The probabilistic elicitation of subjective data

DSTL/TR79234
March 2014

Dstl Portsdown West

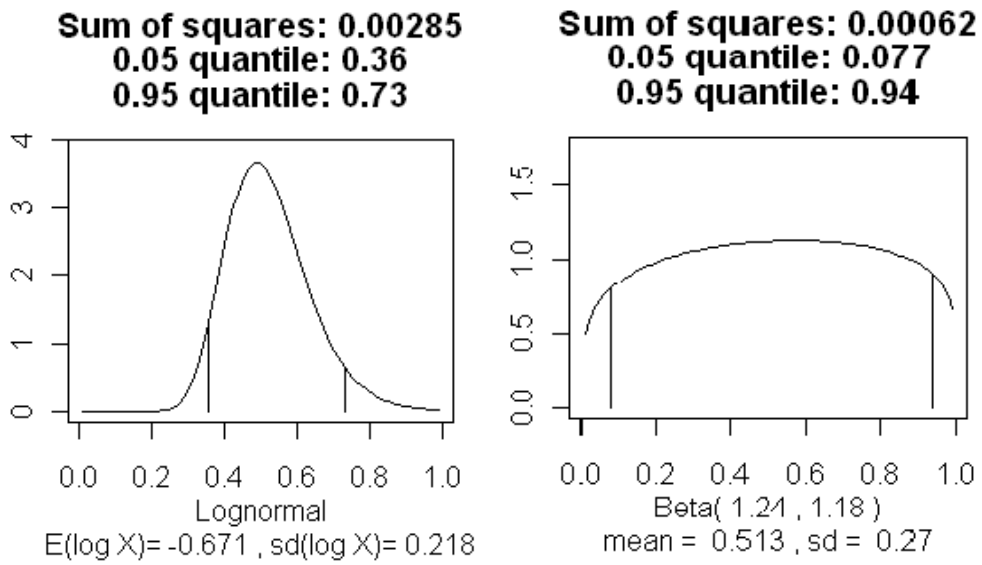
Executive summary

This paper is based on the Defence Science and Technology Laboratory's (Dstl's) best practice advice and has been produced to feed directly into the government's Analytical Quality Assurance book (the cross-governmental book that encourages quality analysis to be delivered by drawing upon the experience gained to share best practice).

Much of Dstl's, and other science and technology organisations', advice is driven by research that utilises subjective quantities, as determined by experts. At present, these subjective quantities are often simply reported as single-point figures; it is not known how confident the experts are in their estimates. Consequently, important decisions might be being made based on estimates for which the experts are very unsure.

The aim of this report is to present established methodology and highlight best practice for representing the uncertainty surrounding experts' subjective judgements by means of probability distributions. If used, decision-makers will have a means of assessing the levels of risk that they are taking by using expert judgement to inform the decision-making process.

To illustrate the importance of the matter: suppose an expert, or a group of experts, is asked to estimate a particular proportion. The diagram below depicts two possible scenarios. In the graph on the left, the expert has estimated a proportion of 0.5 (50%). Similarly, in the graph on the right, the expert has estimated a proportion of 0.5. The horizontal axes on the graphs display the range of possible proportions, from 0 to 1, and the vertical axes represent the likelihood of the unknown quantities being those proportions (do not worry about the differing scales, just focus on the shapes of the graphs). In the graph on the left, the expert is clearly confident that the unknown quantity is around 0.5, yet in the graph on the right, the expert is not confident, and considers that, whilst the most likely proportion is 0.5, it could almost equally likely be any proportion between 0 and 1. Hence, if a customer were presented with the left-hand graph, they would have confidence in using the estimate of 0.5 to inform an important decision. However, if they were presented with the graph on the right, they would know that it would be unwise to use 0.5 to inform an important decision.



In addition to being used to represent the uncertainty surrounding standalone subjective estimates, the techniques considered within this report – termed probabilistic elicitation techniques – should also be used to determine the risk surrounding subjective model inputs and the associated model output(s).

This report introduces methods for representing the uncertainty surrounding subjective judgements.

Table of contents

1 Introduction

1.1 Background

1.1.1 This paper is based on the Defence Science and Technology Laboratory's (Dstl's) best practice advice and has been produced to feed directly into the Analytical Quality Assurance book.

1.1.2 Much of Dstl's, and other science and technology organisations', advice is driven by research that utilises subjective quantities, as determined by experts. At present, these subjective quantities are often simply reported as single-point figures; it is not known how confident the experts are in their estimates. Consequently, important decisions might be being made based on estimates for which the experts are very unsure.

1.1.3 In recent years, techniques for eliciting subjective data from Subject Matter Experts (SMEs) have been developed in academia and have been exploited to great effect within both the public and private sectors. For example, they have been widely used in the design and management of large, complex engineering projects (O'Hagan et al, 2006). Such projects are often essentially unique, so there is very limited experience about the performance of components individually and in combinations. It is therefore natural to draw on expert judgements. In particular, there has been extensive use of elicitation in connection with nuclear installations.

1.1.4 SME elicitation is defined as *'a systematic process for formalising and quantifying, typically in probabilistic terms, expert judgements about uncertain quantities'.* [US Environmental Protection Agency].

1.1.5 For the purpose of this report, SME elicitation, which will be referred to as 'elicitation', will relate specifically to probabilistic elicitation. Elicitation is used when there is no available data, or there is insufficient data, or data that is not fit-for-purpose, that can be used to estimate a particular parameter and its associated uncertainty.

1.1.6 There are specific tools and methodologies for performing elicitations. This paper will explore those tools and methodologies in detail.

2 An overview of the elicitation process

2.1 Context

2.1.1 Suppose an expert was asked to provide an estimate for a particular parameter (assuming that there was no available, or suitable, data that could be used to determine the estimate and its associated uncertainty). Rather than just eliciting the expert's 'best guess' figure, it would be much better to elicit their 'best guess' figure, together with an indication of the confidence that they have in the figure. It could be that the expert is extremely confident about their estimate but, conversely, they might have very little confidence in it. It is important that this uncertainty is captured and taken account of.

2.1.2 Subjective estimates could be required for:-

(i) inputs to models,

(ii) inputs to processes and

(iii) standalone purposes.

2.1.3 Reference (i), elicitation would enable the transition from Figure 1 to Figure 2.

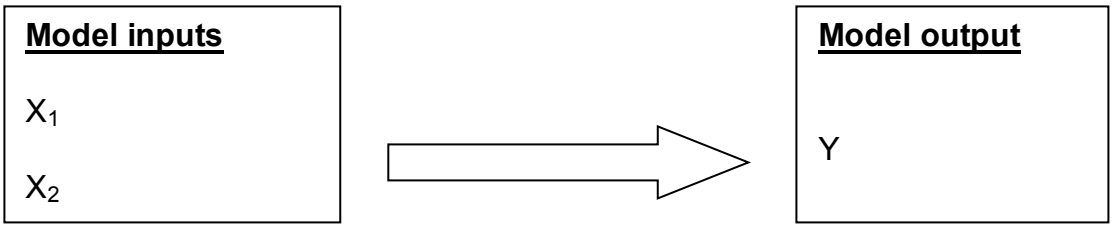


Figure 1: Models using subjective data inputs, without elicitation.

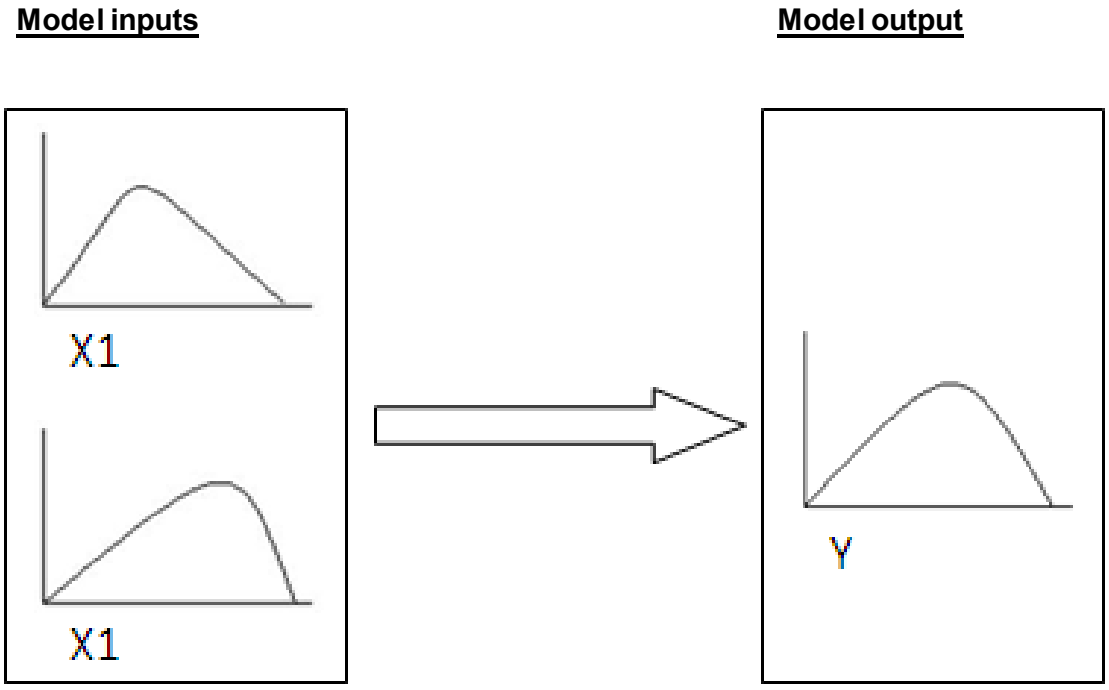


Figure 2: Models using subjective data inputs, with elicitation.

