their judgments—their fuzzy rules, assumptions, and uncertainty ranges—as they so wished. The experts' judgments were summarized and displayed side by side to facilitate comparison.

2.2 Summary Table of Phases and Steps

A summary table of the elicitation phases and steps is given below. Some phases and steps are performed differently for fuzzy and probabilistic elicitations; these are prefaced with an asterisk. For example, *Eliciting and documenting the expert judgment*, involves eliciting the fuzzy rules in the former, and obtaining probability responses, in the latter. Other phases or steps vary according to the situation, such as how the experts are selected. Note that the same fuzzy examples are not used throughout Table 2.1.

Table 2.1. Elicitation Phases, Steps, and Examples.

Phases, Steps	Probability Example: Auto Reliability	Fuzzy Example: Radioisotopes
1. Determining whether expert judgment can be feasibly elicited.	Feasibility indicated by prior (informal) use of experts' judgment.	Feasibility indicated by prior use of expert judgment.
2 . Determining whether expert judgment can be better elicited in a probabilistic or fuzzy framework	Experts thought in terms of numeric likelihoods; the mathematical foundations of subjectivist probabilities were a plus.	Incoming information was imprecise; one advisor expert preferred fuzzy for the quick creation of a robust expert system.
3. Designing the elicitation		
1. Identify the advisor expert(s).	One self-identified advisor expert identified additional advisors at the national and international levels.	One advisor expert volunteered himself and identified another advisor.
2. Construct representations of the way that the experts measure or forecast the phenomena of interest.	Representations included reliability block diagrams, reliability success trees, and failure modes.	Representations focused on features evident in plots of gamma-ray spectrum and of the second derivative of the spectra.
*3. Draft the questions. For fuzzy, this involves: identifying the variables, identifying the inputs and outputs to the system, and disaggregating the inputs and outputs into distinct linguistic variables.	What is your expected, number of incidents per thousand vehicles to fail to meet specifications? Best case number? Worst case number?	What are your fuzzy rules concerning a peak and these linguistic variables: *low*, *medium* and *high energy* and *very very good, very good*, *good*, *somewhat good* or *somewhat somewhat good*?

Phases, Steps	Probability Example: Auto Reliability	Fuzzy Example: Comparing Experts
*4. Plan the interview situation	Team interviews because the experts worked in teams.	Separate interviews followed by structured joint interviews.
5. Select the experts	The advisor selected the auto products for reliability characterization, which determined the selection of teams, already composed of experts.	The advisor identified the two locally-available and recognized experts.
6. Motivate participation by the experts	The advisor carefully drafted the formal request for participation by cover memo and followed up with telephone calls.	The motivation of participation by the advisor was very informal because this was an in-house effort and there were only two experts.
7. Pilot test the questions and interview situation	Extensive pilot tests of the sets of questions and the cover letter (for motivating participation) were performed via teleconference calls.	Pilot tests of the questions were conducted on the advisor expert and led to refinements in how the fuzzy rules were elicited.
***4. Eliciting and documenting the expert judgment**	Advisor and those he designated lead the team interviews, elicited and recorded the subjective probability estimates, assumptions, and failure modes.	The researchers elicited and documented the experts' fuzzy rules, membership functions, the information, and assumptions the experts considered.
***5. Representing the expert judgment for the experts' review and refinement**	Teams' performance estimates were represented as probability distributions. Teams reviewed the probability distributions and updated their estimates as new information became available.	The researchers and experts refined the fuzzy rules and membership functions. The experts refined their fuzzy rules, in structured joint interviews. The experts' reviews led to labels and caveats being placed on their expert judgment.
6. Facilitating the comparison of multiple	Comparisons were done between proposed designs	We compared experts' fuzzy rules, assumptions,

experts' judgments	and options for testing, instead of between experts' judgments.	qualifications, and the difference to the bottom line in using one expert's judgment over another.

2.3 Phases and Steps

Phase 1: Determining Whether Expert Judgment can be Feasibly Elicited
To answer the question of whether expert judgment can be feasibly elicited, consider the following:

- The domain. *Recommendation*: Most domains of science and engineering are amenable to eliciting expert judgment. If, in addition, there are articles on expert judgment, or detailed instructions passed on by word of mouth, it bodes well for eliciting judgment. *Recommendation*: Domains that may **not** be amenable to the elicitation techniques described in this paper are those in which the experts must quickly respond to control a physical process and are unable to explain their responses, even in retrospect. Jet pilots performing flight simulations would be an example (Shiraz and Sammut, 1998).
- The capabilities of the individual experts. Some individuals are less able than others to articulate their thinking. Generally, individuals can describe their thinking if descriptions are elicited while the problem is at the forefront of their thinking rather than in retrospect. *Recommendation:* In our experience, about five per cent of the experts have great difficulty in "thinking aloud," regardless of the elicitation (Meyer and Booker, 1991).

Probability Example: In the automotive application, it was determined that expert judgment could be elicited. Not only was there recognition that "engineering experience," another word for expert judgment, was a valuable resource but it was tacitly being used in team discussions of the reliability of new automotive products.

