# Elicitation of subjective (probabilistic) beliefs: a protocol for three methodological experiments

**FINAL PROTOCOL: 26/09/2018**

## Introduction

Expert judgements are often required to support decision making. Structured expert elicitation (SEE) asks individuals for probabilistic descriptions of their (uncertain) beliefs, so that these can be explicitly integrated within quantitative assessments. Some areas of science have guidance on the use of SEE[1, 2]; others, such as health care decision making (HCDM), have recently shown an interest in producing guidance.[3] This work is developed in the context of HCDM.

Research looking at important methodological choices is mostly inconclusive, for example on: which experts to engage, how to most appropriately elicit distributions, or whether group interaction increases the accuracy of group judgements. The underlying challenge of methodological research in this area is that beliefs are inherently unobservable: the accuracy of elicited judgements in representing the experts’ beliefs cannot be directly established as heterogeneity in knowledge (i.e. the fact that individuals’ beliefs differ) cannot be easily disentangled from the lack of ‘normative’ skills (i.e. individuals not being able to represent their beliefs accurately in probabilistic terms).

This document describes a proposal for experiments aimed at generating evidence that can support some of the methodological choices in elicitation. Whilst the motivation for designing these experiments is to inform the health care decision making context, conclusions drawn have applicability across other areas. Specifically, the experiments aim to evaluate alternative methods of SEE and how they perform in representing parameter uncertainty, to explore individuals’ ability to supply accurate distributions when there is heterogeneity in knowledge (for example, between sub-populations), and to explore how individuals revise their answers when presented with group summaries.

This document describes the general approach for the experiments, and then details aspects of their design.

## General approach to the experiments

The crucial issue in designing the experiments is how the alternative design choices (e.g. elicitation methods) will be compared. There are two possible approaches.

### Almanac questions

The majority of previous evaluations have relied on historical data or scientific facts – so-called ‘almanac’ questions. Almanac questions are quantities known with certainty, for example, weather or population statistics, for which individuals are reasonably expected to know something about. Individuals are asked to express their subjective probabilities[[1]](#footnote-1) for these quantities using alternative methods and performance is measured as the deviation between the known value of the quantity and the elicited beliefs.

There are limitations of this approach. Firstly, the nature of ‘almanac’ quantities differs inherently from the nature of the unknown quantities typically elicited in health care. To qualify as known quantities, almanac question relate to either uncertain events that will be realised in the future (e.g. rainfall tomorrow or survival of individual patients), facts (e.g. the distance between the earth and Mars) or to summaries of datasets (population or sample-based). The typical target questions for SEE in HCDM, however, relate to unknown parameters of a statistical model, quantities not expected to be known with certainty – for example, the probability that UK patients with a wound heal in 6 months when treated with X. This means that source information used by the expert is usually based on sampled data (e.g. published studies or own observations), and hence is itself uncertain, and may also be biased, heterogeneous and lack generalisability. Secondly, given the nature of almanac questions, beliefs about them may be formed on the basis of known facts coupled with analogical reasoning (for example, reasoning about the distance between Mars and Earth could be based on the knowledge that it would take a spacecraft one year to reach Mars). In contrast, experts in HCDM are typically health carers and one of the main sources of information determining their beliefs on the likely target questions are observations of patients and outcomes that are usually of relevance, i.e. knowledge will be mainly from observation. Whilst reasoning will also be required, it is likely that concerns focus more on uncertainty, bias, and heterogeneity.

Thirdly, whilst individuals are expected to have some level of epistemic uncertainty about their answer to almanac questions[[2]](#footnote-2), the accuracy of elicited summaries in representing this uncertainty cannot be established as this is subjective (individual). Instead, multiple almanac questions are often used and the frequency of true values that fall outside the elicited credibility regions are used as a measure of performance. Finally, responses to almanac questions can be inaccurate either if beliefs are themselves inaccurate or if, even with accurate beliefs, the individual struggles to express these in probabilistic terms.[[3]](#footnote-3) While the absolute accuracy of an elicited uncertainty measure cannot be determined using almanac questions, the *relative* accuracy of different elicitation methods could be compared by randomising individuals to different methods, since randomisation will ensure that any systematic differences between groups will be due to the elicitation methods. However, this may require a prohibitively large sample size.

### Simulated learning process

An alternative approach is based on simulated learning processes, such as the experiment described in Wang 2002[5]. Instead of asking for almanac quantities, the individual’s knowledge is determined using a simulated learning process where the trainee interacts with a virtual domain. In the Wang 2002 example[5] the virtual domain was a cat-mouse game that the individual played. A Bayesian (probabilistic) network was used to stochastically update the posterior probability of the games’ parameters conditional on the participant’s actions, to determine the subjects’ next observation. The participant’s knowledge was determined by the observations he/she was exposed to during the game, which were recorded (a database of observations).

There are several advantages to the simulated learning process in this context. Firstly, the process is reflective of the learning process (by observation) we may expect of experts in health which are typically health carers that observe patients over time. We recognise, however, that in real clinical practice information from observation is not considered in isolation, i.e. health carers also draw from published evidence, peer contact or other related evidence, and in real practice there are also consequences of making a wrong decision that might determine how cautious a clinician is in their assessment. However, there are a number of advantages in isolating the component of learning by observation. These will be explored in the following paragraphs.

If the simulated learning process constitutes the single source of information participants receive, if the participant has paid full attention to the task, the dataset should form their subjective beliefs. This is crucial as, under these conditions, elicited probabilities can therefore be directly compared to the posterior probability distribution implied by the observed dataset and prior belief– i.e. accuracy can be measured.

Secondly, the nature of this approach allows determining more conditions of the experiment than with the almanac approach. For example, it can be used to isolate the source of knowledge to only the observations provided, and in this way it can be used to impose an equal knowledge-base across subjects. This standardises the task, reduces between-participant variation and hence requires smaller sample sizes.

Finally, in this approach, the same information is provided to all participants. Any differences in the elicited distributions across individuals therefore can directly be attributed to different levels of normative ability, reflecting the skills needed to extract information from observations and quantify the resulting beliefs in probabilistic terms. This is the particular dimension of normative ability that is captured in these experiments –here called ‘probabilistic’ accuracy.

For these reasons, we here base the experimental approach on simulated learning.

## Overview of the experiments

Three experiments will use a simulated learning process to explore alternative methodological choices in SEE:

* + - The first experiment will compare two methods of elicitation (bisection vs. chips and bins)
    - The second will determine whether individuals are able to accurately extrapolate from their knowledge-base to different populations
    - The third will explore how individuals revise their own probability assessments when presented with Delphi-type group summaries.

### How will information be fed to participants in the experiments?

Participants will be shown a number of observations generated randomly from a statistical model, which are recorded. The game will be as abstract as possible so that all knowledge is acquired from the game and not influenced by external information or separately acquired prior beliefs. The model will represent a generic medical problem with the following features:

Population: Patients present with a (hypothetical) symptomatic disease, aiming for the resolution of their symptoms. A generic description will be used, avoiding any terminology relating to a specific disease.

Interventions: Two treatments, i.e. two pills. The quantity of interest refers to only one of the pills; the other one will be used to allow for an incentive mechanism. This is detailed in a separate section ahead, section 3.5.

Outcomes: Number of patients achieving relief of symptoms after treatment.