2.1.4 By expressing uncertainty in the input variables, the uncertainty in the output variable can be ascertained. It is important to understand that if the input distributions are based on subjective judgements, the uncertainty surrounding both the model inputs and the model output will be epistemic (uncertainty due to lack of knowledge rather than randomness).

2.2 The eight suggested stages of elicitation

2.2.1 There is a suggested multi-staged approach to elicitation:-

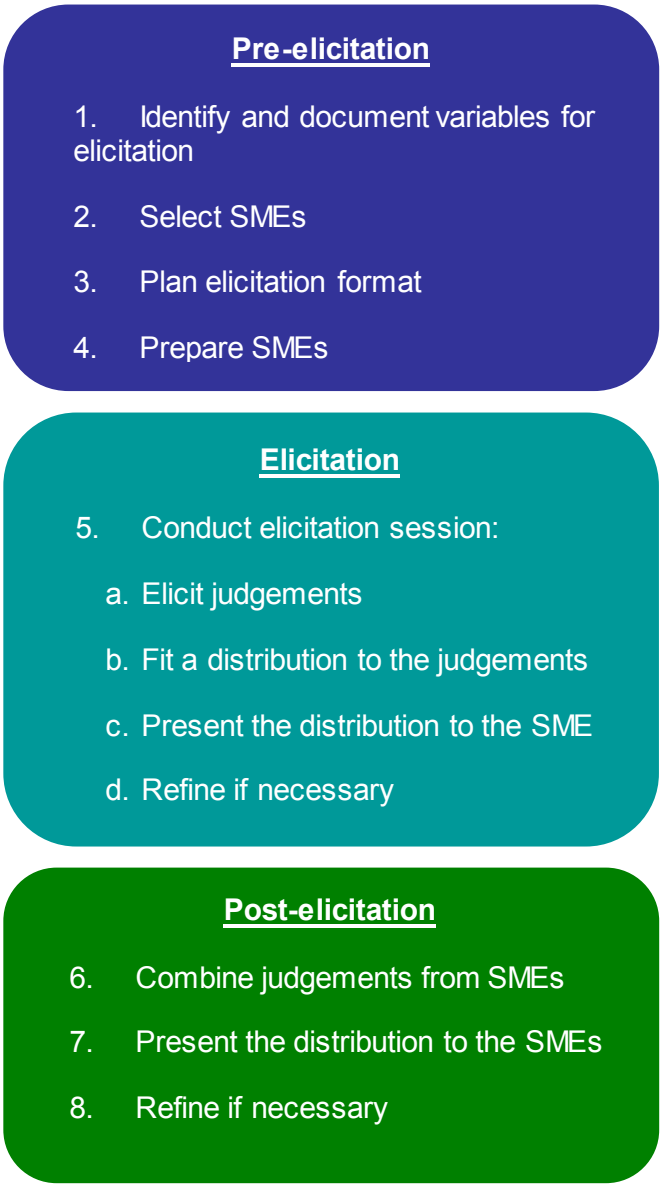


Figure 3: The eight suggested stages of elicitation.

2.2.2 These eight suggested stages were determined during a series of Dstl workshops.

3 Types of values

3.1 Quantitative values

3.1.1 As discussed in the 2006 book 'Uncertain Judgements – Eliciting Experts' Probabilities' (by O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E. and Rakow, T), it is not usual to attempt to elicit measures such as means or variances, or extreme percentiles[1] (i.e. the 1st and 99th) from SMEs, as psychological research (see below) has indicated that people are much more competent at eliciting measures and quantities such as medians, modes and proportions (Garthwaite et al, 2005). It should be noted that this research relates to participants actually seeing samples of data and then being asked to estimate sample statistics. When elicitation is used in Dstl studies, there will be no datasets to visualise. Despite this, the principles discussed below (i.e. the cognitive difficulties involved in estimating means and variances) still apply.

3.1.2 'Uncertain Judgements – Eliciting Experts' Probabilities' talks about experiments that have been conducted that have attempted to investigate peoples' abilities to judge sample proportions (Erlick, 1964; Nash, 1964; Pitz, 1965, 1966; Shuford, 1961; Simpson and Voss, 1961 and Stevens and Galanter, 1957). We are told that Shuford (Shuford, 1961) conducted one particular experiment whereby he projected a series of 20 x 20 matrices onto a screen. Each matrix contained a different combination of red and blue squares and was displayed to the audience for a number of seconds. After displaying each matrix, Shuford asked the audience to estimate the proportion of squares that had been red. In this, and similar experiments, subjects generally assessed the sample proportion very accurately, with the mean of the subjects' estimates differing from the true sample proportion by less than 5% in most cases.

3.1.3 The book then goes on to describe similar experiments that have been conducted to investigate people's abilities at estimating measures of central tendency, such as the mean, median and mode (Beach and Swenson, 1966; Peterson and Miller, 1964 and Spencer, 1961 and 1963). These experiments have usually involved a sample of numbers being displayed to a group of people, who have then been asked to estimate (just by visual inspection) the mode, median and mean of the sample. When the sample distributions were approximately symmetric (and hence the mean, median and mode were numerically similar), the participants' estimates of these measures were very accurate (Beach and Swenson, 1966 and Spencer, 1961). However, an experiment conducted by Peterson and Miller (Peterson and Miller, 1964) used a sample drawn from a population whose distribution was highly skewed[2]. Consequently, the participants' estimates of the median and mode were reasonably accurate, but their estimates of the mean were biased towards the median.

3.1.4 O'Hagan et al then tell us that further psychological research has indicated that people are not very good at understanding what is meant by the term 'variance'. As such, they struggle to assign credible numerical values to it. When estimating relative variability, it has been shown that people are influenced by the mean of the stimuli and hence

[1] Any of the 99 numbered points that divide an ordered set of scores into 100 parts, each of which contains one-hundredth of the total.

[2] A normal distribution that is slanted towards one extreme or the other.

estimate the coefficient of variation[3], rather than the variance. So when the means increase, people's assessments of the variance decrease. Additionally, if a population distribution is bimodal[4], and hence large deviations from the mean predominate, the variance is usually overestimated. On the other hand, if small deviations from the mean predominate (i.e. when the population distribution is normal), then the variance is generally underestimated (Beach and Scopp, 1967).

3.1.5 Further, much empirical research has been conducted to investigate people's abilities to assess the extreme tails of a distribution – and the research has revealed that people struggle to do this. This is largely because it requires the consideration of events that are unlikely, and hence comparisons do not come readily to mind.

3.1.6 To summarise, the extensive research conducted in this area has shown that people can successfully estimate the proportions, modes and medians of samples. However, they are less proficient at assessing sample means if the sample distribution is highly skewed, and they struggle to estimate the tails of distributions. As such, when conducting elicitations, where possible it is sensible to attempt to elicit the former rather than the latter.

3.2 Qualitative values

3.2.1 Within government, it would be of great use to be able to determine the uncertainty surrounding qualitative judgements. For example, in Military Judgement Panels (MJPs) the military experts often have to decide upon things such as how successful a course of action might be, what an actor (a player within a scenario: be it a political figure, warlord, military leader etc.) might do (psychological wrapping) etc. If these qualitative decisions could be 're-framed' to be quantitative judgements, then the experts' uncertainty surrounding these judgements could be represented. For example, if a decision had to be made as to whether a actor might do x or y, then the experts could be asked to say how many times out of ten they think the actor would do x. Given this proportion, their uncertainty surrounding it could then be represented via the elicitation of a probability distribution[5]. Further, when MJPs require SMEs to determine how successful or effective a course of action might be, it would again be possible to frame the question such that their response be quantitative – and hence a probability distribution could be elicited.

The triangular distribution: a note

3.2.2 Before moving to the next section (which describes the various probabilistic elicitation techniques), it is important to mention the triangular distribution. Widely used, the triangular distribution is generated by determining a most likely value, coupled with upper and lower range points, and then joining these points together. Whilst the distribution has utility, the following four methods will enable the user to elicit more detailed information about an expert's uncertainty in virtually the same amount of time.

[3] The coefficient of variation is defined as the standard deviation divided by the mean. It is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other.

[4] A continuous probability distribution with two different modes. These appear as two distinct peaks.

[5] This might seem confusing as the uncertainty surrounding the experts' uncertainty will be being elicited. However, if asked to say how many times out of ten something would happen, the experts might simply not have a clue and hence could pick a figure at random (i.e. five or eight out of ten). If this was the case, the eliciting of a probability distribution for their five or eight out of ten would represent this uncertainty.

4 Four techniques for eliciting univariate distributions from single experts

4.1 Overview

4.1.1 Sheffield University has developed best practice tools and techniques for conducting probabilistic elicitations. This set of tools is referred to as the SHeffield ELicitation Framework (SHELF) and was developed by Professor Tony O'Hagan and Dr Jeremy Oakley from the Department of Probability and Statistics at the University of Sheffield.

4.1.2 The tools run in R and are now also available on the Internet (http://optics.eee.nottingham.ac.uk/match/uncertainty.php). R is a free software environment for statistical computing and graphics and is accessible to anyone who has an internet connection.

4.1.3 There are four SHELF options for eliciting either proportions / medians[6]. Each of these methods is now considered in turn.