Fuzzy Example: In the radioisotope example, gamma ray spectroscopists had historically given their judgments. For example, an expert might say "that looks like a Bismuth spectrum because there are three well-shaped peaks with about the correct energies".

Phase 2: Determining Whether the Expert's Judgments can be Better Elicited in a Probabilistic Framework or a Fuzzy Framework
Consider (1) whether the expert is accustomed to and able to think in terms of probabilities, and (2) to what degree the knowledge being elicited is imprecise.

- Ask the experts how they represent the technical problem and look at the solutions or results of their work. *Recommendation*: If quantitative representations are absent and the experts describe results in qualitative linguistic terms, the experts may prefer the fuzzy approach. If the results are represented as probabilities, percentiles, confidence intervals, or points on a plot, the experts may be accustomed to thinking in probabilities. While many scientists, engineers, and mathematicians are accustomed to formulating their thinking in quantitative terms, even probabilities, they are still prone to the usual biases, such as inconsistency, broadly defined here to mean that their point estimates of mutually exclusive events do not sum to 1.0.

- Also consider the preference of the main experts in determining whether to elicit in a fuzzy or probabilistic form. If the experts have a strong preference for either one, it is generally best to use their preference. First, ask them their reasons for the preference. If their reasons are based on misconceptions about fuzzy or probabilities (for example, contradict some of the recommendations in this section), resolve the misconceptions and ask them to reconsider their preferences.

Additional considerations are the

- Requirement for an expert system: If the application requires a system to run without input from users or experts, (for example, a control system such as that described in Parkinson et. al., 1998) the fuzzy approach may be preferable.
- Changeability of the representation: If the way that the experts identify, measure, or forecast the phenomena of interest is likely to change greatly through time, a fuzzy framework may be more flexible.
- Requirement for probability distributions: Techniques are available for deriving probability distributions from judgments elicited in fuzzy form (Booker et al., 2000; Parkinson et al., 1999; Smith et al., 98 and 97), so this requirement does **not**, by itself, necessitate a probability elicitation.

To check your selection of probability or fuzzy elicitation, ask the experts if they would be able to respond in the form you have selected.

Probability Example: In the automotive application, the auto engineers thought in terms of numeric likelihoods (probabilities) of systems succeeding or failing. Also, the researchers had used a subjectivist probability method on a similar reliability application, and were asked by the automotive sponsor to tailor it to this application. Additionally, the statistical rigor and defensibility of the subjectivist approach appealed to the sponsor.

Fuzzy Example: In the application whose goal was to create an instrument to identify radioisotopes from their gamma-ray pulse-height distributions, the main experts had already considered both probability and fuzzy techniques. They had selected fuzzy and had begun to self-elicit their rules in the fuzzy framework before our first meeting. Their reasons for preferring fuzzy were valid: they were creating an expert system for pattern recognition, and the inputs to the expert system were likely to be imprecise.

Phase 3: Designing the Elicitation

Involve the experts in the steps described below to ensure that the steps reflect the experts' way of thinking about the technical problems.

Step 1: Identify the advisor experts (also known as "champions").

Look for one or two individuals who are knowledgeable about their domain and their culture; who can provide "entree" into their culture, explain its workings, provide guidance on the below-mentioned aspects of the elicitation, motivate wider participation by the experts; and who are willing to act as advisor experts. Often the advisor experts will be the same individuals who initially contacted you.

In our work, we identified the advisor experts, almost after the fact, when they began to push the work forward in their work groups or companies. Once the advisors have been identified, it is helpful to ask them privately what they would personally like to gain from participating in this work and how they will judge its success or failure.

Probability Example: In the automotive application, the individual who was to become the advisor expert volunteered himself when we described the role of the advisor; he worked side by side with us (over the telephone) to conduct the steps below and to involve additional advisors from the company's national and international sites. This advisor defined success in terms of his company adopting the process for characterizing reliability as their new way of doing business and applying it to new products. This advisor defined success in terms of: 1) developing the approach to the point where the reality of higher reliability products could be demonstrated in the field, and 2) his company applying the approach to all new product development programs.

Fuzzy Example: In the identifying radioisotopes application, the first expert who contacted us volunteered to act as an advisor expert. He, then, involved the principal investigator of this project as another advisor expert. The advisor experts wished our elicitation to lead to some additional rules that would help them distinguish valid peak shapes within a particularly confusing energy region.

Step 2: Construct representations of the way that experts measure or forecast the phenomena of interest.

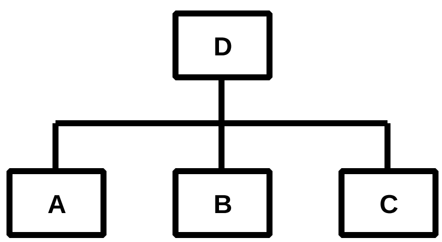
The experts' organization may already have an officially accepted representation of the phenomena of interest, or the experts may have tacit understandings of what the phenomena are. If the representations do not already exist, they can be created in interaction with the experts. (Note that this step may be done in parallel with step 3.) While these representations are not absolutely necessary to conducting elicitations, they are highly desirable if the goal of the work is to form a common basis of understanding or effect a change in the way of doing business (for example, making decisions). Additionally, the representations provide a mechanism for incorporating all available information and a framework for displaying the results of the expert judgment.