Description of the game: Participants act as practitioners over a number of clinic days (the number of days is defined in each experiment, see details ahead). In each day, participants care for a variable number of patients (6 to 13). The number of patients varies randomly per day, to make it more difficult for individuals to algebraically evaluate the proportion of successes by counting. This is desirable as it mimics more closely the knowledge from observation typical of the likely experts in HCDM. Participants are presented with the two pills and are asked which they will use to treat patients that day. Participants will be presented with the following representation:

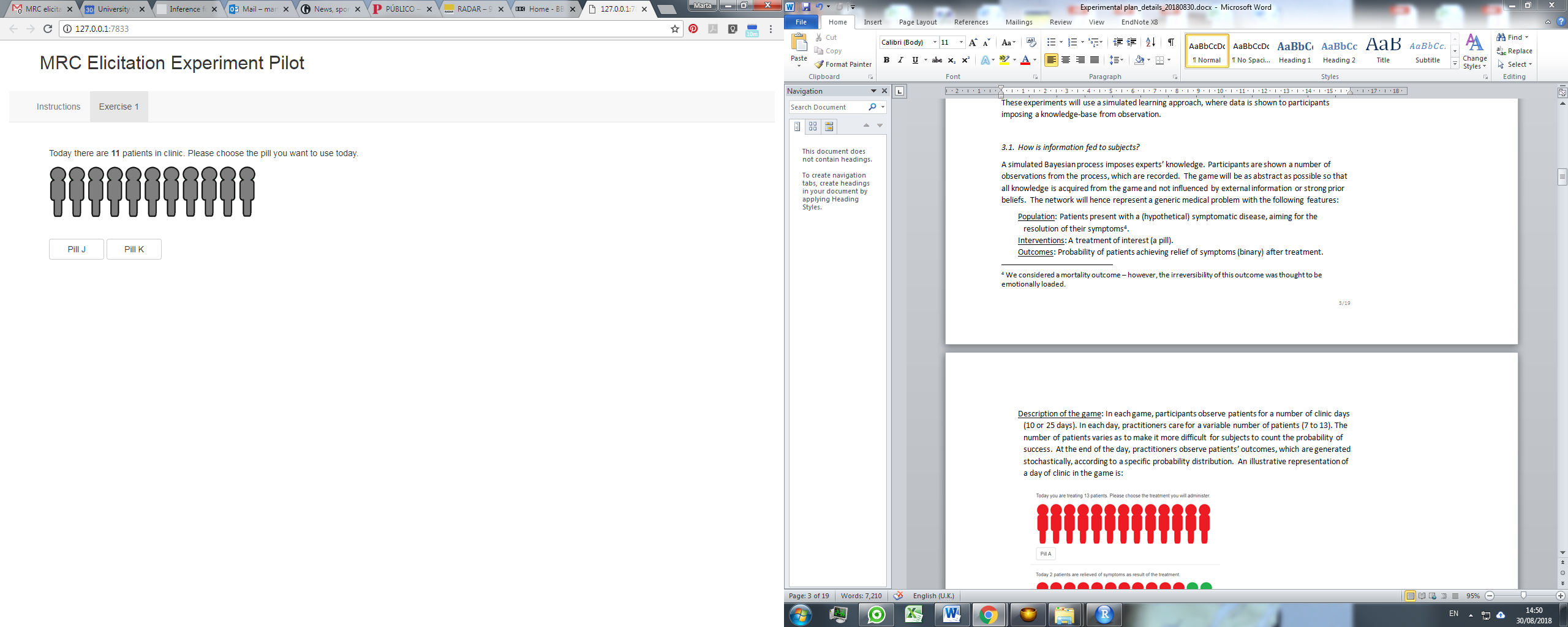


Figure 1: Snapshot of the R shiny app (1).

Once the participant chooses the pill to use, a new screen appears (see Figure 2 below) showing participants how many of the patients achieved symptom resolution that day with the chosen pill (number of successes). This number is generated randomly, according to a specific model (presented in more detail for each experiment).

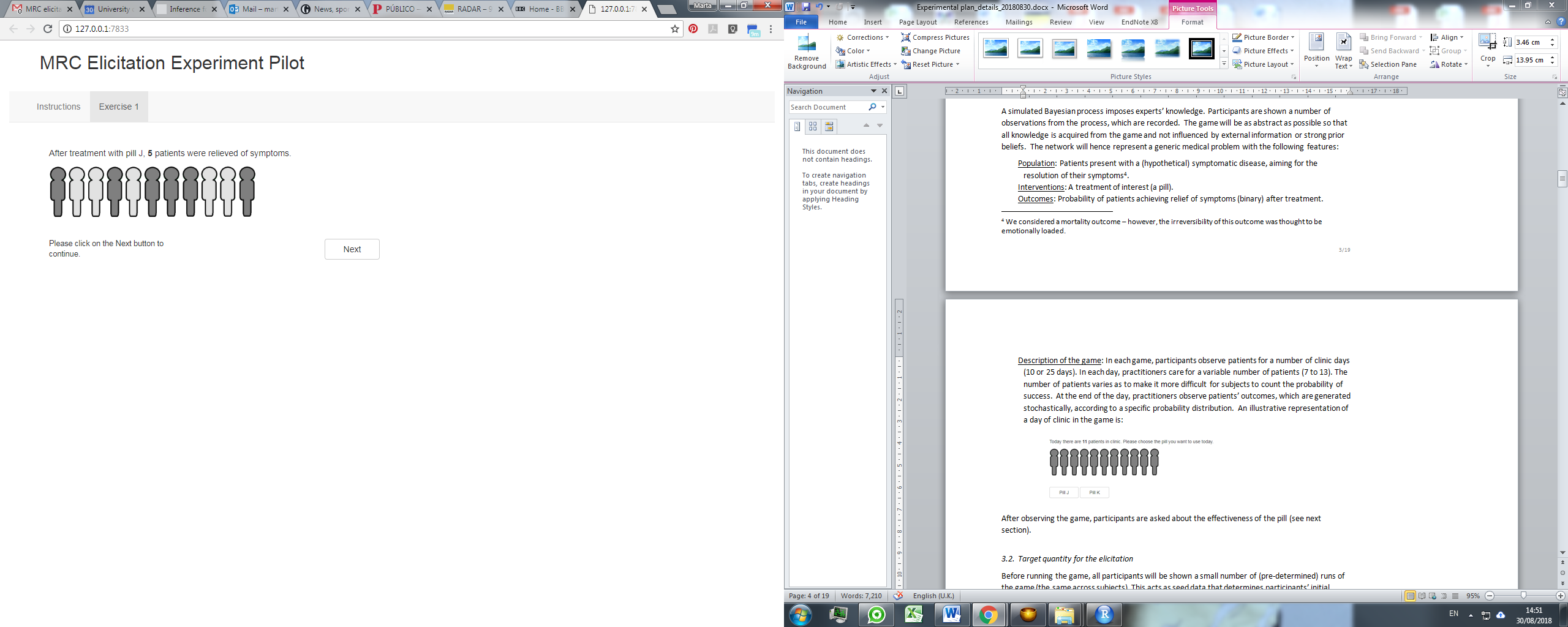


Figure 2: Snapshot of the R shiny app (2).

After observing the game, participants are asked about the effectiveness of the most effective pill (see next section).

### Target quantity for the elicitation

After running the game, participants will be asked to elicit the expected population values for the probability of success (probability of achieving symptom-relief) with the pill of interest.[[4]](#footnote-4) This refers to the wider population from which the patients observed in the different clinic days are drawn from. Specifically, participants will be asked:

*If you were able to treat all patients in the population, what proportion would you expect to become symptom-free?*

Before running the game, all participants will be shown a small number of pre-determined runs of the game (equal across subjects).[[5]](#footnote-5) This determines participants’ initial beliefs about the probability of treatment being successful (prior information). It aims to homogenise participants’ priors and mitigates sensitivity of results to different uninformative or vague priors that subjects may use. Participants’ beliefs on the quantity of interest should reflect the posterior probability of success, i.e. the (approximate) prior beliefs fed to participants updated with the data observed from the different clinic days. This information will be presented to participants as follows:

*The two pills were previously tested in a clinical study that recruited 6 patients:*

*. Of the three that were randomised to take pill A, two got symptom relief.*

*. Of the three that were randomised to take pill B, one got symptom relief.*

*(Note that the study was very small and hence at this point you should still be uncertain about the effectiveness of the two pills.)*

### Participants and sample size

Given that the subjects’ knowledge-base is ‘built up’ in the approach chosen for the experiments, it is not relevant for individuals recruited to be experts in any particular knowledge-area, and it is more important that they are representative of the type of normative skills we expect from experts in HCDM. Hence, we will recruit medical students from the Hull York Medical School (HYMS) and nursing and midwifery students from the department of Health Sciences (HS) at the University of York.

We will aim to recruit a sample size of 64 participants, which we believe is feasible given the funding available for the experiments. This number will allow experiment 1 to have a double balanced design will all sequences applied being different (details of the design of experiment 1 are given in section 4.6). No formal sample size calculation was undertaken given the lack of evidence on the potential magnitude of effect size and its variance.