4.2 The bisection method

4.2.1 Also commonly referred to as 'the quartiles method', this method for eliciting a median (which can be expressed as a proportion) uses a series of equal-odds judgements. The expert is asked to do the following:-

(i) Determine a feasible range for the unknown quantity. In this example, the SME sets the range as being 0 to 100%.

(ii) Determine a median[7], m, such that $P(\theta < m) = 0.5$ (θ being the unknown quantity). So as not to confuse the SME, they would simply be asked what they thought the most likely proportion might be. In the example below (Figure 4), it can be seen that the SME chose 0.3 (30%).

[6] Other methods involve the elicitation of uniform or triangular distributions but such distributions provide little resolution or information.

[7] The median, also called the second quartile, cuts the ordered dataset in half. The lower quartile cuts off the lowest 25% of the ordered dataset. The upper quartile cuts off the highest 25% of the ordered dataset.

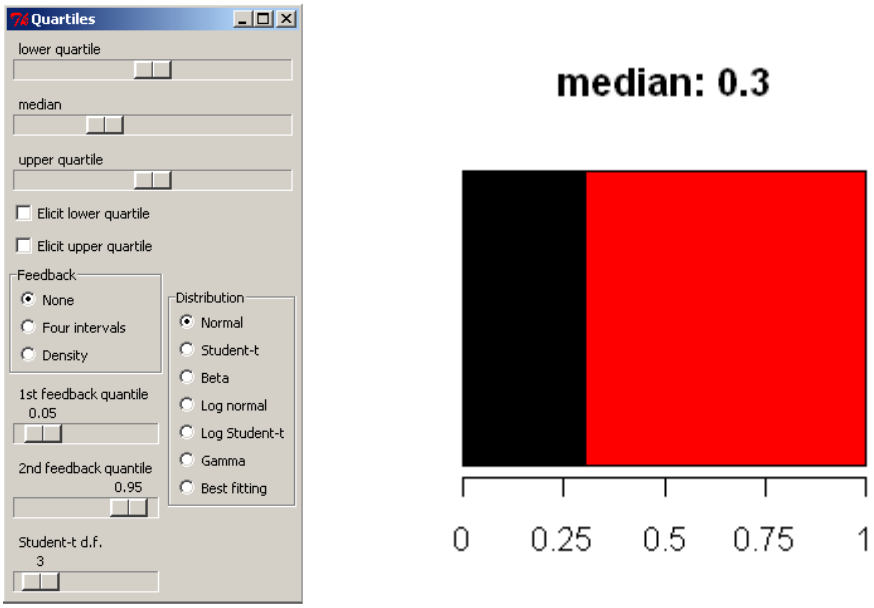


Figure 4: Eliciting a median, bisection method.

(iii) Determine a lower quartile[6], l, such that $P(\theta < l) = 0.25$. Again, the SME would not explicitly be asked to elicit a lower quartile. Rather, they would be asked to set the divider between the green and blue (in Figure 5) to the point where they thought it equally likely that the unknown quantity could reside in either the green or the blue area. Alternatively, the facilitator could ask the SME to place the divider at the point such that, if they had £10, they would be equally happy to use it to bet that θ would lie in the green area as they would to bet that θ would lie in the blue area.

In the example below, it can be seen that the SME chose 0.2 (20%). Hence, they deemed it equally likely that θ might lie between 0 and 20% as between 20 and 30%.

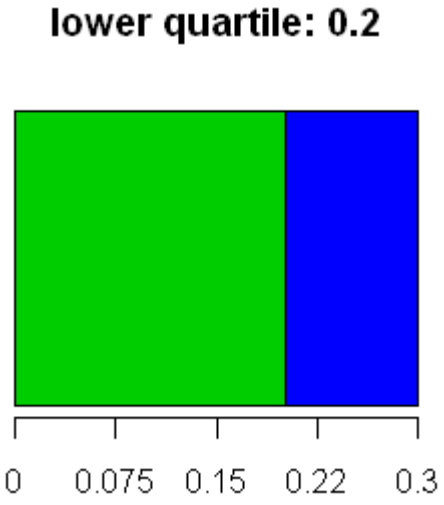


Figure 5: Eliciting a lower quartile, bisection method.

(iv) Determine an upper value[6], u, such that $P(\theta < u) = 0.75$. As in part (iii), the SME would not explicitly be asked to elicit an upper quartile. Rather, they would be

asked to set the divider between the pink and light blue (in Figure 6) to the point where they thought it equally likely that the unknown quantity could reside in either the pink or the light blue area. In the example below, it can be seen that the SME chose 0.35 (35%). Hence they deemed it equally likely that θ might lie between 30 and 35% as between 35 and 100%.

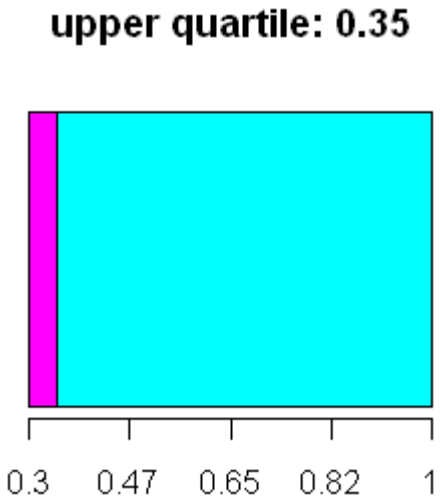


Figure 6: Eliciting an upper quartile, bisection method.

(v) Check for consistency: does the SME believe that the four intervals [0,l], [l,m], [m,u] and [u,1] are equally likely? If not, they may want to modify their median and / or quartiles. Do they consider θ ϵ [l,u] to be as likely as θ∕ϵ [l,u]? Again, if not, they might want to modify their median / quartiles.

4.2.2 The facilitator for the elicitation then determines the distribution that best fits the SME's elicited beliefs. This is done via least squares. Least squares is a mathematical procedure for finding the curve that best fits a given set of points by minimising the sum of the squares of the offsets (also known as the residuals – i.e. the distances of the points from the curve).

4.2.3 When determining the best-fitting curve, SHELF has the following probability distributions at its disposal:-

- **Normal** – *a probability distribution shaped like a bell, often found in statistical samples. The distribution of the curve implies that, for a large population of independent random numbers, the majority of the population often cluster near a central value, and the frequencies of higher and lower values taper off smoothly.*

- **Student-t** – *a probability distribution that is symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to producing values that fall far from its mean.*

- **Beta** – *a family of continuous probability distributions defined on the interval (0,1), parameterised by two shape parameters, typically denoted by α and β .*

- **Log Normal** – *a probability distribution in which the log of the random variable is normally distributed, meaning it has a bell curve.*

- **Log Student-t** – *a probability distribution in which the log of the random variable follows a Student-t distribution.*

- **Gamma** – *a probability distribution with a shape and scale parameter. It is a left-skewed distribution with light tails.*

4.2.4 At least one of these distributions should provide a reasonable best-fit.

4.2.5 Figure 7 shows that in this particular instance the best-fitting distribution was the normal distribution, with a mean of 0.29 (29%) and a standard deviation of 0.11 (11%). This distribution is shown to the SME. The SME is asked to consider the 5th and 95th quantiles[8] and consider why θ is unlikely to be this small and large, respectively. The SME can then conduct further modifications if necessary.

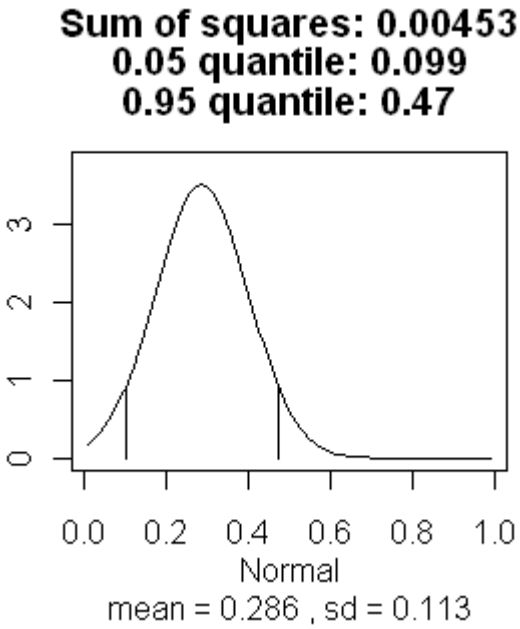


Figure 7: Best-fitting distribution, bisection method.

4.3 The tertile method

4.3.1 This method is similar to the bisection / quartiles method but asks the SMEs to elicit tertiles[9], as opposed to quartiles. Hence, once again, this method for eliciting a median uses a series of equal-odds judgements. The expert is asked to do the following:-

(i) Determine a feasible range for the unknown quantity. In this example, the SME sets the range as being 0 to 50%.

(ii) Determine a median, m, such that $P(\theta < m) = 0.5$. So the SME would simply be asked what they thought the most likely proportion might be. In the example below, it can see be seen that the SME chose 0.2 (20%).

[8] The fifth quantile cuts off the lowest 5% of the ordered dataset. The 95th quantile cuts off the highest 5% of the ordered dataset.

[9] The lower tertile cuts off the lowest 33% of the ordered dataset. The upper quartile cuts off the highest 33% of the ordered dataset.