Explain the need for the representation to the advisor, define it, and ask the advisor if such a thing exists. If the advisor is unsure about what you are requesting, or whether such a representation exists, ask: 1) for examples of the information that the experts have on the problem; 2) what information they receive, in what form, and from whom; and 3) what they give as a product of their expertise. *Recommendation:* In our experience, discussions about representing the phenomena of interest have led to detailed explanations of the problem by the advisor, often taking days. We recommend allowing time for these explanations of the problem because they will lead to better representations of the phenomena.

Probability Example: In an automotive application, the representations reflected the goal of the project—characterizing the reliability of new automotive products during their developmental programs (Kerscher et. al., 2000). The representations already existed in part and were further elaborated on. The representations had a particularly important role in this application: they provided the common language between the auto employees and us—the "roadmap" for doing business—and the mechanism for incorporating new information as it became available.

The representation focused on the automotive products whose reliability was being modeled. Because the reliability of a product depends on the reliability of its parts—component, subsystem, and system—this step involved representing these parts and their logical relationships in models. These models took the form of reliability logic flow diagrams, typically, reliability block diagrams or reliability success trees. For example, Figure 2.1 (Kerscher et al., 2000) shows a simple generic subsystem D that is composed of components A, B, and C. If the components A, B, and C in the diagram are all in series, the reliability of subsystem D will be the product of the reliabilities of the components.

Figure 2.1. Reliability success tree diagram.



Fuzzy Example: In the application for identifying radioisotopes, the key representations were plots of features of gamma-ray spectra.

To learn about the representations in this application, an advisor expert was asked to describe the information the experts have available and how they give their expert judgment. The information available to the experts, the gamma spectroscopists, is pulse-height distributions of the observed gamma rays, the detector response functions, and libraries of photo peak energies associated with specific radioisotopes. The pulse-height distribution is used to identify the observed gamma-ray peaks. The observed peaks are compared to the detector response function to determine if the observed peak is consistent with the detector response or is due to statistics or noise. The experts identify all the features in a pulse-height distribution; that is, features other than peaks are present, such as Compton edges, which are due to the detector response. These additional features are identified as "not a peak" and are eliminated as peaks using fuzzy membership. The library is the knowledge base upon which a particular pulse-height distribution is categorized. For example, an expert might say "that looks like a Bismuth pulse-height distribution because of the observed peaks, and the extra features look like the Compton edges associated with the observed peaks."

In addition, an advisor expert was asked to roughly list the steps that he anticipated the expert system would perform. Theses steps are: 1) identify "peak shape" from that which is "not a peak;" 2) compare peak energy to library energy using fuzzy membership, 3) tally the peak matches for all isotopes in the library; and 4) determine the best match, to identify the isotope.

Step 3: Draft the questions.
As a starting point for drafting the questions, ask the advisor "What are the phenomena (variables) of interest, how do you assess these, what metrics or natural language terms do you use?" The endpoint of this step differs for fuzzy and probability elicitations, as illustrated in Table I: for a probability elicitation, it leads to the creation of technical questions that the experts will later answer; for a fuzzy elicitation, it provides the linguistic variables that the experts will use as building blocks in constructing their fuzzy rules. (For details on linguistic variables and fuzzy rules, see Ross, 1995). *Recommendation:* For a fuzzy elicitation, it helps to ask the advisor expert to identify the variables of interest, identify the inputs and outputs to the system, and disaggregate these inputs and outputs into linguistic variables. Note that variables that are to be handled as fuzzy or crisp may emerge at this point. For instance, if the linguistic variables do not have a fuzzy continuum of values but one set value, these may be treated as crisp. For example, one of the early determinations made by the fuzzy expert system for identifying radioisotopes involves a crisp value. The expert system essentially asks "Are there enough counts in a specific energy region to say whether there is a feature, such as peak shape. The crisp value is defined to be three standard deviations of net counts above the background as determined by adjacent energy regions. *Recommendation:* In drafting the questions for probability elicitation, include consistency checks (for example, that the mutually exclusive events sum to 1.0).

***Probability Example*:** In the automotive application, the advisor described how the design and process (manufacturing) engineers thought about product reliability or performance. The design engineers thought in the metric of incidents per thousand vehicles failing to meet specifications. The process engineers thought in the metric of parts per million. The advisor further described how the experts thought in terms of what caused the product to fail, or its "failure modes."

We drafted separate questions for the design and process engineers. For example, the design engineers were asked to estimate the number of incidents per thousand vehicles that they expected that the specifications would fail to be met. The design engineers were also asked to provide a subjective range on this estimate. We elicited the range by asking them for a reasonable worst case and best case number of incidents. We also asked the engineers to provide textual descriptions of the potential failure modes for the product and to estimate their likelihood (Kerscher, et al., 2000;1999; 98; PREDICT, 1999).