Between-subject variation will be explored using participants’ characteristics, such as age , gender, whether they are nursing, midwifery or medical students, undergraduate or postgraduate, the year of studies, whether they have a quantitative A level or AS level, and how confident the participant is in using probabilities (very confident to absolutely unconfident). We will also collect responses to the Berlin Numeracy Test[6, 7] and the Scott and Bruce's General Decision-making Style questionnaire[8].

The Berlin numeracy scale is test of statistical numeracy and risk literacy, a component of numeracy that is thought to be important for informed and accurate risky decision making such as required in engineering or in health. Statistical numeracy specifically refers to an understanding of the operations of probabilistic and statistical computation, such as comparing and transforming probabilities and proportions. The 7-item questionnaire was used –reproduced in Appendix 1. Each correct answer is attributed a point and the total score is the sum of the points. The numeracy scale is to be applied at the end of the experiments to avoid promotion of under- or over-confidence.

The General Decision-making Style questionnaire[8] includes five subscales examining five decision-making styles: rational (thorough information search and logical evaluation of alternatives), intuitive (reliance on gut feelings and hunches), dependent (advice seeking and reliance on others), avoidant (a tendency to escape and avoid decision situations) and spontaneous ((a tendency to make fast and speedy decisions). The questionnaire includes 25 questions, five in each subscale. All questions are measured on a scale between strongly disagree (1) to strongly agree (5). Higher scores in the subscale (the sum of the items) mean that this style is used more frequently. See Appendix 1 where the questionnaire is reproduced

### Outcomes

*Primary outcome*: Across the three experiments, the primary outcome will be a measure of the difference between the elicited distribution and the posterior distribution from the prior and data provided to participants. To evaluate the difference between the two distributions the Kullback-Leibler divergence (throughout referred to as KL), a measure of the information lost when one distribution, q(x), is approximated by another, p(x), will be used. It is defined as:

|  |  |
| --- | --- |
| KL(q(x)||p(x)) = ∫p(x)log(p(x)/q(x))dx | (Equation 1) |

This metric is grounded in information theory. It is invariant to transformations and obeys the chain rule of conditional probability. It is non-symmetric, in that KL(q(x)||p(x)) is different from DKL(p(x)||q(x)).

*Secondary outcomes*: Two secondary outcomes will be considered: bias, defined as the difference in the means of the true and elicited (and fitted) distributions; and uncertainty, defined as the variance ratio of the two distributions. Additional qualitative information may also be requested.

### Monetary incentives

A number of areas of experimental behavioural research involving human subjects (such as in experimental economics) advocate the implementation of monetary incentives so that participants perform tasks to the best of their ability, as they would in ‘real’ circumstances.[9] This is important here as the students we aim to recruit may not be motivated to complete the task as assiduously as would be expected of professionals.

This incentive structure is expected to be key in incentivising engagement to the game, and in learning the proportion of success. Specifically, the incentives we will use are:

* a show up fee of £15,
* additionally, an individual reward, to be attributed on the day, that will approximately vary between £0 and £25.

The individual reward will be based on a reward scheme that will aim to incentivise the two basic tasks individuals are required to perform: 1) learn from playing the game and 2) be as accurate as possible in the elicitation.

Accuracy will be incentivised by based on the KL scores for each of the elicitations conducted by the participant. To incentivise learning from playing the game, a choice component is added where the participant is requested to choose which of the two pills to use each day. Such a choice component also mimics clinicians’ choices in the care provided to patients, and increases engagement as it introduces an active task for participants. Specifically, of the two pills offered as an option in each day of clinic, one is the pill of interest for the elicitation and the other is less effective and is not of interest (i.e. the probability of success will not be elicited for this pill)[[6]](#footnote-6). At the start of the game, participants will be told that one of the pills is less effective than the other and that, from playing the game, they will be likely to identify the least effective pill at which point they should stop using it. The sooner participants stop using the least effective pill, the greater the financial reward they receive.

### Other aspects of conduct

A pilot will be undertaken with colleagues to evaluate the feasibility of the experiments, time to completion, and to determine the optimal experimental conditions (e.g. value of the probability parameters).

The experiment will be designed in R using the SHINY package. A bespoke training package will be developed and delivered to participants (see section 7 for more details).

The experiment is planned to use a face-to-face session lasting around 1.5 to 2 hours. Subjects will, however, be asked to complete all the tasks individually (i.e. the game and the elicitations). We will record what states participants were presented with during the experiment and all actions taken by the participants.

Ethics approval has been sought and granted prior to the conduct of the experiment. Participants will also be asked to fill in a consent form, which aims to comply with the General Data Protection Regulation (GDPR).

### Aims (and set-up) for each of the experiments

The aims of the experiments are:

* Experiment 1: Comparison of two alternative methods of elicitation.

This experiment will use a set-up (set-up 1) that imposes knowledge directly on the target quantity. This is as described in section 3.1, where participants observe directly data on the target question in 3.2.

* Experiment 2: To determine if individuals’ are able to ‘extrapolate’ from their knowledge-base to different populations.

This experiment will use a different set-up (set-up 2), where the imposed knowledge is relevant to, but is not directly on, the target parameter. Specifically, participants will observe outcomes for subgroups in a population and will be asked to quantify the probability of treatment being successful for another population comprising of a mixture of the two subgroups.

* Experiments 3: To understand how individuals review their own probabilistic assessments when presented with Delphi-type group summaries.
  + 3.1: Does individuals’ performance affect the extent of revision?

This experiment uses experimental set-up 1, but further presents individuals with a (hypothetical) group summary. It aims to evaluate whether the extent of revision by individuals, after being presented with the group summary, is determined by their own performance.

* + 3.2: How does between-expert variation within a group affects individuals’ revision?

This experiment uses experimental set-up 1, and presents individuals with, not just the (hypothetical) group summary, but also summaries of the elicited quantities by each member of the group. It aims to assess the extent of revision when there are different levels of between-expert variation in the group.

These will be described in more detail in sections 4, 5 and 6.

## Experiment 1: comparing different methods of elicitation

The aim of the experiment is to assess how well the elicited probability distributions derived from two different methods of elicitation match description(s) of uncertainty. This experiment uses set-up 1, where the knowledge offered to participants is directly on the target quantity, that is, where the data observed by participants constitutes direct samples of the population and outcomes that the target question concerns.

Participants will be randomly allocated to one of the two methods. Differences between the groups will reflect the method’s ability to allow experts to represent their beliefs probabilistically: this is here referred to as ‘probabilistic’ accuracy. Experiment 1 will hence compare the probabilistic accuracy of two alternative methods of elicitation.

### Methods of elicitation

The two methods of elicitation, the bisection and the chips and bins (or histogram), are both widely used for SEE in HCDM[10]. The bisection method is a variable interval method (VIM) and asks experts to assess the three quartiles of the distribution. The chips and bins is a fixed interval method (FIM) that defines a larger number of intervals (typically up to 20 bins) and asks the expert to distribute a fixed number of chips across these intervals. The more chips placed in a particular interval, the stronger the belief that the true value of the quantity of interest lies in that interval.

Garthwaite[11] identifies that existing methodological research is conflicting as to which type of method, FIM or VIM, is best and cites one study in favour of FIM[12] and another with opposite conclusions[13]. In HCDM, the general understanding is that the bisection method returns wider representations of uncertainty (argued to be more appropriate in representing within-expert uncertainty) than the chips and bins, although the latter has been proposed to be more intuitive for less quantitative experts to grasp[10, 14]. Further detail on how each method has been implemented is reported in section 4.5.

Participants will be asked to express uncertainty in the quantities elicited. In the context of learning from observation, to express uncertainty, individuals will need to have some consideration for the number of samples they have observed: the more samples they observe the lowest level of uncertainty they should have about the expected population value. Additionally, participants should also consider the level of variability of continuous quantities (i.e. the level of variation observed across patients), or the level of over-dispersion in probability parameters (greater variation than expected in the probability of success between subsamples of patients). For example, for the task at hand in these experiments, there may be over-dispersion if there is heterogeneity in the population sampled, that is, if the population differs day-by-day. For the same sample size, higher levels of variability or over-dispersion leads to higher levels of epistemic uncertainty.