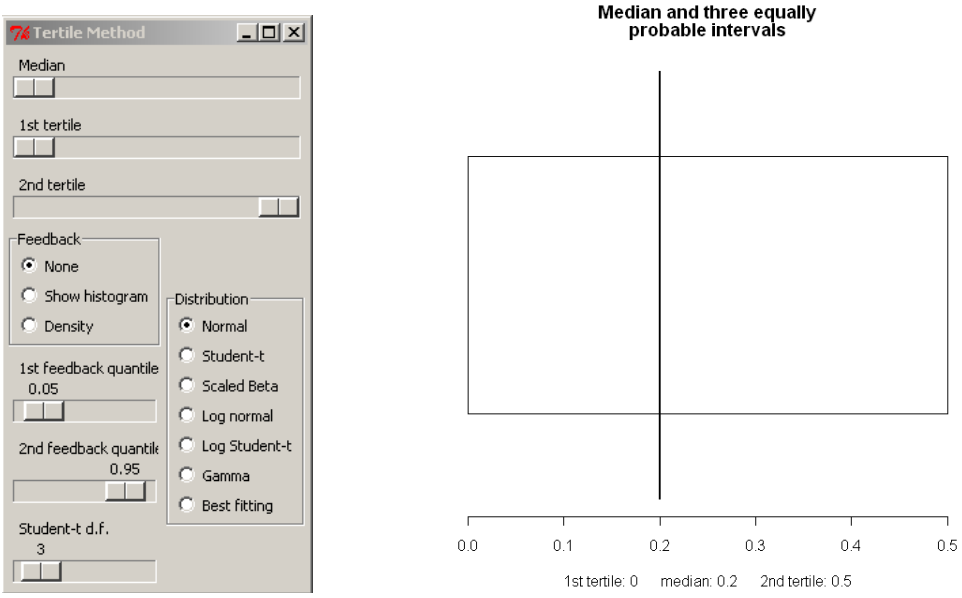


Figure 8: Eliciting a median, tertile method.

(iii) Determine a first (lower) tertile[8], t_1, such that $P(\theta < t_1) = 1/3$ and determine a second (upper) tertile[8], t_2, such that $P(\theta < t_2) = 2/3$. Hence the SME would be asked to determine three equally likely intervals. So, if they had to bet £10, they would be equally happy to use the £10 to bet that θ would lie between their lower range and t_1 as they would to bet that θ would lie between t_1 and t_2, or between t_2 and t_3. In the example below, the expert chose 0.15 (15%) and 0.24 (24%) as their first and second tertiles, respectively.

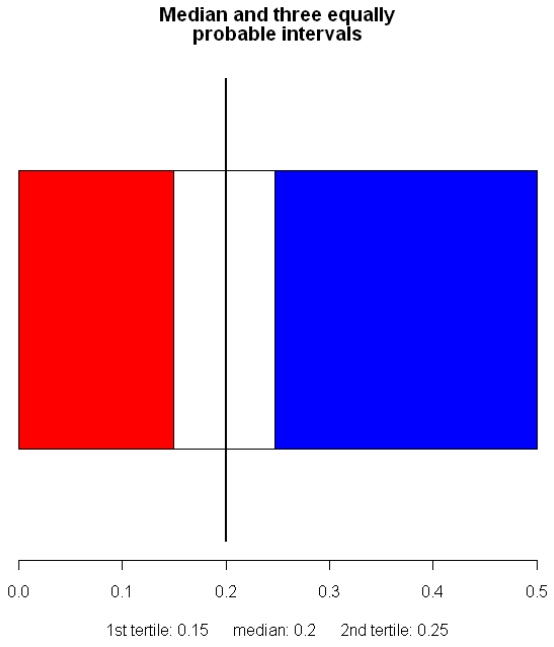


Figure 9: Eliciting lower and upper tertiles, tertile method.

(iv) Perform a final check: does the SME believe that the three intervals $[0,t_1]$, $[t_1, t_2]$ and $[t_2,1]$ are equally likely? So, in the example, does the SME believe that θ is as likely to reside between 0 to 15% as it is to reside between 15 and 25% and 25 and 50%? Additionally, do they believe that θ is as likely to lie between 0 and

20% as it is between 20 and 50%? If not, they may want to modify their median and / or tertiles. It can also be helpful to show the SME a histogram of their beliefs, which is piecewise uniform across the three intervals (Figure 10). On seeing this, the SME might wish to adjust their median or tertiles.

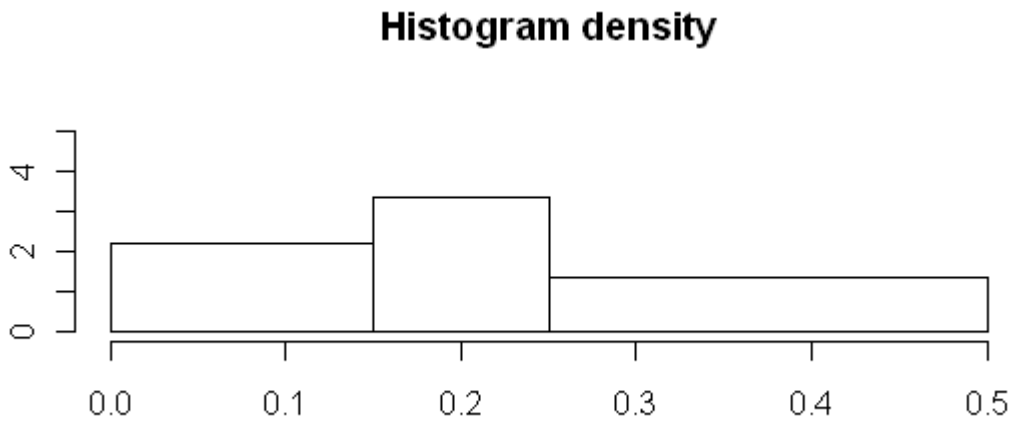


Figure 10: Histogram, tertile method.

4.3.2 The facilitator then determines the distribution that best-fits the SME's elicited beliefs.

4.3.3 In this instance, the best-fitting distribution is the Beta distribution (Figure 11), with a mean of 0.21 (21%) and a standard deviation of 0.10 (10%). This distribution is shown to the SME. The SME is asked to consider the 5th and 95th quantiles[7] and consider why $\boldsymbol{\theta}$ is unlikely to be this small and large, respectively. They can then conduct further modifications if necessary. Once the SME is happy that the distribution accurately represents their beliefs, the elicitation process is over.

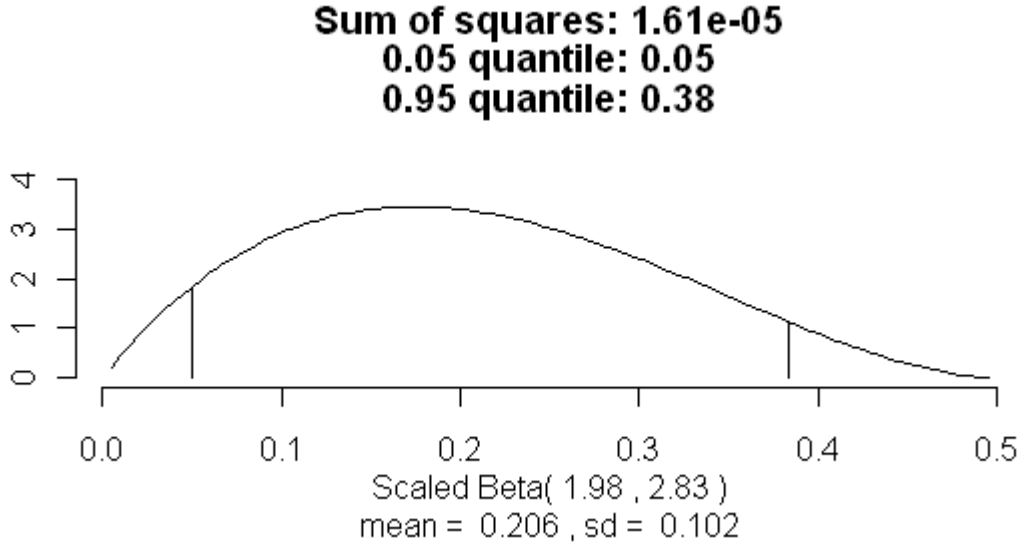


Figure 11: Best-fitting distribution, tertile method.

4.4 The probability method

4.4.1 A third method for eliciting a median from an SME, and capturing their surrounding uncertainty, is termed the 'probability method'. It is also sometimes referred to as the 'hybrid method'. For this method, the SME is asked to:-

(i) Determine a feasible range for the unknown quantity. In this example, the SME sets the range as being 100 to 300 units.

(ii) Determine a median, m, such that $P(\theta < m) = 0.5$. So the SME would simply be asked what they thought the most likely value might be. In the example below, it can be seen that the SME chose 160.