Once we developed the basic questions for those working in design engineering, we modified the questions for those working in process engineering. Later, we modified the questions even more for those working in software engineering, that is, on automotive parts run by software).

Fuzzy Example: To draft the questions to elicit the fuzzy rules for this radioisotope application, the advisor expert was asked to do the following.

- Identify the variables of interest. The advisor named peak energy, peak width, peak shape, and peak area as the variables of interest and added that the detector response function had several variables—energy calibration, efficiency, and resolution—that could affect these variables.
- Identify the inputs and outputs to the system. The advisor considered the information coming into the detector and its output (the detector response function) and provided the following list of inputs to the expert system (see Table 2.2 below).

Table 2.2. Inputs to the Expert System.

Library Data	Detector Response	Observation
photon energy	calibration	peak energy
	resolution	peak width
	Compton scattering	peak shape
intensity	efficiency	peak area

The advisor listed the following (see Tables 2.3. and 2.4. below) as outputs that he would like the expert system to provide for expert and novice users, respectively. For expert users, the last row lists the possible isotopes—Barium 133, Xenon 133, and Iodine 131—as determined by a fuzzy step, followed by a curve fitting step. The fuzzy step, labeled “Fuzzy ~Match” involves matching the shape and energy between the observed gamma-ray spectra and the library of the photon peak energies. In this case, Barium 133 is the most likely radioisotope because it has the best (highest) match with Barium’s peak energy. For novice users, the expert system will only display the best match between the observed gamma-ray spectrums and the library of photon peak energies, in this case Barium 133.

Table 2.3. Sample Outputs for Expert Users.

Peak Energy	Peak ~Shape	Barium 133 Energy	Barium 133 Attribute	Barium 133 Fuzzy ~Match	Xenon 133 Energy	Xenon 133 Attribute	Xenon 133 Fuzzy ~Match	Iodine 131 Energy	Iodine 131 Attribute	Iodine 131 Fuzzy ~Match
82.8	1.00	53.2	UNLIKELY	0.00	81.0	MUST_HAVE	1.00	80.2	MIGHT_HAVE	1.00
157.7	0.56	81.0	WILL_HAVE	1.00				284.3	MUST_HAVE	0.87
279.0	1.00	160.6	UNLIKELY	0.56				364.5	MUST_HAVE	0.97
306.9	0.96	223.2	UNLIKELY	0.00				637.0	MIGHT_HAVE	0.00
359.3	1.00	276.4	MIGHT_HAVE	1.00				722.9	MIGHT_HAVE	0.00
389.3	1.00	302.0	MUST_HAVE	0.96						
689.7	0.26	356.0	MUST_HAVE	1.00						
		383.9	MIGHT_HAVE	0.93						
PEAK NUMBER = 5.78		ISOTOPE NUMBER= 5.44			ISOTOPE NUMBER= 1.00			ISOTOPE NUMBER = 2.84		
		Barium133 RESIDUAL= 0.34			Xenon133 RESIDUAL= 4.78			Iodine131 RESIDUAL = 2.94		
		Barium 133 MATCH = 0.94			Xenon 133 MATCH = 0.17			Iodine 131 MATCH = 0.49		

Table 2.4. Sample Output for Novice Users.

Best Match
Barium 133

- Disaggregate these inputs and outputs into distinct linguistic variables. The linguistic variables for the inputs of the observed peak energy could be *low*, *medium*, and *high*, and the output linguistic variables for the fuzzy peak shape membership could be *very very good*, *very good*, *good*, *somewhat good* or *somewhat somewhat good*.

Step 4: Plan the interview situation.

Consult the advisors on the interview situation that will fit their culture, their way of thinking and doing business. Advisors will probably suggest whatever situation has worked well in the past. While it may seem obvious to tailor the interview to the experts' culture, we have observed many researchers who attempt the reverse. They try to fit the experts to a particular kind of interview that they, the researchers, prefer. This approach often results in the experts' not participating or the undermining of the credibility of the project.

Describe the main ways in which experts can be interviewed, and ask the advisor whether

- the experts are to arrive at consensus, such as in a team meeting, or not (for example, such as when their judgments are later to be combined by statistical means);
- a problem is likely to arise with experts unconsciously or unwillingly adjusting their own judgment to match others' judgments (Meyer and Booker, 1991). While this bias is not usually a problem in our applications, it can occur, usually in interview situations in which the participants are not of equal status (for example, managers and their employees, military officers and their staff, mentors and those they have mentored);
- the experts' names are to be associated with individual judgments or whether individual judgments are to be anonymous;
- the expert judgment is to be provided on paper (for example, in response to written sets of questions), during face-to-face meetings, or by electronic means; or
- the researchers or the experts themselves are to document the expert judgment.

