SHELF Methods

The objective of elicitation is to construct a probability distribution to represent, as accurately as possible, the knowledge and beliefs of an expert or group of experts regarding a Quantity of Interest (QoI) X. A SHELF elicitation will always involve two rounds of judgements, each expert first makes their own individual judgements, and then after discussion of these judgements a "consensus" probability distribution is constructed from group judgements. However, depending on the nature of the quantity or quantities of interest, there are a variety of different judgements that they may be asked to make.

We discuss here the judgements that are appropriate for eliciting a probability distribution for a single *continuous* QoI, i.e. one that can take any value in some range.

In eliciting a distribution for a continuous QoI, two important practical considerations arise.

1. *Fitting*. To specify a probability distribution completely would require an impossibly large number of judgements. In effect, we would need to specify the expert's probability that $X \leq x$, for all possible values x that X might take, and when X is continuous this is a very large, or indeed infinitely large, number of probabilities. In practice, we elicit only a small number of quantitative judgements from experts, and then fit a convenient distribution to those judgements.
2. *Methods*. A substantial body of research in the psychology of judgement identifies various common biases that are introduced by use of particular kinds of judgements, and also by the phrasing and sequencing of questions to elicit those judgements. No unique best set of judgements to elicit emerges from this research but SHELF offers four *methods* that are designed to avoid such biases, and can be said to represent best practice in elicitation for continuous quantities. Each method comprises a particular sequence of judgements.

In this document, we describe these four methods, with guidance for the facilitator, explanation of the underlying psychology and discussion of the pros and cons of each method.

We begin by presenting the three methods that SHELF offers for use in the individual judgement phase. For convenience in exposition, we shall suppose that the expert is female, while the facilitator is male.

Plausible limits

All the methods begin by asking the expert for her plausible limits for X, comprising a lower plausible limit L and an upper plausible limit U. These should be such that although it may be theoretically possible for X to lie outside these limits she would regard that as extremely unlikely.

There are two reasons to ask for plausible limits as the expert's first judgement. First, a bias that has often been identified in the psychology literature is over-confidence, whereby the expert does not assign enough probability to extreme, unexpected values of X. Second, this is particularly marked when the expert has previously been asked for an estimate, due to the phenomenon of *anchoring*, whereby a value that is already in the expert's mind assumes too much credibility as a likely value for X.

This is why SHELF begins by asking the expert for plausible limits L and U, and the facilitator is advised to challenge her judgements. See the slide set "Plausible Limits" and the guidance to the facilitator in the notes of those slides.

Tertile method

The Tertile method next asks the expert for her median value for X, and then her tertile values.

The expert's median is a value M such that she judges it to be equally likely that X is above or below M. Her lower tertile T1 and upper tertile T2 should divide the possible range of X values into *three* equally likely intervals – below T1, between T1 and T2, and above T2. Each will have probability one-third. Furthermore, it follows that the two intervals from T1 to M and from M to T2 should be judged equally likely, each with probability one-sixth.

The Tertile method is based on several psychological factors.

- The median and tertiles are known as *quantiles*. If the facilitator were to ask the expert to specify her probability that X would be below some value x, then he would establish x in her mind as an anchor, and this would influence her judgements. By asking instead for the expert to specify values of X corresponding to stated probabilities, he avoids this bias.
- Having the plausible limits L and U stated first, the expert's judgement of M is anchored, but it is anchored on both sides, thereby minimising any possible bias in her judgement of M. Similarly, the judgements of T1 and T2 are also anchored on both sides (by L and M, and by M and U, respectively).
- With the three values T1, M and T2 (in addition to L and U), the facilitator can fit a variety of standard distributions. However, the most widely used families of distributions (such as the normal, beta

and gamma families) have two parameters and so in principle can be fitted from just two values. With three elicited judgements, it is generally not possible to fit any distribution from these families perfectly, so the fitting process involves some compromise. This allows the facilitator to see how well any such distribution fits the expert's judgements, and hence to assess whether any fitted distribution is an acceptable representation of the expert's opinions.