It is hence important to also acknowledge the possibility that the two methods do not perform equally under different levels of uncertainty. For example, bisection may overestimate uncertainty under low uncertainty and the chips and bins method lead to more accurate elicitations, whilst under high uncertainty the opposite may happen and the chips and bins may underestimate uncertainty while the bisection is more accurate. Hence this experiment 1 will consider two different levels of uncertainty, determined by a different number of observations and a different level of over-dispersion. Further details on how these will be specified are considered in the next section, 4.3.

### Hypothesis to be tested

In this experiment we aim to test two main hypotheses:

H1: The bisection method leads to more accurate elicitation (in a scenario of lower uncertainty).

H2: In a scenario of increased uncertainty, the bisection leads to less accurate elicitation.

### Experimental set-up

To test the abovementioned hypotheses, levels of the following factors will be specified:

Method of elicitation, X: chips and bins (X= 0) and bisection (X= 1)

Level of uncertainty, U: Two levels of uncertainty are defined.

* + Low uncertainty scenario, U=0

Participants observe data for N days treatment with the most effective pill, specifically N=25 days (clinical days where participants choose the least effective are additional to N). On each single day, i, participants observe n[i] patients. The possible values assumed for n[i] are integers from 6 to 13 {6,7,8,9,10,11,12,13} drawn randomly with equal probability. The Binomial distribution governs the outcomes in each day of clinical practice. Its probability parameter for patients treated with the pill of interest, p0, is constant within a game, and across games may assume one of the following values {0.3,0.4,0.6,0.7}. This represents a situation where there is no over dispersion. Specifically, the number of patients achieving symptom resolution, O[i], is given by O[i] ~ Bin(p0, n[i]).

The second, less effective, pill (implemented to accommodate an incentive mechanism for engagement with the game, section 3.5), is assumed to have a risk difference of ­‑0.25.

* + High uncertainty scenario, U=1

In this scenario, the set-up is similar but participants also observe fewer days of treatment with the pill of interest than in the previous scenario, specifically M= 10 days (clinical days where participants choose the least effective are additional to N). The model governing outcomes is also Beta-Binomial (over-dispersion model); this means that the number of patients achieving symptom resolution, O[i], is given by O[i] ~ Bin(p[i], n[i]), with the probability parameter for day i being p[i] ~ Beta(p0\*m, (1 –p0)\*m), where p0 is the mean probability and m a measure of (lack of) heterogeneity (bigger m leads to lower over dispersion). As in the previous scenario, p0 may assume one of the following values {0.3,0.4,0.6,0.7}. The constant m assumes a value of 2. The data can be alternatively expressed as a number of independent draws from a Beta-Binomial(n[i], p0, m).

The risk difference for the second pill is equal to that assumed in the low uncertainty scenario with the corresponding outcome will be drawn from a Binomial model (a Binomial model instead of Beta-Binomial, is here used to allow participants to distinguish this pill from the pill of interest).

### Target quantity

Participants will be asked to elicit their expected value for the probability parameter in the overall population (section 3.2) – that is, uncertainty over p0. For this, participants will be asked to consider the wider population – this is equivalent to eliciting for a large number of samples.

The wording of the question is

*About pill C, the most effective of the two pills... If you were able to treat all patients in the population with this pill, what proportion would you expect to become symptom-free?*

The target question requires that participants update their prior beliefs (determined by the seed data) with the data observed from the different clinic days. The prior is Beta (defined as the posterior from combining the seed data with a Beta prior), and hence for the low uncertainty scenario the posterior distribution can be derived using conjugacy, and is also a Beta distribution. For the high uncertainty scenario, however, a conjugacy relationship is not available.  Markov-Chain Monte-Carlo methods (MCMC) will be used to obtain this distribution.

### Method of elicitation

Experts will be presented with the target question and will be asked to elicit using one of the two alternative methods. Both will be preceded by asking participants for bounds, which are presented as:

*“I believe that it's very unlikely that:*

* + *the proportion is less than \_\_\_\_\_\_\_\_\_\_ percent,*
  + *the proportion is greater than \_\_\_\_\_\_\_\_ percent.”*

The **bisection method** will elicit quartiles (the median, M, the 1st quartile, Q1 and the third quartile, Q2), and will be worded as follows:

*“Can you determine a value (your midpoint) such that the proportion is equally likely to be less than or greater than this value? \_\_\_\_\_\_\_\_\_\_\_\_\_.”* (M)

*Suppose you were told that the proportion is below your assessed midpoint. Can you now provide a new value (your lower quartile) so that the proportion of patients is equally likely to be less than or greater than this value? \_\_\_\_\_\_\_\_\_\_\_\_\_.”* (Q1)

*Suppose you were told that the proportion is above your assessed midpoint. Can you now provide a new value (your upper quartile) so that the proportion of patients is equally likely to be less than or greater than this value? \_\_\_\_\_\_\_\_\_\_\_\_\_ .”* (Q3)

After eliciting, the participant will be presented with the following summary:

*“Your answers imply that the proportion is equally likely to be less than and greater than* <<M>>, *and it is equally likely to be between* <<Q1>> *and* <<Q3>> *as it is to be outside this range”.*

The participant is then invited to revise his/her answer.

For the **chips and bins method**, the bounds initially elicited will be used to constrain the range presented for the histogram. The number of bins will be variable to produce reasonable bin widths (1%, 2%, 5% or 10%) that fit the bounds given. The number of chips will be the twice the number of bins. The graphical display implemented is illustrated next.

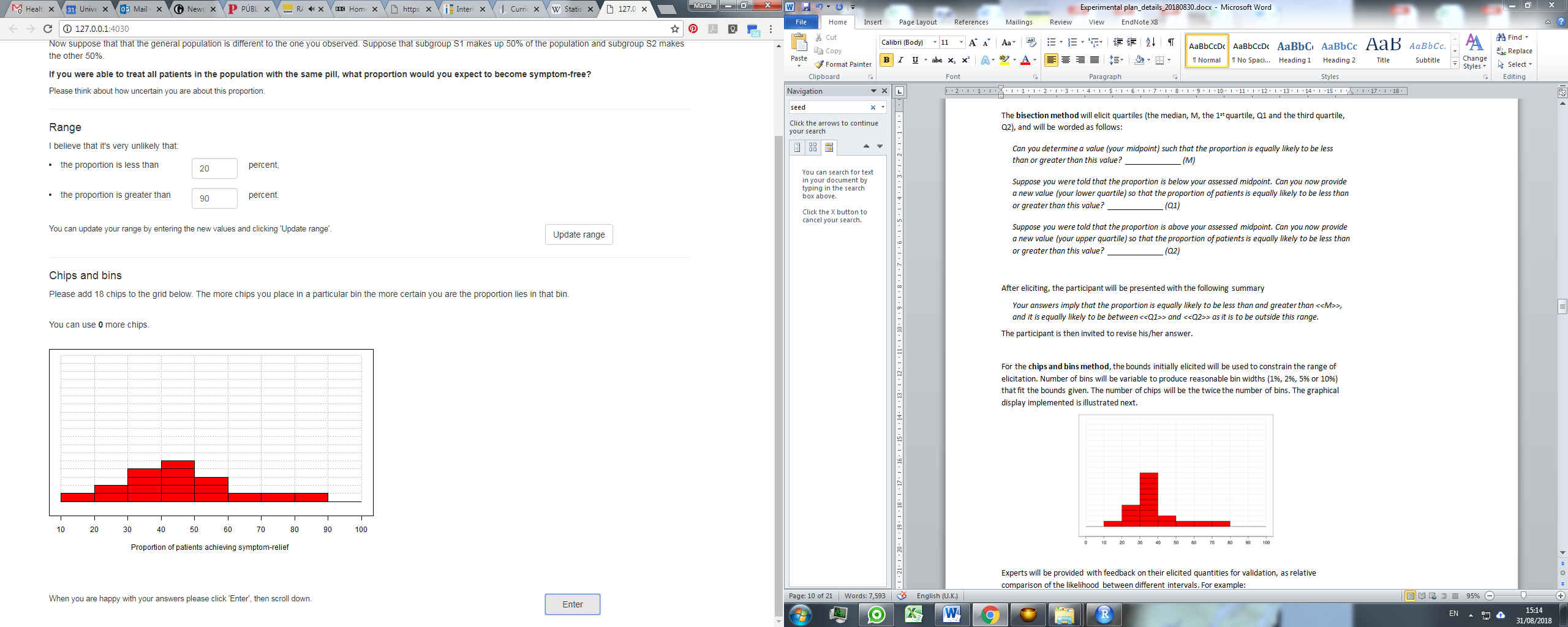


Figure 3: Snapshot of the R shiny app (3).