At this step, you will often still have questions about which interview situation is best. We recommend that you

- ask the advisors which interview situation they recommend;
- consider using a combination of interview situations (for example, initially interview the experts separately, then bring them and the records of their elicitation together, and allow them to amend their judgments); or
- pilot-test the interview situation and let the results answer any remaining questions.

Recommendation: It is easier to conduct detailed and lengthy interviews with one expert at a time. For this reason, experts are typically interviewed separately, at least initially, for fuzzy elicitations. Also, there is often only one expert locally available, anyway. *Recommendation*: It's generally best to interview the experts in the same situation in which they work (for example, if the experts usually work individually, interview them separately; if they work as teams to arrive at judgments, interview them as teams). *Recommendation*: If it is likely that the experts will unconsciously or unwillingly adjust their judgments to those of other experts, it is best to elicit their judgments separately (Meyer and Booker, 1991). *Recommendation*: The experts should document their own judgment, if they will be updating their judgments through time (for example, as in the automotive application). Otherwise, the researchers should do it because they tend to be more thorough and because it relieves the experts of the burden.

Probabilistic Example: On the automotive application, the advisor explained that the experts worked in teams to design and manufacture the products, that team members typically met face-to-face, and reached some consensus in their product planning. The advisor was not unduly concerned about the possibility of group think—having team members unconsciously acquiesce to a dominant members' decision. However, to minimize the chances of this bias occurring and to maximize the diversity of the judgments, we decided to have the team members individually self-elicit their own judgments. The plan was to have team members individually complete worksheets asking for their expected, worst case and best cases estimates of incidents per thousand vehicles. Also on the worksheets, the team members would list the failure modes and their associated likelihood of occurrence. They were to bring these worksheets to the face-to-face meetings of their team. Each team was to be led by the advisor, or the advisor's designee, who would record the team's judgments on flip charts for the team's review.

***Fuzzy Example*:** On the expert comparison application, we knew that experts had been consulted formally and informally about what a particular real or potential "condition might do to the technology's performance." We planned to interview the experts separately because they were often consulted individually. Also, there were two indicators that our interviews of the two experts should, at least initially, be separate: first, fuzzy rules would be elicited which meant the interviews would be intensive; and second, one of the experts had earlier mentored the other which could have meant that the newer expert was prone to group think bias.

We planned to interview both experts separately for their fuzzy rules and membership functions and to also record their assumptions and sources of information. We would then bring the experts and their judgments together in structured interviews where the experts could view each others' responses and amend their own as they wished. We also planned to monitor the structured joint sessions for signs of the bias, such as the newer expert deferring to the other expert. (For further information on this bias, see Meyer and Booker, 1991, p. 134-5).

Step 5: Select the experts.

The advisor selects the experts to be elicited or advises on the selection strategy (for example, whether other experts are to be selected on the basis of publications, experience, the organization or work group to which they belong, and/or their availability). This step varies more according to the circumstances of the elicitation, rather than whether the elicitation is fuzzy or probabilistic. When few experts are available or the application is in-house, the process of selecting the experts can be informal. For example, in the study comparing expert and trainees in their performance predictions, the advisor expert recommended selecting individuals with a range of years of experience and mentors. The advisor then provided their names, information on their years of experience, and who had mentored them.

Probabilistic Example: The advisor selected the auto products for reliability characterization, which in turn, determined the teams that would be interviewed. There were teams for each component in a subsystem or system. The teams were typically composed of four to eight experts who saw their part of the auto product through its development cycle.

Fuzzy Example: In the expert comparison application, the advisor identified the locally-available and recognized experts. There were only two experts, one of whom was the advisor.

Step 6: Motivate participation by the experts.
Ask the advisor whether problems will arise in getting the experts to participate, and if so, how best to motivate participation. For example, in the expert-trainee comparison, the advisor thought experts would be more likely to participate, if they knew elicitations would only take an hour. The advisor talked to the individuals he had mentally selected and encouraged them to participate. Given that these individuals were incredibly busy, we probably could not have obtained participation by other more formal means. In essence, the advisor, through his standing in this culture, motivated the participation of about seven individuals.

If problems are anticipated, or if the experts have to be motivated formally (as is often the case if they are employed by industry), ask the advisor

- how the request should be delivered: verbally (in person or via telephone), by hard copy memo or electronic communication, or by some combination of these;
- from whom the communication should come and which letterhead should be used;
- the order in which the communication will be routed to possible participants; and
- the timing of the communication (for example, before or after a meeting describing this endeavor).

Show the advisor the checklist of things (Meyer and Booker, 1991, pp. 90–92) that individuals typically want to know in deciding whether they will participate: how they were selected, who is sponsoring the effort, how long it will take, what tasks they will perform, and the anticipated product of the effort and their access to it. This information is usually provided to the experts in a cover letter or e-mail requesting their participation.

Probability Example**:** In the automotive application, we were outsiders, unfamiliar with the culture, and thus relied on what the advisor thought would motivate participation. The advisor carefully drafted the request for participation using the checklist mentioned above. The request for participation was a cover memo followed by a series of telephone calls. The advisor also apprised the participants of the progress of the project, in particular, how well their initial reliability judgments predicted the later test data. Receiving this information helped motivate a large number of experts to participate over several years.