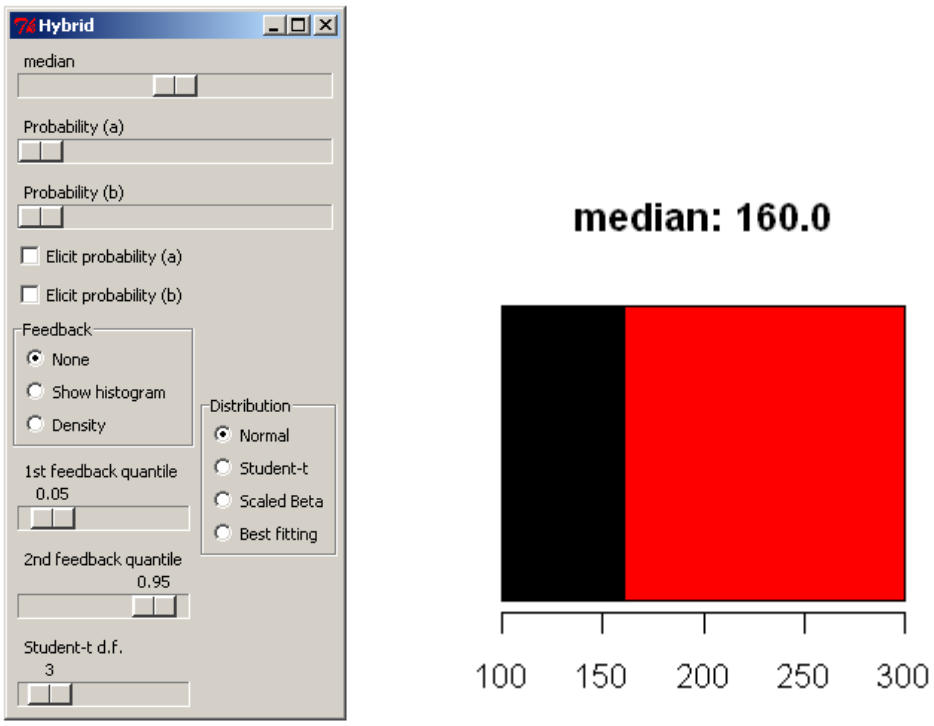


Figure 12: Eliciting a median, probability method.

(iii) Elicit a first probability such that $P(L < \theta < (2m + L)/3)$, where L is the bottom of the range (so in this case 100). Cognitively, this might seem like a difficult task, but the SME would not be presented with this formula; rather, in this particular example, the SME would be asked to estimate the probability that $P(100 < \theta < 140)$.

(iv) Elicit a second probability such that $P((2m + U)/3 < \theta < U)$, where U is the top of the range (so in this case 300). Again, cognitively, this might seem like a difficult task, but the SME would not be presented with this formula; rather, in this particular example, the SME would be asked to estimate the probability that $P(210 < \theta < 300)$.

In the example case, the SME chose first and second probabilities of 0.2 and 0.1, respectively (Figure 13).

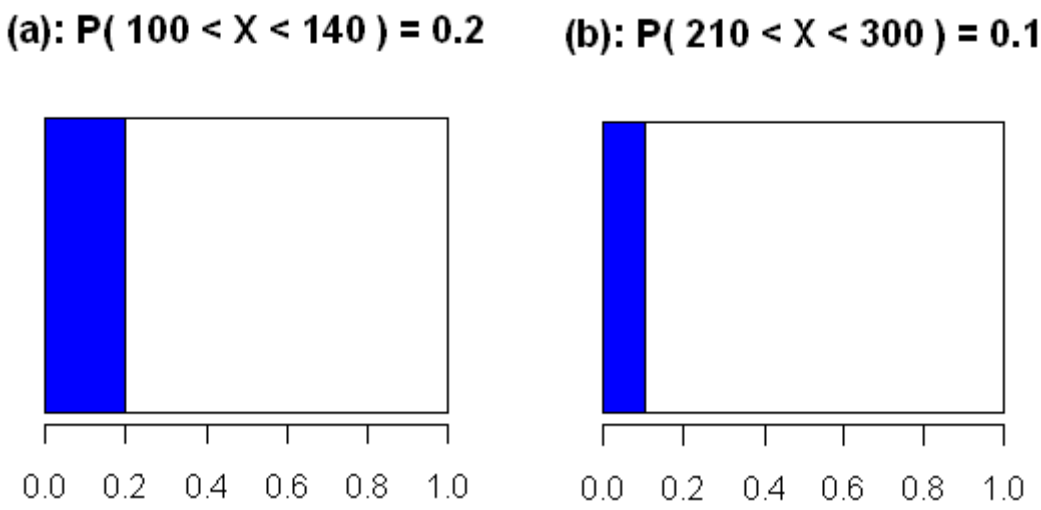


Figure 13: Eliciting probabilities, probability method.

(v) Perform a final check by viewing the histogram of their beliefs and considering its shape (Figure 14). On seeing this, the SME might wish to adjust their median or probabilities.

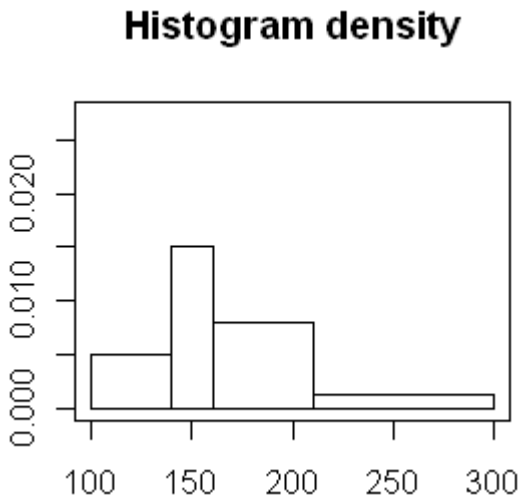


Figure 14: Histogram, probability method.

4.4.2 The facilitator then determines the distribution that best-fits the SME's elicited beliefs.

4.4.3 In the specific example, the best-fitting distribution is the Log Student-t distribution, with a mean of 160.77[10] units and a standard deviation of 1.28[11] units (Figure 15). This distribution is shown to the SME. The SME is asked to consider the 5th and 95th quantiles and consider why θ is unlikely to be this small and large, respectively. They can then conduct further modifications if necessary. Once the SME is content that the distribution accurately reflects their beliefs, the elicitation process is over.

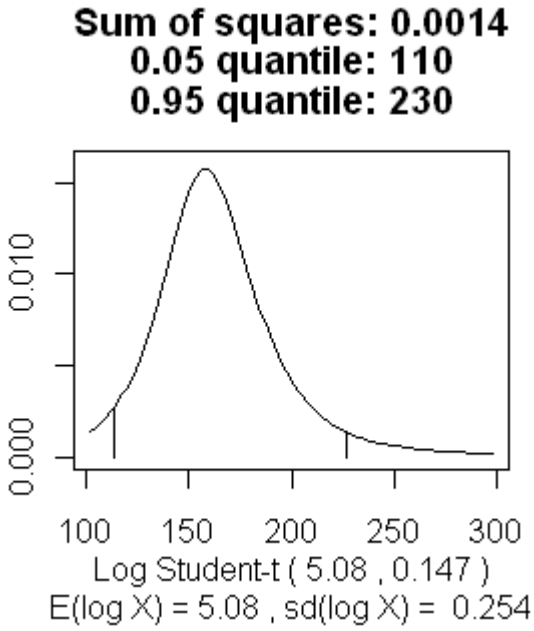


Figure 15: Best-fitting distribution, probability method.

4.5 The trial roulette method

4.5.1 The fourth and final method for eliciting a median from an SME, and its surrounding uncertainty, is termed the 'trial roulette method'. The SME is asked to:-

(i) Determine a feasible range for the unknown quantity. In this example, the SME sets the range as being 0 to 100 units. The facilitator then divides this range into ten equally sized intervals (termed 'bins').

(ii) Allocate 'chips' (as this is the trial roulette method) to bins, so that the proportion of chips in each bin represents the probability that θ lies in each particular bin (see below). It is up to the facilitator as to how many chips are used, but a sensible suggestion would be to use between ten and 20.

[10] The exponential of 5.08.
[11] The exponential of 0.25.

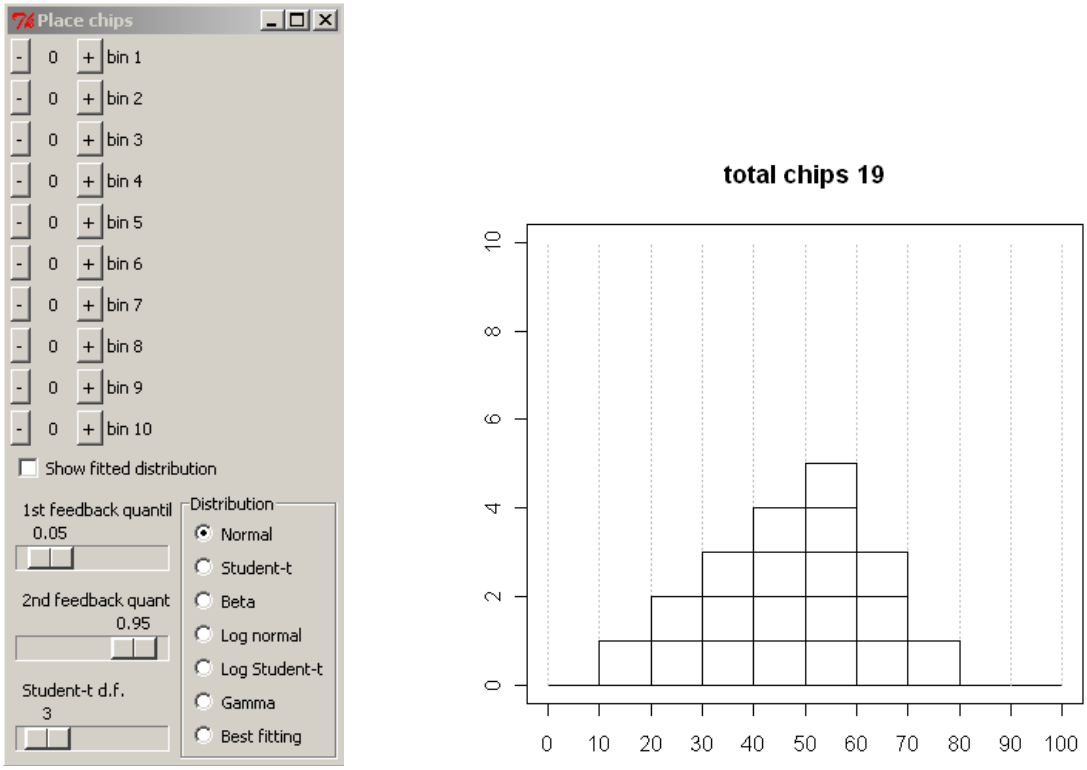


Figure 16: Eliciting expert's distribution, trial roulette method.