Experts will also be provided with summary feedback on their elicited quantities. The feedback will highlight the probability assigned to the modal bin and to values above/below the modal bin. For the histogram above this would be:

*“Your answers imply that:*

* *There is a 28% probability that the proportion is between 40 and 50.*
* *There is a 39% probability that the proportion is between 10 and 40.*
* *There is a 33% probability that the proportion is between 50 and 90.”*

After being presented with the summary, participants are given the opportunity to revise their initial assessments.

Fitting will not be undertaken within the elicitation session.

### Design

The experiment will use a full factorial design with all 4 combinations of the two levels, U and X: 00 (U=0,X=0), 01 (U=0,X=1), 10 (U=1,X=0), 11(U=1,X=1). To increase the number of observations, all participants will provide four different quantities (over 4 periods) representing each of the different experimental conditions. Given each individual will provide 4 quantities, the parameters of the models for each experiment will use different probability values (p0 values) of: 0.3, 0.4, 0.6 and 0.7. The experiment will need to block the probability value used at each period. Hence, a 4x4 Graeco-Latin design will be implemented. The design is shown below for 4 subjects, identifying the scenario of the two levels U and X, and probability values appear in brackets.

Table 1: Example of a 4 by 4 Greco-Latin square identifying UX (p), where U the uncertainty scenario {0,1), X is the method {0,1}, and the probability value p0 is in parenthesis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Subject | | | |
|  |  | 1 | 2 | 3 | 4 |
| period | 1 | 00 (0.2) | 10 (0.6) | 01 (0.8) | 11 (0.4) |
| 2 | 10 (0.4) | 00 (0.8) | 11 (0.6) | 01 (0.2) |
| 3 | 01 (0.6) | 11 (0.2) | 00 (0.4) | 10 (0.8) |
| 4 | 11 (0.8) | 01 (0.4) | 10 (0.2) | 00 (0.6) |

The design proposed above generates 4 probability distributions per subject recruited –i.e. a sample of 64 subjects would generate 256 probability distributions. The list with the randomisation sequences is provided in Appendix 2. Note that we generated 69 randomised sequences in case more participants were recruited

### Primary outcome measure: evaluation and modelling approach.

Evaluating KL (including fitting): As stated in section 3.4, the primary outcome metric is the KL divergence of an elicited distribution in relation to a true distribution. The true distribution for the low uncertainty scenario can be obtained using conjugacy, but for the high uncertainty it will be obtained by MCMC and approximated by a kernel density estimate. The elicited distribution will be derived from the elicited summaries by fitting a Beta distribution to the bounds and the elicited summaries using each method. Sensitivity analyses will test the use of alternative distributions (such as the generalised Beta or spline-based distributions). The KL will be computed using numerical methods of integration. Appendix 3 presents the R code used to compute KL divergence.

Model specification and model selection: The dependent variable for modelling is yij, the KL of subject j in period i (the period reflects the order of the quantities elicited by every particular subject). The KL is non-negative, which is important in determining a model to use. To our knowledge, KL has not been modelled before in the context of elicitation, and hence the likely form of its distribution and the distribution of residuals in the regression are unknown to us. For this reason we will use generalised linear models, where we will start with using a Normal distribution with identity link, and then explore alternative link functions (g) and/or distributions if necessary so that the residuals will be roughly Normal. Model specifications that may be of particular relevance, given the support of the KL metric, are the Normal distribution with log link or the Gamma distribution with identity link (note that the Gamma distribution does not include zero in its support – however, we do not expect any of the observed KLs to be zero).

In choosing between the above-mentioned specifications, we will compare models and assess fit using: i) penalized-likelihood criteria such as the AIC and BIC and ii) using residuals for random error and random effects, looking for outliers and distributional features.

The full model will be:

|  |  |
| --- | --- |
| E(yij) = g-1(mu + pi + sj + qprobij + b1\*Xij + b2\*Uij + b3\*Uij\*Xij + eij). | (Equation 2) |

Here, pi is the fixed effect of the ith period (where p1 = 0), sj the effect of the jth subject described using a fixed effect, qprobij the fixed effect of the probability value used by subject j in period i where probij assumes values {0,1,2,3} for values of probability of {0.3, 0.4, 0.6, 0.7}, Xij identifies the method of elicitation and is 0 if chips and bins is used by subject j in period i, or 1 if bisection is used, and Uij identifies the uncertainty scenario taking a value of 0 if subject j in period i completes an experiment for the low uncertainty scenario, and 1 otherwise.

First, we will evaluate whether subject effects can be assumed as random effects with standard deviation (sigma). Then we will undertake covariate selection, starting with the full model above, and using backward selection of the subject and period effects, then interaction terms, and only then of the main effects.

The fully balanced and strong repeated measures design means that period and carry-over effects can still be identified. We do not expect these to be significant; the training package will also minimise these effects by asking participants to run two practice elicitations using both methods. Period effects are included in the full model above. First order carry-over effects will be tested in a sensitivity analysis.

The two main hypotheses to be tested with this experiment are:

Ha1: b1 ≠ 0 (test of the coefficient for the elicitation method).

Ha2: b3 ≠ 0 (test for the coefficient of the interaction term)

### Secondary outcome measures: evaluation and modelling approach.

Bias and the variance ratio will be evaluated from the data. GLM’s will be used to determine whether a Gaussian distribution can be used or whether other distributions and link functions fit the data better. For the variance ratio a log link function may be used. We will also explore using bivariate models.

At the end of the experiment, participants will also be asked if they found each of the methods (generally) easy to complete (Response options: easy, challenging, very difficult) and if they had any preferences regarding the elicitation method used (“*If, in a future elicitation, you were given a choice between chips and bins or bisection, which would you choose*?” Response options: chips and bins, bisection, indifferent. *Please justify your choice*. Response in free text). This will be descriptively analysed. An open text box for further comments will also be provided. Comments left by participants will be qualitatively described.

## Experiment 2: Are individuals’ able to ‘extrapolate’ from their knowledge-base?

In SEE, variation in elicited judgements across experts may arise from experts having a different knowledge-base from which they form their beliefs. To provide a probability distribution for a common target quantity, individual experts will need to adjust (‘extrapolate’) their beliefs using some form of analytical reasoning. A simple example is where a health care professional observes a sample comprising two subgroups, but the subgroup distribution observed is different to that of the overall population over which the target question focusses on. The expert has a knowledge base that is relevant (in that he/she observes both subgroups) but when asked to elicit at the population level they need to adjust (or re-weight) what they directly observed. This is the situation that this experiment will focus on. This experiment examines how well individuals make such adjustments. This can inform the design of SEE exercises in practice and recruitment of experts. For example, if experts are not able to make such adjustments reliably, then one could recommend always eliciting conditional quantities. This experiment could also inform whether SEE should prioritise experts with good knowledge of an area but poorer performance, or experts with imperfect knowledge of an area but who have better normative skills and are good extrapolators.

Further details on the experiment are provided below. Note that common aspects to experiment 1 (e.g. evaluation of outcome measure) will not be mentioned here.

### Hypothesis tested

The specific hypotheses tested in the experiment are:

H1: Accuracy when extrapolation is required is associated with probabilistic accuracy.

H2: Accuracy when extrapolation is required is associated with the extent of the extrapolation (difference in the split between observed and target populations).

The second hypothesis implies that the experiment needs to set up different levels of the extent of extrapolation required, that is, different levels of ‘representativeness’ of the observed knowledge base in relation to the target population.