Fuzzy Example**:** In the expert comparison situation, this step was informal because the work was done being done in-house and only two experts were involved, one of whom was the advisor expert. The advisor expert was responsible for motivating his and the other expert's involvement in this effort.

Step 7: Pilot-test the questions and the interview situation.
The pilot test provides the last check on the elicitation design before it is conducted. Aspects of the elicitation that need pilot testing are the experts' understanding of the technical question, the response mode—fuzzy or

probabilities—and any directions, such as how to complete the set of questions. If the expert judgment will serves as input to another process, such as decision making, the decision makers should be included in the pilot tests to ensure that the judgments are on the desired phenomena, at the right level, and in the needed form.

Pilot tests are conducted on the advisor expert and on any other experts or users that the advisor recommends. Pilot tests involve having the selected individuals answer the draft questions, with one major addition. The pilot testers are to "think aloud" as they go through the elicitation to allow the researchers to pinpoint problems in the elicitation, such as where the metrics caused confusion (Meyer and Booker, pp. 155–56). *Recommendation:* Pilot testing is generally a good idea if you will be interviewing more than a few experts or intensively interviewing more than one expert through time.

Probability Example: On the automotive application, pilot tests were necessary because about forty design engineers would be interviewed, and they were to be followed by process, and software engineers. We conducted, during conference calls, extensive pilot tests of the sets of questions (work sheets) and the cover letter for the design engineers. The advisor called other advisor experts, introduced us and the idea of pilot testing to them, and listened in on the pilots, as a means of learning how to conduct these. Later, the advisor conducted the pilot tests of worksheets himself, when these needed to be tailored to the process and software engineers.

Fuzzy Example: On the expert comparison application, pilot tests were performed because the fuzzy elicitations were expected to be intensive. (The elicitations later averaged 20 hours per expert.) Pilot tests of the questions were conducted on the advisor expert and led to refinements in how the fuzzy rules were elicited (for example, the linguistic variables describing conditions and performance).

Phase 4. Eliciting and Documenting the Expert Judgment

This stage basically involves administering and documenting the questions as designed and pilot-tested in the earlier stages. Note that this phase differs for the fuzzy and probabilistic elicitations. For fuzzy elicitations, this phase involves eliciting the fuzzy rules, while for probabilistic elicitations, it involves obtaining probability responses to the technical questions. *Recommendation*: Try to keep interviews to about an hour in length so as not to tire the experts. If it appears that that the interview will run over the allotted time, ask the experts if they wish to continue or to schedule another interview.

Probabilistic Example: In the automotive application, the advisor and those he designated lead the team interviews, and elicited and recorded the subjective probability estimates, assumptions, and failure modes.

Fuzzy Example: In the expert comparison application, the researchers elicited and documented the experts' fuzzy rules, membership functions, their information sources, and assumptions.

Phase 5. Representing the Expert Judgment for the Expert's Review and Refinement

The representations mentioned in step 2 (phase 3) can form a framework for documenting the expert judgment. Review the experts' judgments with the experts and refine these. In probabilistic applications, the refinements are likely to be to the probability responses and to caveats concerning their use. In fuzzy applications, the refinements tend to focus on the fuzzy rules, membership functions and caveats. *Recommendation:* If a membership function is very narrow, this may be the time to redefine it as crisp. *Recommendation:* If it is not possible to do the reviews in person, send the transcriptions of the expert's judgment to each expert for review. We recommend setting a date by which they are to respond and explaining that you will assume that they accept the judgment as is, unless you hear otherwise.

Probabilistic Example: To probabilistically represent the expert judgment in the automotive application, we transformed the teams' initial performance estimates into probability distributions using Monte Carlo computer simulations. The teams' estimates referred to their part of the automotive product, such as a component. The probability distributions for each part of the automotive product were combined according to how the parts fit together, as represented in step 2, (for example, see Figure 2.1). This combining of distributions provided an overall reliability characterization for the auto product at a point in time and a means for determining if the target reliability was reached. These reliability characterizations were given to the advisor, who passed them to the participating experts, for review. The advisor kept the experts apprised of how well their initial reliability judgments predicted the later test data on the product.

The teams refined their judgments through time as test data became available. Equations such as Bayes Theorem were used to mathematically update the teams' performance estimates (PREDICT, 1999).

Fuzzy Example: In the expert comparison application, representation and refinement included working with the experts separately to refine their fuzzy rules and membership functions. As part of the review and refinement, we plotted the results of applying the fuzzy rules, showing what the predicted performance would be for a given condition or combination of conditions. We reviewed these plots with the experts separately as a check that their rules delivered results that were consistent with their expectations. For instance, one membership function was found to be broad and the experts wanted it have a crisp edge with no overlap, so it was redone accordingly.

Then we brought the two experts together in a structured interview situation to share their sources of information and judgments. In some cases, the newer expert had not seen some of the simulations of the other expert and modified his fuzzy rules in light of these. We recorded these changes and then, in a last joint meeting, had the experts review the written records again. Then, later as articles were written about this project (Meyer et. al., 1999), the experts again reviewed manuscripts.