4.5.2 The facilitator then determines the distribution that best-fits the SME's elicited beliefs.

4.5.3 In the case study, the best-fitting distribution is the normal distribution, with a mean of 47.6 units and a standard deviation of 15.9 units (Figure 17). This distribution is shown to the SME. The SME is asked to consider the 5th and 95th quantiles and consider why θ is unlikely to be this small and large, respectively. They can then conduct further modifications if necessary. Once the SME is happy with the distribution, the elicitation process is over.

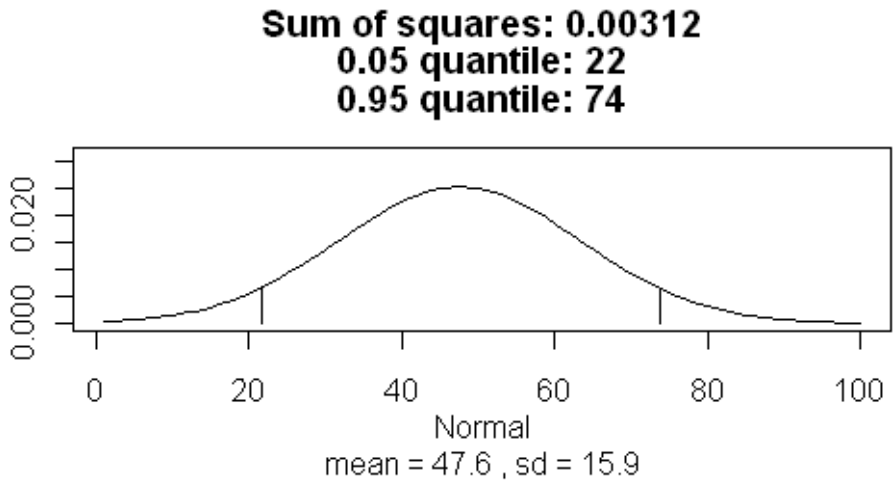


Figure 17: Best-fitting distribution, trial roulette method.

5 Multiple experts

5.1 Aggregation methods

5.1.1 When important decisions or inferences are to be made, it is a good idea to obtain the opinion of more than one expert. Hence it is necessary to obtain a single probability distribution that encapsulates the views of all of these experts. There are two ways to do this: by mathematical or behavioural aggregation.

5.1.2 With mathematical aggregation, a distribution is elicited from each expert, independently of the other experts, and these distributions are then combined mathematically into a single distribution.

5.1.3 With behavioural aggregation, the group of experts strives to reach consensus and then a single distribution is elicited for the whole group.

5.1.4 Academics from both Sheffield and Strathclyde University concur that behavioural aggregation is the best approach to take. Both universities also recommend that distributions are elicited from each expert first (and the results documented). Hence, a quasi form of the Delphi method is being used.

5.1.5 The Delphi method is a formal technique that seeks to get the best from group discussions. It is an iterative procedure that works as follows:-

(i) Each expert's opinion (distribution) is elicited, independently of the other experts.

(ii) The experts are then privy to all the other experts' opinions (distributions).

(iii) In view of what they have now heard, and seen, the experts are invited to revise their initial opinions (and distributions).

5.1.6 Behavioural aggregation is the preferred approach for a number of reasons. First, expertise can be shared and healthy discussions can be had. Second, very 'informed' SMEs have the opportunity to impart their knowledge and experience upon the less well-'informed' members of the group. Third, the arbitrary decision as to which mathematical aggregation technique to use does not have to be made. And, fourth, and perhaps most crucially, behavioural aggregation is a much more subtle form of aggregation than mathematical aggregation.

5.1.7 Tables 1 and 2 summarise the pros and cons of each form of aggregation.

Mathematical aggregation

Pros	Cons
Everyone's beliefs are used	If weightings are not used, a much less 'informed' expert will be making the same contribution as a much more 'informed' expert
More 'informed' experts can be given greater weightings	A form of 'double counting' of expertise could occur if the knowledge of some of the experts overlaps substantially
There is no limit on the number of experts that can be used	It is not clear whose opinion (if anyone's) the resulting probability distribution represents
	Very 'informed' experts do not get the chance to share their experience with the others
	An arbitrary choice has to be made as to which mathematical aggregation technique to use

Table 1: The pros and cons of mathematical aggregation.

Behavioural aggregation

Pros	Cons
Expertise can be shared	If not managed properly, strong personalities might have too much input into the discussion, and subsequent decision-making, and reticent experts too little
Very 'informed' experts have the chance to share their knowledge and experience	The pressure to reach consensus might result in some experts hiding their real (dissenting) views
No arbitrary choice has to be made as to which mathematical aggregation technique to use	It might not be possible to reach a consensus

Allows for more subtle forms of aggregation	The process does not really work if there are more than eight experts
	Sometimes group opinion can be overconfident (as people feel as though they have less personal responsibility)

Table 2: The pros and cons of behavioural aggregation.

5.1.8 Experienced elicitation practitioners have deduced that it is best to have between four and eight SMEs when using behavioural aggregation as, if there are more than eight experts there are too many opinions being shared, which becomes unmanageable. This deduction is supported by a recent internet article (Kelsey, 2009), which concluded that the optimal number of people needed for an effective meeting is between five and nine. The article discussed how a group of two can have insufficient resources, whilst a group of three is often unstable, with one person controlling the others by being the 'split' vote. A group of four often devolves into two pairs. However, with five to eight people, it is possible to have a meeting where everyone can speak out. When the number is higher than eight, not enough attention is given to each person and meetings risk becoming too noisy, too boring and too long – or some combination thereof! It should be noted, however, that the smaller the group, the lesser the statistical power.

5.2 An example of behavioural aggregation using SHELF

5.2.1 Suppose that an attempt were made to elicit a probability distribution, relating to a particular unknown quantity, when there were five SMEs. To begin, probability distributions would be elicited from each expert. These distributions would then be documented. In the example below, the bisection method has been used. The row relating to the probability of 0.25 presents the lower quartiles that were elicited from each of the experts. The row relating to the probability of 0.5 presents the medians that were elicited from each of the experts. The row relating to the probability of 0.75 presents the upper quartiles that were elicited from the experts. The rows pertaining to the probabilities of 0 and 1 represent the bottom and top of the ranges, respectively. Hence, in the example below, all of the experts set the range to be between 0 and 0.5.

	Probability	Expert1.value	Expert2.value	Expert3.value	Expert4.value	Expert5.value
1	0	0	0	0	0	0
2	0.25	0.08	0.15	0.12	0.15	0.085
3	0.5	0.1	0.2	0.15	0.2	0.1
4	0.75	0.25	0.3	0.18	0.3	0.35
5	1	0.5	0.5	0.5	0.5	0.5

Table 3: The medians and lower and upper quartiles for each of the five experts.

5.2.2 On obtaining the medians and quartiles from the SMEs, their elicited views are converted into best-fitting probability distributions (Figure 18).

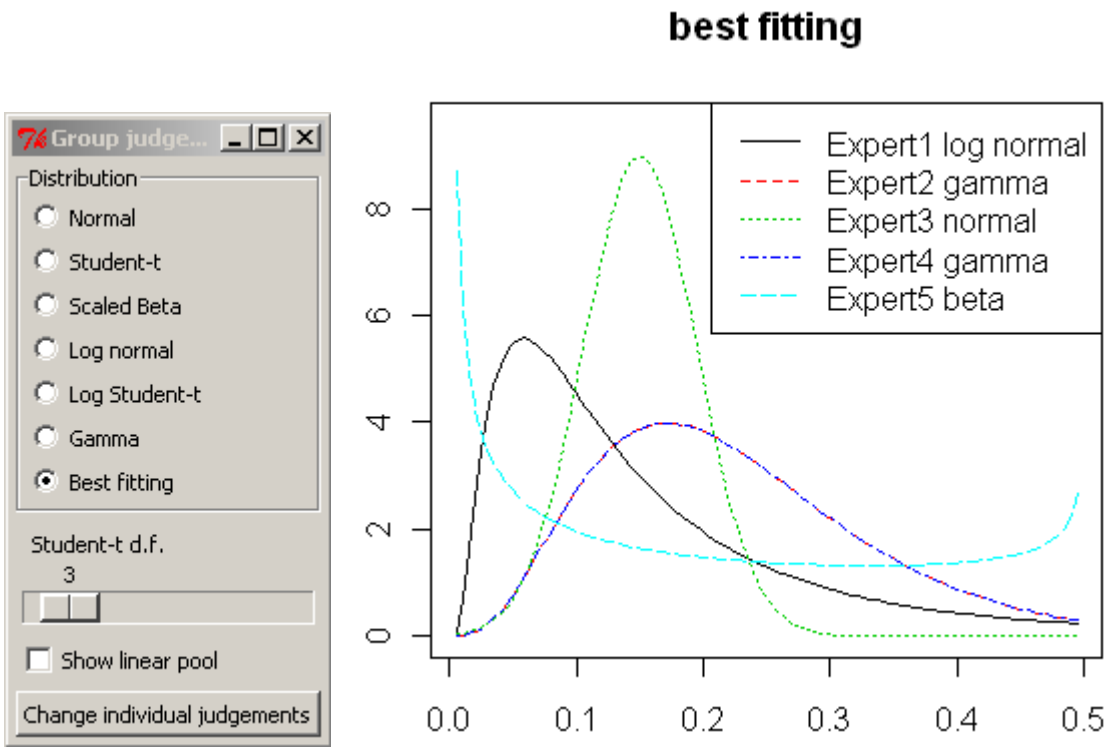


Figure 18: Best-fitting distributions for each expert.