### Experimental set-up

This experiment will use a different set-up than that of experiment 1. At each clinic day, participants will be shown a total number of patients (n\_total) that, analogously to experiment 1, are sampled randomly from the set {6,7,8,9,10,11,12,13} with equal probability of 1/8. However, patients will be grouped into two groups (S1 and S2). Participants will be told that one group has better chance of symptom resolution than the other, but at the beginning of the experiment participants will not know whether this is S1 or S2. The number from each subgroup in each clinic day is generated randomly, with a pre-determined probability parameter, p\_split.

n\_s1 ~ Bin(p\_split, n\_total); n\_s2 = N – n\_s1

Different levels of p\_split will be examined to allow testing for hypothesis 2 corresponding to the following odds 80:20, 70:30, 60:40. The first subgroup (observed in the largest proportion) will be the least severe patient group. An illustrative representation of how participants will see this information is:

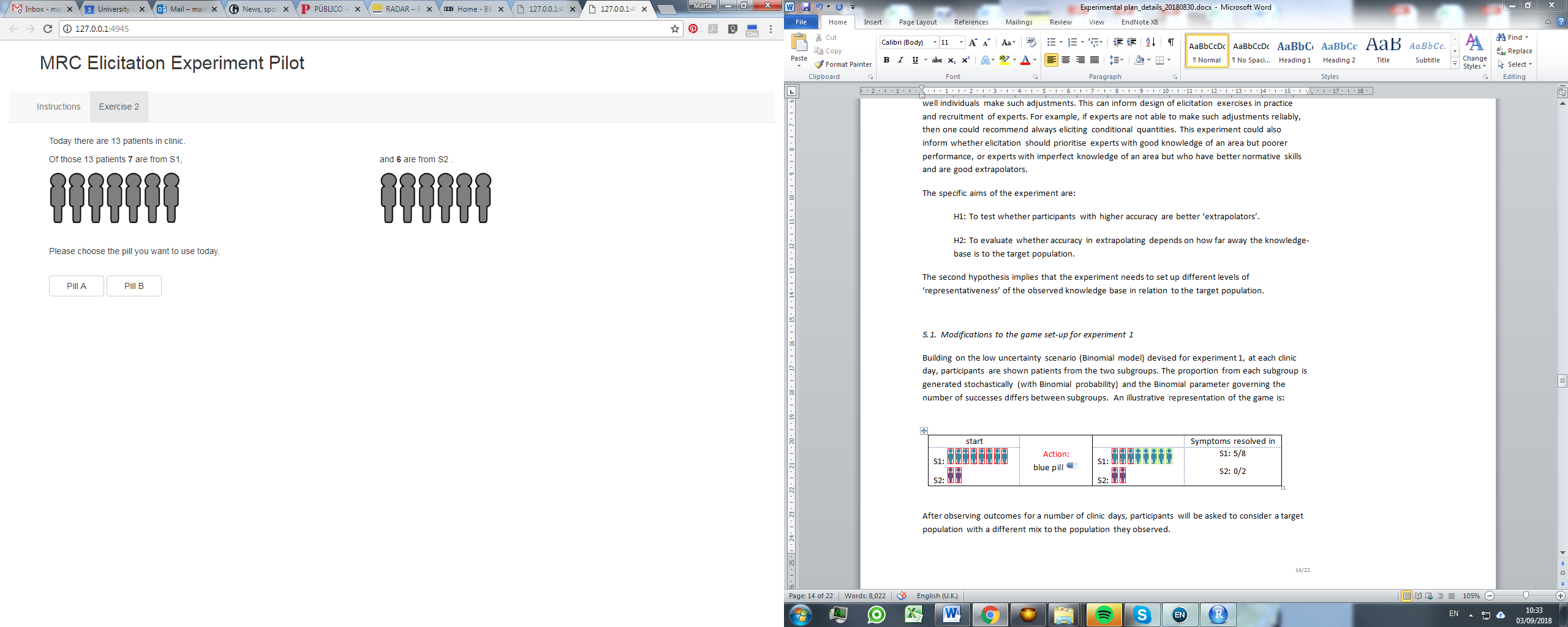


Figure 4: Snapshot of the R shiny app (4).

There will still be two new pills available in clinical practice, and these will be used by the participant in exactly the same way as in the first experiment. One of the pills is less effective, and it will be less effective in both groups of patients (the same risk difference will be applied to both subgroups). Once the participant chooses the pill they wish to use, they will observe outcomes for the two subgroups. The number of successes is governed by a Binomial model (as in the low uncertainty scenario in experiment 1). The probability parameter governing the number of successes for the pill of interest differs between subgroups:

r\_s1 ~ Bin(p1, n\_s1); r\_s2 ~ Bin(p2, n\_s2).

In this experiment, the largest subgroup will have a probability parameter of 0.35 and the smallest subgroup of 0.65. The second pill will have a risk difference of 0.3.

A representation of how participants will see this information is:

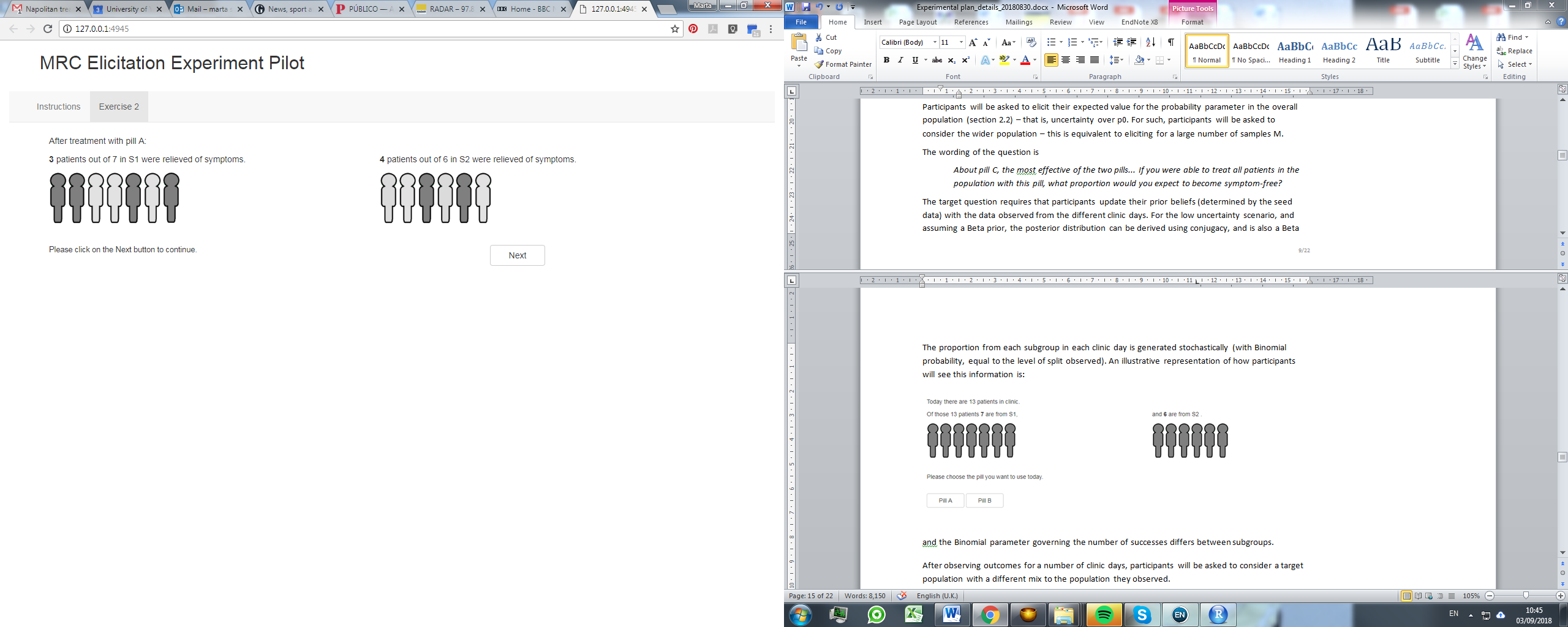


Figure 5: Snapshot of the R shiny app (5).

The participants will observe outcomes for the pill of interest for 15 clinic days.