As a result of these reviews, we refined our presentation of the results. The experts had expressed concern that the plotted results, particularly for the more extreme hypothetical conditions, might be misinterpreted as coming from frequentist statistics; that is, from statistically sampled experimental data, which as mentioned earlier did not exist. To address this concern, the labels of *subjective* or *expert judgment* were applied to all results and a caveat was added cautioning individuals on the origin and recommended use of this information.

Phase 6. Facilitating the Comparison of Multiple Experts' Judgments

If there are judgments from multiple experts, decision makers or application sponsors may want to evaluate for themselves whether the experts provided significantly different judgments, and if so, the basis for these differences.

We recommend side-by-side comparisons, in which the user can compare different experts' judgments (fuzzy rules, membership functions, probabilities, subjectively-estimated uncertainties, assumptions, and sources of information) and if appropriate, the expert's qualifications (years of experience and why they were selected for participation). If possible, provide information on whether the use of one expert's judgment over another makes a difference to the bottom line. In addition, it is often helpful to have statements in the experts' own words as to whether and why they thought that they differed.

Probabilistic Example: In the automotive application, experts were not compared because teams were the unit of elicitation, and there were not duplicate teams whose judgments could be compared. Instead, comparisons were done between proposed designs, such as if a component were made of aluminum as opposed to plastic, to see what the resulting reliabilities would be. These comparisons were called *what ifs* and were also done to anticipate which components should be prototyped and tested.

Fuzzy Example: In the expert comparison application, we presented a side-by-side display of the experts' judgments because the comparison of expert judgments had been of particular interest to our sponsors. The experts' rules were shown side by side in simple tables to allow viewers to visually compare them. In addition, a summary table was given of: the experts' description in their words about whether they were basically in agreement or not, the assumptions they made, their qualifications, and the main areas of disagreement in the fuzzy rules. Note that the experts' assumptions were given because these have been found to drive the experts' answers (Ascher 1978, Booker and Meyer, 1988) and because decision makers may choose one expert's judgment over another, depending on whose assumptions they most agree with. The expert's qualifications, in this case, years and type of experience, were presented because this information is of interest to decision makers, particularly when the experts give differing judgments.

Finally, a table (Table 2.5) was presented which showed whether the use of one expert's judgment over the others', would make a difference to the bottom-line answer for a particular condition. (Note that the performance scale for Table 2.5 is from 0.0 to 1.00, 1.00 meaning meets performance specifications.) Our reason for showing that there was, in this case, no difference to the bottom line is that it simplifies matters. For instance, it allows the decision makers to forgo the difficult tasks of selecting one expert's judgment or of mathematically combining the judgments.

Table 2.5. Subjective Performance Ranges Based on Each Expert's Fuzzy Rules and a Hypothetical Condition.

	Median	Range
Expert 1	1.00	(.92, 1.00)
Expert 2	1.00	(.96, 1.00)

3. Summary

The driving consideration in eliciting expert judgments in fuzzy or probabilistic forms is the expert's preferences. Key points in performing the elicitation are as follows:

- Using the advisor, the expert "insider" who can advise on how to conduct the elicitation so as to fit the experts' way of thinking and doing business. Using the advisor expert embodies the principle of "asking how to ask" from cultural anthropology—the idea that the researcher may be an outsider to a culture, unaware of the special dialect and customs, and therefore may need to ask an insider how to ask the questions (Meyer and Paton, 2000).
- Pilot testing, if the judgments of more than a few experts will be elicited.
- Documenting as much as possible the experts' thinking and sources of information, as well as the results.
- Involving the experts in the review, analysis, interpretation, and presentation of the expert judgment. The experts' involvement in, or even ownership of, the process is crucial, particularly if the expert judgment must be elicited periodically to reflect the latest knowledge. If expert judgment will be repeatedly elicited, the researcher should aim to have the elicitations led by 1) the advisor, 2) a core group of trained experts, or 3) the experts themselves, through self-elicitation of their own judgments.

4. Acknowledgements

We gratefully acknowledge our subject matter experts and advisors, particularly William Kerscher of Delphi Automotive Systems, for without their knowledge, no elicitation would be possible. Our thanks must also go to Jerry Parkinson (ESA-EPE) for introducing us to the area of fuzzy sets. We are also grateful for the assistance of skilled interviewers, particularly Laura McNamara (CIC-10) and Jerry Morzinski (TSA-1). For the idea of Table 1, we give credit to William

Murray. Finally, we gratefully acknowledge Vivian Romero (TSA-1) for her excellent work in formatting this chapter.

5. References

Ascher, W., *Long-Range Forecasting: From Crystal Ball to Computer*. New York: Wiley-Interscience, 1981.

Booker, J.M and Meyer. M.A., 1988, "Sources and Effects of Interexpert Correlation: An Empirical Study," LA-UR-87-2998, *IEEE Transactions on Systems, Man and Cybernetics*, **8**, 135-142.