5.2.3 If mathematical aggregation was the preferred option (which it is not), a linear opinion pool (where the sum of the distributions is taken and divided by the total number of distributions) could be used. SHELF only allows the calculation of an equally weighted linear opinion poll (i.e. where the experts have equal inputs).

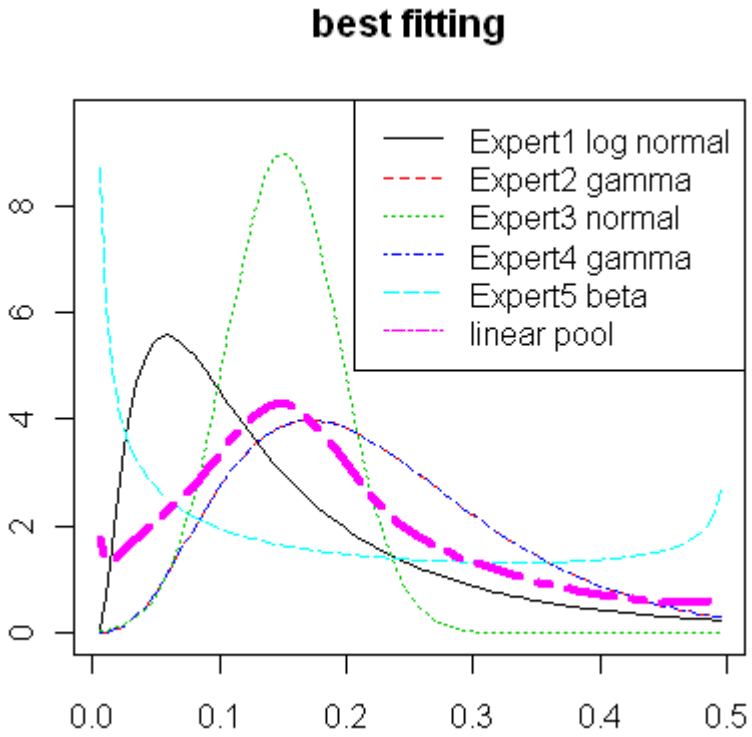


Figure 19: Best-fitting distributions for each expert, with a mathematical aggregation.

5.2.4 Returning to behavioural aggregation: having elicited medians and quartiles from each expert, and having generated the corresponding best-fitting distributions, the experts are now shown all of the other group members' elicited distributions. The meeting chair then facilitates a group discussion. As part of this discussion, each expert explains why they gave the answers they did. Expertise and opinion is shared and the chair aims to get the group to arrive at a consensus with regard to what the median and the quartiles should be.

5.2.5 Suppose, in the example, that, after a well-facilitated session, the group agreed that the median should be 0.15 and the lower and upper quartiles should be 0.10 and 0.25, respectively. The chair would now use these values to generate a single, best-fitting distribution (Figure 20).

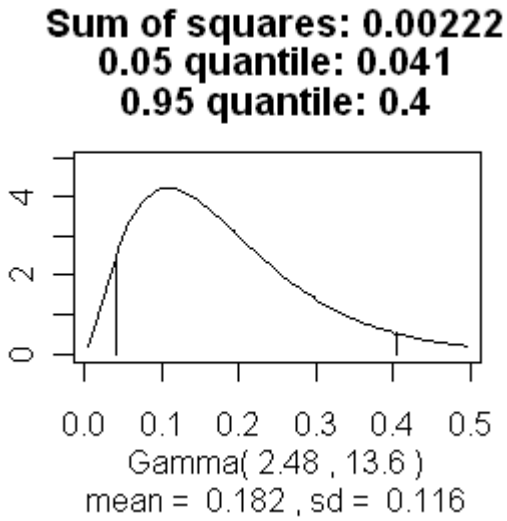


Figure 20: Best-fitting distribution for the group.

5.2.6 The best-fitting distribution is a Gamma distribution, with a mean of 0.18 and a standard deviation of 0.12. The group of experts is shown this distribution and refinements can be made until the group is happy that the resultant distribution accurately reflects their collective view.

6 A case study and trial

6.1 Trial session

6.1.1 In January 2011, a trial elicitation session was held, during which four analysts were asked to estimate the shortest distance between Wembley Stadium and Old Trafford (in miles, by road). The script for the session can be found at Annex A.

6.1.2 The bisection method was used.

6.2 Questionnaire feedback

6.2.1 At the end of the session, the delegates were asked to fill out a questionnaire, which sought to (i) understand if they were clear as to what values they were being asked to provide, (ii) see if it was easy to reach a consensus within their group and (iii) determine whether or not they could see the value of elicitation. The results of the survey are presented below:-

6.2.2 In response to the statement: 'In my one-to-one session, it was clear what values I was being asked to provide', two of the analysts strongly agreed, one agreed and one neither agreed nor disagreed.

6.2.3 Reference the statement: 'It was easy to reach a consensus in the group session', three of the analysts strongly agreed and one agreed.

6.2.4 Two of the analysts strongly agreed with the statement: 'I can see why the elicitation of probability distributions for subjective judgements is important', the other two agreed.

6.2.5 When asked if elicitation would be relevant and beneficial to their areas of work, one analyst strongly agreed, two agreed and one neither agreed nor disagreed.

6.2.6 All four analysts agreed that elicitation could have many applications across the whole of their business area (two strongly).

6.2.7 In terms of the free text responses, the general consensus was that, at first, the analysts struggled conceptually with understanding what was required when asked for the lower and upper quartiles.

6.2.8 In summary, the questionnaire feedback suggested that the session went very well and it also indicated that the analysts could see the utility of the technique. In terms of development, for future elicitation sessions more preparation is needed to ensure that the SMEs are clear as to what exactly is required from then when they are asked for their lower and upper quartiles.

7 Utility

7.1 Standalone estimates

7.1.1 Often, a probability distribution relating to an unknown value will not be used as a model input but, rather, will simply be reported to a customer as a standalone distribution. This standalone reporting has the potential to be hugely informative. It is best to demonstrate how informative it could be by means of an example ...

7.1.2 Suppose that an SME was asked to estimate a particular proportion (given the absence of any data) and that a probability distribution representing the expert's uncertainty was elicited using the bisection method (Section 4.2). Figure 21 depicts two possible scenarios. In both scenarios a median of 0.50 has been elicited. However, in the example on the left, a lower quartile of 0.45 and an upper quartile of 0.60 has been elicited and, in the example on the right, a lower quartile of 0.30 and an upper quartile of 0.75 has been elicited. Best-fitting distributions have then been determined. If the first scenario were reality, the distribution on the left would be sent to the customer and the customer would have confidence that the expert was fairly sure that the proportion would be around 0.50. However, if the second scenario were reality, the distribution on the right would be sent to the customer and the customer would see that, whilst the median proportion was 0.50, it is also almost equally likely that the proportion could be anywhere between 0 and 1. Hence the customer would see that the expert had very low confidence in their estimate of 0.50. Given the scenario on the left, the customer would have reasonable grounds to use the SME's beliefs when making important decisions, yet, given the scenario on the right, it would be unwise for the customer to place much impetus on the SME's beliefs. If the second scenario was reality, and a probability distribution had not been elicited, the customer would just receive an estimate of 0.50 and hence would be unaware of the expert's extreme uncertainty surrounding this estimate. As a consequence, the 0.50 could be used to inform important decisions.

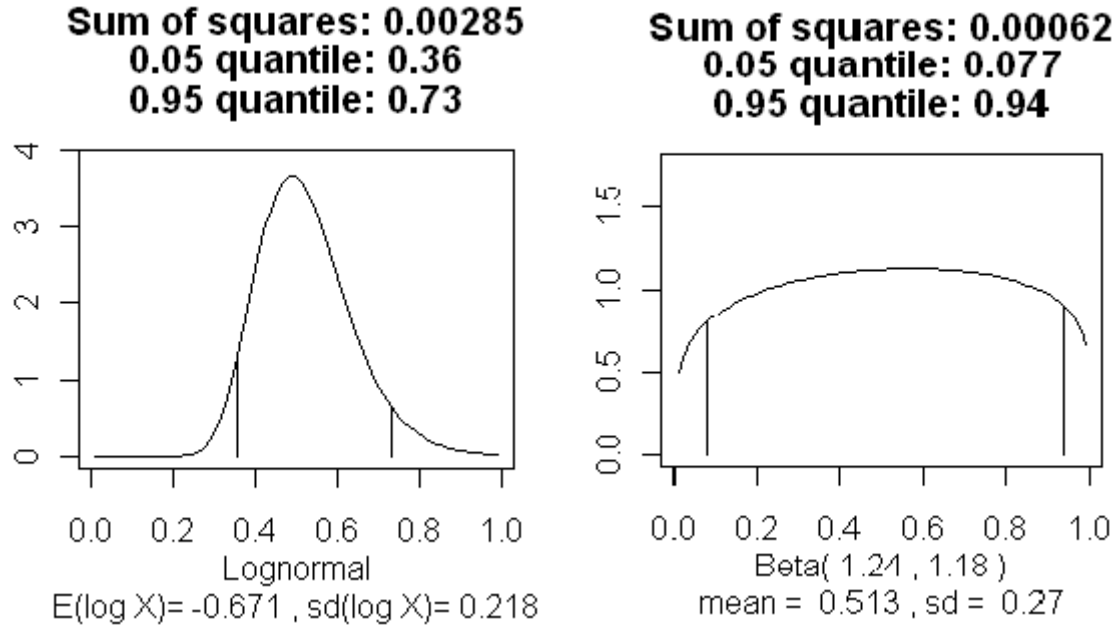


Figure 21: Two probability distributions with medians of 0.50.