### Design

The experiment uses a randomised block design. The factor of primary interest is the different observed subgroup splits: 80(80:20), 70(70:30), 60(60:40). Participants observe outcomes, by subgroup, for a pre-defined number of clinic days. Participants are asked to elicit first for a target population with same split (T0) and then for a 50(50:50) split (T1) – see diagram below.

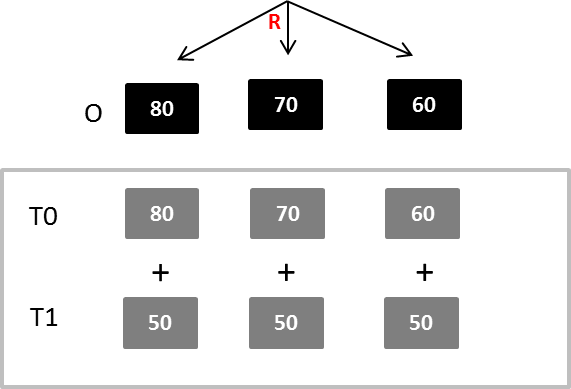


Figure 6: Summary of the processes for experiment 2.

The nuisance factor is the method of elicitation, X (chips and bins or bisection).

### Target quantity

As detailed above, participants will firstly be asked about the effect of the treatment for a population with a split of patients that is equal to what they have observed.

*“About pill A, the most effective of the two pills...*

*Suppose that the patients you have just observed were representative of the general population, that is, the split of S1 and S2 patients is unchanged.*

*If you were able to treat all patients in the population with this pill, what proportion would you expect to become symptom-free?”*

Then, participants will be asked to re-elicit for a 50:50 split of patients, which differs from the split they have observed.

*“Now suppose that that the general population is different to the sample you observed. Suppose that subgroup S1 makes up 50% of the population and subgroup S2 makes the other 50%.*

*If you were able to treat all patients in the population with the same pill, what proportion would you expect to become symptom-free? “*

### Inference

The primary outcome is the KL measure at T1, y1 (i.e. for the target population with a 50:50 split). KL will also be evaluated for T0, where the target split is equal to the observed split, to generate a measure of probabilistic accuracy, y0.

The full model will be:

|  |  |
| --- | --- |
| E(y1j) = g-1(mu + b1\*y0j + b2\*I(Oj==70) + b3\*I(Oj==80) + b4\*Xj + ej | (Equation 3) |

where j is the subject, y0j is individual’s j KL for T0, y1j his/her KL for T1, Oj is the split individual j was randomised to, I() the indicator function that assumes a value of 1 of the enclosed statement is true and 0 if false, and Xj the method of elicitation used by subject j (0 if chips and bins is used or 1 if bisection is used). As with the above experiment, a GLM framework will be used allowing different link functions (g) and distributions to be tested.

The hypotheses being tested are:

Ha1: b1≠0

Ha2: b2 ≠ 0 or b3 ≠ 0

## Experiments 3: To understand how individuals review their own probabilistic assessments when presented with Delphi-type summaries.

Experiments 3 aim to gain an understanding of how individuals revise elicited distributions when presented with Delphi-type summaries. We use the recent modified EFSA Delphi method that allows quantifying uncertainty in the form of probability distributions. As with the original Delphi, this method makes use of multiple (sequential) questionnaires (called ‘rounds‘), and at every round experts are fed back an anonymised summary of the information collected in the previous round. This form of interaction between experts is controlled, and advocates of the Delphi argue that it allows for the benefits of the sharing of information without the risks of personal factors influencing judgements inappropriately. In contrast to the original Delphi, the modified EFSA version does not aim to achieve consensus; instead, after all rounds are completed, a final distribution is obtained by using mathematical aggregation with equal weighting.

Exploring Delphi type methods is important for SEE in health care decision making because experienced facilitation is not required and the exercise can be completed remotely.

Despite the benefits of reduced interaction between elements of the group, how individuals revise their estimates in a Delphi process is not well understood. The literature hence recommends research to focus on how and why an individual changes her judgement.[15] This is the focus of the two sub-experiments proposed here, which specifically evaluate:

Experiment 3.1: does the extent of revision depend on the individual’s probabilistic accuracy?

Experiment 3.2: is the extent of revision associated with discrepancy with the group’s assessment?

Each of these experiments is detailed next.

### Experiment 3.1: Is the extent of revision associated to discrepancy with the group, and does the individual’s probabilistic accuracy determine the extent of revision?

#### Hypothesis to be tested

This experiment aims to evaluate if low performers (in terms of probabilistic accuracy) revise their answers to a greater extent (to approximate the group’s distribution) than high performers, and explore how different features of the group distribution determine the extent of revision.

Specifically, the hypothesis being tested are:

Hypothesis 1: Do participants revise their judgements to different extents according to the type of group summary they are presented with?

Hypothesis 2: Does the extent of revision depend on the level of performance of individuals?

The results can have important implications for recommendations on the use of the Delphi process. If accurate individuals are less prone to revising their initial distribution to approximate the groups’ summary, then the group distribution may converge to being more accurate than the individually pooled initial estimates. This is a desirable feature of the Delphi process as it means the revision process dilutes the effect of low performers. It is worth highlighting that the Delphi process also gets participants to explain the reasoning behind their distributions, and this information is shared (anonymously). In our experimental design, however, this has not been implemented.

#### Experimental set-up

This experiment uses the set-up of experiment 1. Participants are asked to play a ‘low uncertainty game’ (Binomial model) with a p0 value of 0.5 and asked to elicit uncertainty for p0, the target quantity. This will be used as an initial distribution for individual *j*. Its associated KL measurement (y0,j), as with experiment 1, is a measure of probabilistic accuracy.

Participants then receive one of 3 possible group summaries on the quantity of interest: concordant with their initial probability distribution, or discordant and either more or less precise than their own. The group summaries provided are hypothetical (groups will not be formed). Discordance is defined based on the median, and precision on the variance. Specifically, the three scenarios for the group summaries are:

* Sc1: the group summary is consistent with the distribution of the individual expert (left hand panel in Figure 7) – the group hypothetical distribution presented to participants has a standard deviation equal to the individual’s distribution (the distribution fitted to the elicited summaries), and its median is the 55% percentile of the individual’s distribution if the elicited is biased downwards or the 45% percentile if biased upwards.[[7]](#footnote-7)
* Sc2: the group summary is inconsistent with the individual expert in that it presents a clearly different median (the median is the 85%/15% percentile of the individual’s distribution if the individual’s is biased downwards/upwards). The precision of the group’s judgement is the same as the individual’s (central panel in Figure 7).
* Sc3: the group summary is inconsistent with the individual expert and is *more* precise than the summary of the individual (right hand panel in Figure 7). It presents a clearly different median (the median is the 85%/15% percentile of the individual’s distribution if the individual’s is biased downwards/upwards), but the standard deviation is now 1/2 of the individual’s distribution.

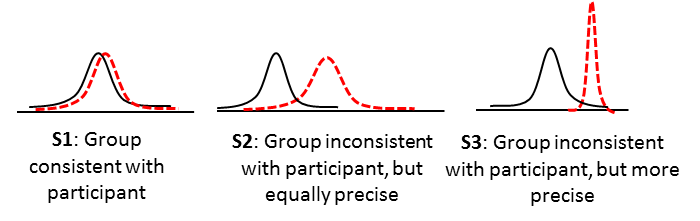


Figure 7: Illustrative example of the scenarios evaluated in experiment 3.1.

After observing the group summaries, participants are asked if they wish to revise their elicited distributions in light of the group summary presented using the same method of elicitation.

An example of scenario 3 using the bisection method is:

*“For this same pill, we have also asked six other (anonymous) participants for their beliefs, in exactly the same way we asked you.*

*The group's summary has been produced by averaging the answers for all six participants. This means that if the group’s combined responses were the same, the group distribution would be the same as the individual’s distribution. It also means that if at least one individual indicated that the proportion may take a particular extreme value, the group distribution will reflect this.*

*As a reminder, you have previously replied that:*

*The proportion of patients is equally likely to be less than and greater than 50 and it is equally likely to be between 40 and 65 as it is to be outside this range.*

*The group has replied that:*

*The proportion of patients is equally likely to be less than and greater than 32 and it is equally likely to be between 26 and 38 as it is to be outside this range.*

*In light of the group assessments, would you like to revise your own assessment?”*

#### Design of experiment

The experiment uses a randomised block design. The factor of interest here is the 3 scenarios for the group summaries defined above (section 6.1.2). We want to still use both methods of elicitation (bisection and chips and bins) and hence the method is considered a nuisance blocking factor. A summary diagram for the tasks and process of sub-experiment 3.1 is shown in Figure 8.