- Judgements of equal probability are easier for the expert to make than assessing specific probability values.

Nevertheless, these are still not easy judgements for an expert to make, particularly the tertiles. See the slide sets "Median" and "Tertiles" (with their notes for the facilitator) for guidance on how the expert may be instructed in making these judgements.

Quartile method

The Quartile method asks the expert for her median and then her quartiles.

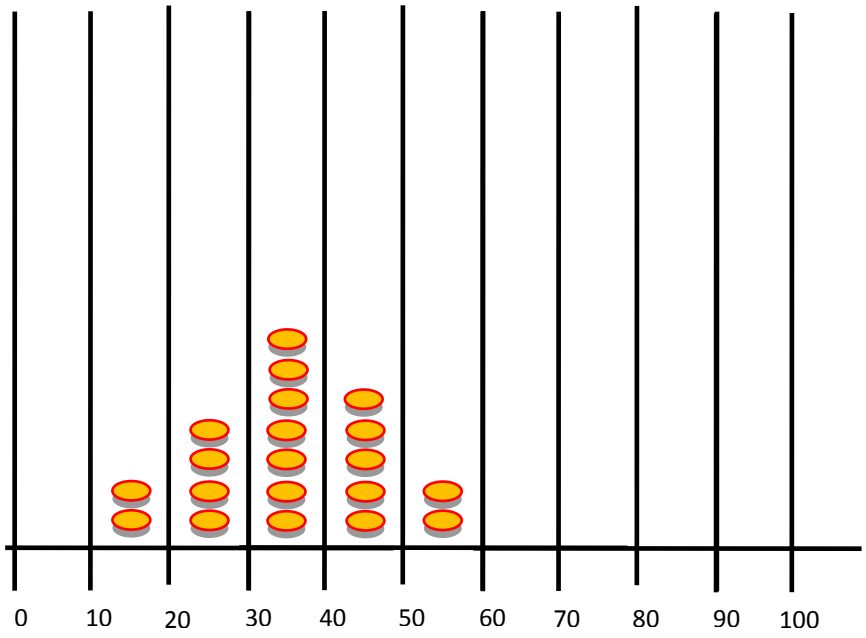
Her lower quartile Q1 is a value between her L and M values such that she judges it to be equally likely that X should be below Q1 or between Q1 and M. Similarly, Q3 is a value between her M and U values such that she judges it to be equally likely that X should be between M and Q3 or above Q3.

The median and quartiles divide the range of possible values of X into four equally likely intervals – below Q1, between Q1 and M, between M and Q3, and above Q3. Each interval therefore has probability 0.25.

The rationale for the Quartile method is essentially the same as for the Tertile method. The slide sets "Median" and "Quartiles" provide guidance on making these judgements.

Roulette method

The Roulette method is quite different. The figure on the next page shows the final result of an expert using the Roulette method. She has placed small counters, called *probs*, on a grid to indicate her probability for the true value of X being in each of a number of intervals, or *bins*. For instance, she has placed four probs in the bin with boundaries labelled 20 and 30. Since she has used 20 probs in all, each one represents a probability of 0.05, and her four probs in that bin mean that she assigns a probability of 0.2 to X being between 20 and 30.



In order to use the Roulette method, the facilitator first gives each expert a bag of probs and a sheet of paper marked with a blank grid. The recommended number of columns in the grid is 10 to 12. The recommended number of probs for each expert is 20 or 25. As with the other methods, the facilitator first asks the experts to specify their plausible limits L and U. He then determines suitable bin boundaries. The lowest bin boundary must be less than or equal to the minimum of the experts' L values, and the highest boundary must be greater than or equal to the maximum of the experts' U values. To make it easier for the experts to think about probabilities, bin boundaries should be rounded to simple numbers if possible.

It is not always necessary to use all the columns on the grid. It is important that there are at least 5 columns in each expert's plausible range, in order for the expert to give a realistic placement of probs. So if the experts differ substantially over their plausible ranges, the full set of 10 to 12 columns should be used, but if there is good agreement over the plausible range then as few as 7 columns could be adequate.