7.1.3 It could be argued that, instead of using formal elicitation techniques, an SME could simply give a 'best guess' figure, along with a verbal indication as to how confident they are in their 'guess' (i.e. *'I'm pretty sure that this is about right'*). There are a number of reasons why this course of action would be unwise. First, what does 'pretty sure' actually mean? One person's understanding and interpretation of 'pretty sure' is likely to be significantly different to another person's. With elicitation, the subjectivity surrounding qualitative assessments of uncertainty is not an issue. Second, a customer might, unwisely, simply take the SME's 'best guess' figure and choose not to report how confident they indicated they were in their estimate. With elicitation this would not be an issue as probability distributions such as those displayed at Figure 21 would be reported.

7.2 Model inputs

7.2.1 Section 7.1 discusses the use and presentation of elicited probability distributions when considering standalone subjective estimates. However, often, probability distributions will be used as model inputs. For example, imagine a model with two independent subjective input parameters, X and Y, and one output parameter, Z, such that X + 2Y = Z. Probability distributions for X and Y could be generated using one of the elicitation methods described in Section 4. These elicitation methods would also provide means and standard deviations for the input variables. A Monte Carlo simulation[12], or other appropriate simulation technique, could then be performed, using suitable software, to generate an output distribution for Z. An appropriate percentile (e.g. the 70th) could then be easily identified from this cumulative distribution.

[12] A method that estimates possible outcomes from a set of random variables by simulating a process a large number of times and observing the outcomes.

8 More complicated elicitations

8.1 The elicitation of an unknown parameter, and its associated probability distribution, has now been considered. However, suppose there was a requirement to elicit a joint probability distribution (for two or more uncertain quantities) from experts. The simplest of all joint probabilities would be the level of association between two variables. However, things now get complicated. It would be unwise to ask the experts to simply estimate a correlation coefficient, as even people with statistical backgrounds struggle with this (Morgan et al, 1990; Kadane et al, 1998 and Gokhale et al, 1982). Therefore, something called a quadrant probability is elicited. This is done as follows:-

(i) Suppose there are two variables, X_1 and X_2, for which the medians, M_1 and M_2, are elicited from the expert(s).

(ii) An attempt is made to elicit the probability that both variables exceed their medians (so $P(X_1 > M_1 \text{and } X_2 > M_2)$). This is known as the quadrant probability and will lie somewhere between 0 and 0.5.

(iii) If the variables are independent, the quadrant probability will be 0.25 (0.5 x 0.5).

(iv) If the elicited probability is greater than 0.25, this indicates positive association, as one variable is likely to exceed its median if the other does.

(v) If the elicited probability is less than 0.25, this indicates a negative association, as one variable is not likely to exceed its median if the other does.

8.2 The elicitation of joint probability distributions is difficult, as it is hard, cognitively, for SMEs to elicit what is being asked of them. Additionally, there are few choices of joint distributional forms available when fitting a distribution. Therefore, for now, it is recommended that the analyst focuses solely on the elicitation of univariate distributions.

9 Conclusions and recommendations

9.1 This report has outlined a number of ways to elicit probability distributions from SMEs, in order to represent the uncertainty surrounding their subjective estimates. The examples and case studies have demonstrated that it is not time consuming to conduct such elicitations (given one is pragmatic about the number of uncertain parameters for which probability distributions are elicited within any one workshop). Additionally, there is no requirement to buy expensive software to conduct elicitations, as it can all be done using the free statistical package R (which is downloadable from the internet) or using the aforementioned internet tool.

9.2 As discussed in depth within the main body of the report, the main benefit of elicitation is that customers will be able to make an assessment as to the levels of risk they are taking by using expert judgement to inform the decision-making process. It is therefore important that the technique becomes standard practice.

10 Annex A

10.1.1 The following script was used at the trial elicitation session:-

Trial elicitation

Going to go through a number of stages ...

Am wanting to elicit the following answer from you ...

How far is it (in miles) from Wembley Stadium to Old Trafford (by road, taking the shortest route)?

(1) Have a think about roughly how far you think it is but, to begin, **do not discuss your thoughts with anyone**. Don't worry if you haven't got a clue. This is fine. The whole point of elicitation is to show how certain, or uncertain, people are of their estimates.

(2) One by one, I want you to come to my desk and I will elicit four things from you ...
   (i) A range. So the highest and lowest values that you think it could possibly be ... so the values for which you think it is very, very unlikely that the unknown quantity would lie above or below, respectively.

   (ii) Your 'best guess' answer.

   (iii) A lower quartile (don't worry about what this is ... all will be revealed).

   (iv) An upper quartile (as above).

   So, whilst you're waiting, you can begin to independently (remember not to speak to anyone!) think about what answers you'll give for (i) and (ii).

<Distributions elicited from everyone>.

(3) Now that I have elicited distributions from you all, I'm going to show you each other's results. One by one, I want you to explain to the group why you gave the answers you did. Be honest. There's no shame in admitting that you didn't have a clue.

<Delegates explain why they gave the answers they gave>.

(4) Now that you've heard from each of your colleagues, I want you to have a group discussion and come up with one range, one 'best guess' answer, one lower quartile and one upper quartile that you're all happy with. So I want you to reach a consensus.

<Group discussion>.

(5) Thank you. I'll now show you the corresponding distribution to check that you're happy with it.

<Derivation of a distribution, any necessary iterations, followed by a 'summing up' of the final distribution>.

(6) I'll now reveal to you what the actual answer is and you can see how you did. Remember that success might not be getting close to the actual value – it could be coming up with an estimate that's way off but deriving a probability distribution that accurately reflects the uncertainty.

<Reveal answer and compare to the final distribution>.

Thank you very much for your time. If I could now just get you to fill out the short questionnaire below, that would be much appreciated. Your feedback will be incorporated into a report that both explores various elicitation techniques and considers their applicability to the Strategy and Capability domain. Your anonymity is assured.

Questionnaire

Please indicate how much you agree or disagree with the following statements ...

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
In my one-to-one session, it was clear what values I was being asked to provide.					
It was easy to reach a consensus in the group session.					
I can see why the elicitation of probability distributions for subjective judgements is important.					
Elicitation would be relevant and beneficial to my areas of work.					
I can see elicitation having many applications across the whole of Dstl.					

Where you gave a particularly high or low score (i.e. strongly agree or strongly disagree), please can you comment further ...

_____

_____

_____

_____

_____

Do you have any further comments?

_____
_____
_____
_____
_____
_____

13 References

Beach, L.R. and Scopp, T.S. (1967). 'Intuitive statistical inferences about variances'.

Beach, L.R. and Swenson, R.G. (1966). 'Intuitive estimation of means'. Psychonomic Science.

Erlick, D.E. (1964). 'Absolute judgement of discrete quantities randomly distributed over time'. Journal of Experimental Psychology.

Garthwaite, P.H., Kadane, J.B. and O'Hagan A. (2005). 'Statistical methods for eliciting probability distributions'.

Gillard, J. (2011). 'A framework for the elicitation of subjective data'. Dstl.

Gokhale, D.V. and Press, S.J. (1982). 'Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution'. Journal of the Royal Statistical Society.

Iswaran, N. (2009). 'Eliciting and Modelling Expert Assessments for Capability Modelling'. Dstl.

Kadane, J. B. and Wolfson, L.J. (1998). 'Experiences in elicitation'. The Statistician.

Kelsey, A. (2009). 'Does the size of your group matter?'. Internet article: http://www.leadingleaders.net/articles/entry/does_the_size_of_your_group_matter//

Morgan, M. G. and Henrion, M. (1990). 'Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis'. Cambridge University Press.

Nash, H. (1964). 'The judgement of linear proportions'. American Journal of Psychology.

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

Peterson, C.R. and Miller, A. (1964). 'Mode, median and mean as optimal strategies'. Journal of Experimental Psychology.

Pitz, G.F. (1966). 'The sequential judgement of proportion'. Psychonomic Science.

Pitz, G.F. (1965). 'Response variables in the estimation of relative frequency'. Perceptual and Motor Skills.

Shuford, E.H. (1961). 'Percentage estimation of proportion as a function of element type, exposure type and task. Journal of Experimental Psychology.

Simpson, W. and Voss, J.F. (1961). 'Psychophysical judgements of probabilistic stimulus sequences'. Journal of Experimental Psychology.

Spencer, J. (1963). 'A further study of estimating averages'. Ergonomics.

Spencer, J. (1961). 'Estimating averages'. Ergonomics.

Stevens, S.S. and Galanter, E.H. (1957). 'Ratio scales and category scales for a dozen perpetual continua'. Journal of Experimental Psychology.

Wesson, C.J. and Pulford, B.D. (2009). 'Verbal expressions of confidence and doubt'. Psychological reports.

This page is intentionally blank