Booker, J. M., R. E. Smith, T. R. Bement, and S. M. Parker, "A Statistical Approach to Fuzzy Control System Methods for Uncertainty Distributions," WAC 2000, Maui, Hawaii, June 2–4, 2000.

French, S., M.D. McKay, and M. A. Meyer, "Critique of and Limitations on the Use of Expert Judgments in Accident Consequence Uncertainty Analysis," Workshop on Expert Judgment and Accident Consequence Uncertainty Analysis, Avifauna, The Netherlands, June 21–24, 1999. Also in *Radiation Protection Dosimetry*. In Press.

Keeny, R. L., and D. von Winderfeldt, 1989, "On the Users of Expert Judgment on Complex Technical Problems," IEEE Transactions on Engineering Management, **36**, 83–86.

Kerscher, W. J., J. M. Booker, T. R. Bement, and M. A. Meyer, "Characterizing Reliability During a Product Development Program," European Safety and Reliability (ESREL) 2000, Edinburgh, Scotland, May 15–17, 2000.

Kerscher, W. J., J. M. Booker, T. R. Bement, and M. A. Meyer, "Characterizing Reliability in a Product/Process Design Assurance Program," invited paper Spring Research Conference, Minneapolis, MN, June 2–4, 1999.

Kerscher, W. J., J. M. Booker, T. R. Bement, and M. A. Meyer, "Characterizing Reliability in a Product/Process Design-Assurance Program," *Proceedings of the International Symposium on Product Quality and Integrity*, Anaheim, California, January 1998, and Los Alamos National Laboratory report LA-UR-97-36.

Meyer, M. A., and J. M. Booker, *Eliciting and Analyzing Expert Judgment: A Practical Guide*, Academic Press: London, 1991 (revised from Los Alamos National Laboratory report LA-11667-MS and NUREG/CR-5424 report).

Meyer, M.A. and R.C. Paton, "Interpreting, Representing and Integrating Scientific Knowledge from Interdisciplinary Projects," *Theoria Historia Scientarium*. Nicholaus Copernicus University: Poland. Forthcoming.

Meyer, M.A., Booker, J.M., and Smith, R.E. and Bement, T.R., "Enhanced Reliability Update," *Weapon Insider* **6:** 1, January/February, 1999, Los Alamos National Laboratory, Los Alamos, New Mexico, 87545.

Ortiz, N. R. , T. A. Wheeler, R. J. Breeding, S. Hora, M. A. Meyer, and R. L. Keeney, "The Use of Expert Judgment in NUREG-1150," *Nuclear Engineering and Design*, **126**, 313-331, 1991 (revised from Sandia National

Laboratories report SAND88-2253C, and Nuclear Regulatory Commission report NUREG/CP-0097 **5**, 1–25.

Parkinson, W. J., R. E. Smith, and N. Miller, "A Fuzzy Controlled Three-Phase Centrifuge for Waste Separation," World Automation Congress, Anchorage, Alaska, May 1998.

Parkinson, W. J., K. W. Hench, M. R. Miller, R. E Smith, T. R. Bement, M. A. Meyer, J. M. Booker, and P. J. Wantuck, "The Use of Fuzzy Expertise to Develop Uncertainty Distributions for Cutting Tool Wear." (LA-UR 99-2787), 3rd International Conference on Engineering Design and Automation (EDA 1999), Vancouver, B. C., Canada, August 1–4, 1999.

"PREDICT—A New Approach to Product Development," (LA-UR-00-543) Los Alamos National Laboratory and Delphi Automotive Systems, 1999 R&D 100 Joint Entry, Los Alamos National Laboratory and Delphi Automotive Systems, March 1999 (M. A. Meyer, J. M. Booker, T. R. Bement, and W. Kerscher III).

Ross, T.J. *Fuzzy Logic With Engineering Applications*, McGraw-Hill: New York, 1995.

Shiraz, G. M., and C. A. Sammut, "Acquiring Control Knowledge from Examples Using Ripple-down Rules and Machine Learning," *Proceedings from the 11th Workshop on Knowledge Acquisition, Modeling and Management*, Banff, Canada, April 18–23, 1998.

Smith, R. E., J. M. Booker, T. R. Bement, M. A. Meyer, W. J. Parkinson, and M. Jamshidi, "The Use of Fuzzy Control System Methods for Characterizing Expert Judgment Uncertainty Distributions," *Proceedings PSAM 4 International Conference*, September, 497–502, 1998.

Smith, R. E., T. R. Bement, W. J. Parkinson, F. N. Mortensen, S. A. Becker, and M. A. Meyer, "The Use of Fuzzy Control System Techniques to Develop Uncertainty Distributions," (LA-UR-97-3087), Joint Statistical Meetings, Anaheim, California, Aug. 10–14, 1997.

Stanbro, W.D. and Budlong-Slyvestor, K., "The Role of Expert Judgment in Safeguards," *Journal of Nuclear Materials Management*, Summer 2000:17-20.