The facilitator then instructs each expert to complete her grid by placing probs to represent her judgements about which values of X are more or less likely. See the slide set "Roulette".

Pros and cons – individual

The facilitator may choose to any of these three methods for the individual judgements stage. They each have advantages and disadvantages.

Tertile and Quartile

The Tertile and Quartile methods are very similar. They share the advantage that they closely respect what is known about the psychology of probabilistic judgements. Their principal disadvantage is that experts find the judgements difficult, particularly the quartiles and tertiles, despite the guidance given in the corresponding slide sets.

For instance, when asking for quartiles, the facilitator needs to make it clear to the expert that Q1 generally should not be placed mid-way between L and M (or Q3 mid-way between M and U). It should be closer to M than to L, and by how close she places it to M the expert is indicating the strength of her knowledge about X. Even with such guidance, this is always a judgement that experts struggle with.

One reason for recommending the Tertile method is that experts may not have so much difficulty placing T1 and T2, although this method has not been used often enough in practice to see whether it has a genuine advantage over the Quartile method. The Quartile judgement task may be easier because asking the expert to divide an interval into three equally probable parts is more complex than asking her to divide into two equally probable parts.

The difficulty of the tertile/quartile judgements often results in the experts making initial individual judgements that do not accurately represent their beliefs, and this generally reveals itself in strange-looking fitted distributions. This is discussed in the document "Facilitator Skills".

Roulette

Experts usually like the Roulette task and find it much less challenging than the median and tertile/quartile judgements, and this is clearly an advantage. However, the Roulette method is in a sense *too* easy. Experts like the way that their probs make a visual representation approximating to a probability density curve, but in practice they may be just making a nice picture and not thinking about probabilities. There is also a psychological bias called the *range-frequency compromise*, according to which experts tend to spread their probabilities too evenly over a set of options (such as the bins in the Roulette grid).

So the Roulette method has the disadvantage that it is more prone to psychological effects than the other two methods.

Probability method

For the group judgements, the facilitator may choose to use the Quartile or Tertile methods, but may also use the Probability method. In this method, the facilitator asks the experts to make group judgements of three probabilities. He first selects a value X1 and asks the experts for their probability that X is less than X1. He then nominates another value X2 and asks for the experts' probability that X is *greater than* X2. Finally, he

chooses a third value X0 and requests the experts' probability that X is less than X0.

The facilitator bases his choices of these three values on the experts' initial individual judgements and the subsequent discussion. X1 should be a relatively low value, such that the facilitator expects the experts to give a probability of about 0.2 or 0.3 that X is less than X1, while X2 is a relatively high value such that he expects the experts to give a probability of about 0.2 or 0.3 to X being greater than X2. The final value X0 should be a value between X1 and X2, such that the facilitator feels the experts might give a probability of around 0.4 or 0.6 to X being below X0.

The reason for offering a switch to the Probability method for the group judgements is that it should encourage the experts to think directly about probabilities. In their individual judgements stage, and during the subsequent discussion, many specific values of X will have been mentioned, so that there is no risk of the facilitator introducing anchors by his choices of X1, X2 and X0.

The sequence of questions, including the switch from asking for the probability of X being below X1 to asking for the probability that it is above X2, is designed to encourage the experts to think carefully about these group judgements.

Pros and cons – group

At the group stage, experts need to reach "consensus" judgements from the perspective of a Rational Impartial Observer (see the document "Facilitator Skills" and the slide set "RIO"). A disadvantage of the Roulette method in this context is the need for the experts to debate the placement of each individual prob, and for this reason Roulette is not an option in SHELF at the group stage. If Roulette has been used at the individual stage, any of the other three methods would be suitable for the group stage.

If either the Quartile or Tertile method has been used for individual judgements, a disadvantage of using these methods at the group stage is that experts are likely to make their judgements by trying to reach compromise values rather than thinking about dividing intervals into equally probable parts. For instance, when asked for their group median, the discussion is likely to focus on their individual medians, and an attempt to reach a compromise/average value. The Probability method at the group stage is proposed to avoid this disadvantage.