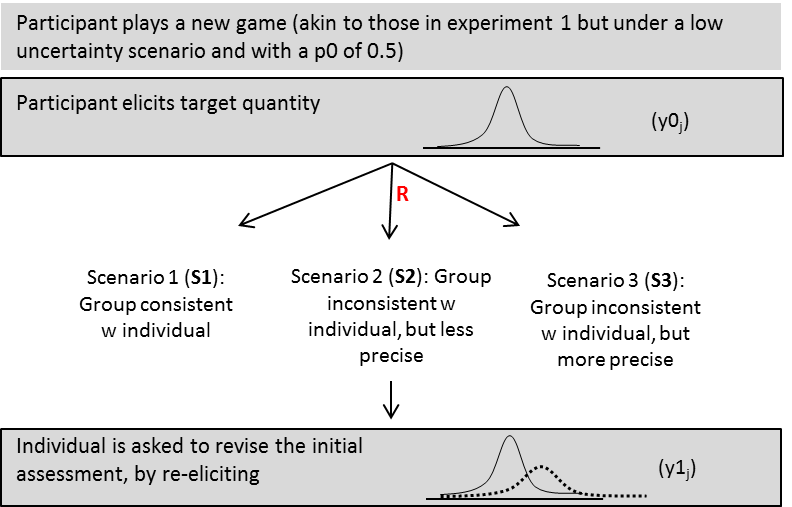


Figure 8: Summary diagram of the processes and tasks for experiment 3.1

#### Inference

Analyses will examine the extent of revision of distributions elicited at the individual level, from before to after the participant being presented with group summaries. The extent of revision will be measured by the difference in KL,  for each individual j by comparing their initial and final (revised) distributions. The following model will be used:

|  |  |
| --- | --- |
| E(j)= g-1(mu + b1\*y0j + b2\*I(Sj==2) + b3\*I(Sj==3) + b4\*Xj + ej), | (Equation 3) |

where j is the subject, y0j is individual’s j KL for the initial distribution, Sj is the group distribution individual j was randomised to, I() the indicator function that assumes a value of 1 of the enclosed statement is true and 0 if false, and Xj the method of elicitation used by subject j (0 if chips and bins is used or 1 if bisection is used, here a blocking factor). As with experiments 1 and 2, a GLM framework will be used allowing different link functions (g) and distributions to be tested.

The hypotheses tested in this experiment are:

Ha1: b1≠0 or b2≠0

Ha2: a≠0

### Experiment 3.2: How does between-expert variation within a group affect individuals’ revision.

#### Hypothesis to be tested

The level of uncertainty in a group’s summary can reflect either higher levels of within-expert uncertainty or higher levels of between-expert variation. In this experiment, we aim is to examine how individuals revise their estimates when uncertainty in the group distribution arises from either within-expert variation or from between-expert variation. For such, participants will be presented with disaggregated results for each member of a group, where the group distribution is discordant with the individual’s. Two alternative scenarios will be used to generate the same group estimate:

. Sc1: In the first scenario (the top panel in Figure 9), there is low variation in the judgements of other members of the group (they are concordant amongst themselves but not with the participant); uncertainty is hence mostly derived from uncertainty in the individual responses.

. Sc2: In the second scenario (the bottom panel in Figure 9), there is high variation between experts (they are discordant amongst themselves). However, uncertainty in the group’s distribution is mainly due to between-expert variation, and the hence the individual judgements of other members are more precise than in scenario 1.

|  |
| --- |
| Scenario 2: Higher between-expert variation  Scenario 1: Higher within-expert uncertainty  Figure 9: Illustration of the scenarios defined for experiment 3.2 |

Three hypotheses will be tested in this experiment:

H1: Do individuals revise their answers when faced with different levels of between-expert variation in a group?

H2: Does the extent of revision depend on initial performance?

H3: Does the extent of revision depend on how different the split in the sample is from the target population?

#### Experimental set-up

This experiment uses the set-up of experiment 1. Participants are asked to play a ‘low uncertainty game’ with a p0 value of 0.3 and asked to elicit uncertainty for p0, the target quantity. This will be used as an initial distribution for individual *j*. Its associated KL measurement (y0,j), as with experiment 1, is a measure of probabilistic accuracy. The key to this experiment is that participants are then presented with the distributions of other individuals, and asked whether they’d like to revise their judgements in light of this information. This differs from experiment 3.1 as individual summaries, rather than only a group summary, are presented to participants. Participants will either see that other members of the group may be 1) consistent but imprecise, or 2) discordant but precise.

In experiment 3.2, the group is composed of 4 members (the participant plus other 3 hypothetical individuals). The group distribution is implicit, i.e. it is not presented to the participant. The group mean is discordant with the individual’s elicited distribution (the group mean is set to the 80th/20th percentile of the individual’s when biased downwards/upwards, respectively).

-- Scenario 2 (members of the group are discordant). For one element, the mean is equal to the group’s, for another the mean is equal to the 99.5%/0.5% percentile of the participant’s distribution if biased downwards/upwards, and the third will assume the value required to maintain the group mean. The variance of the additional members of the group is 1/10 of the participant’s variance. With this information, the variance of the combined (group) distribution can be calculated from the within-member variance and the between-member variance of the means. We make the assumption that the implied sample size for the individual distributions () is equal, and that the distributions from different individuals are independent. The estimated combined variance is:

|  |  |
| --- | --- |
|  | (Equation 4) |

where is the mean for group and is the overall mean.

The scenario where other members of the group are concordant assumes the same group summaries. The mean of the distribution for one of the additional member is the group’s and for the remaining two is equal and its value is what is required to maintain the group mean. The means for these latter two individuals are made to differ slightly (+-5% of the difference between the participant’s mean and the group mean). The variance of the distributions of the additional members is assumed equal amongst them and is calculated from the relation established in Equation 4.

The distributions obtained for each expert are then converted onto elicited summaries for the method being used (either chips and bins or bisection).

The following exemplifies the type of information participants will be presented with in scenario 2 and when using the chips and bins method:



Figure 10: Snapshot of the R shiny app (6).

After seeing the summaries, participants are asked to revise (if they wish) their original estimates.

#### Design

Participants are randomised to receive one of the two scenarios described above; randomisation is blocked according to the method of elicitation. A flow diagram of the process and tasks required of participants is shown in Figure 11.

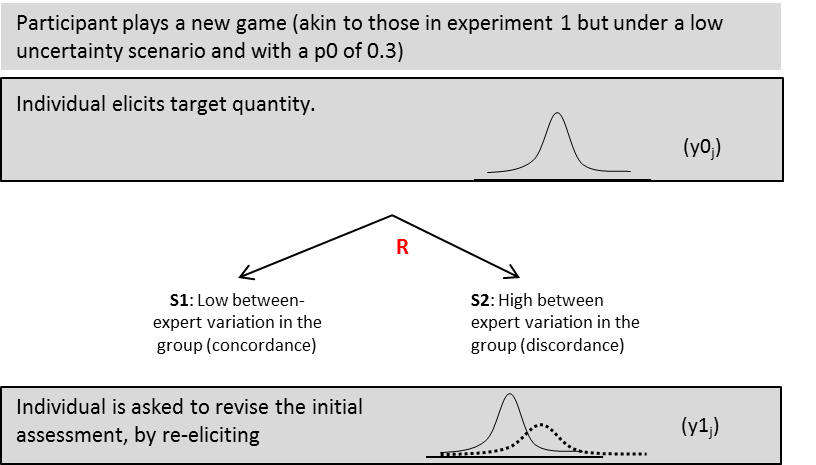


Figure 11: Summary diagram of the processes and tasks for experiment 3.2

#### Model for experiment 3.2

The dependent variable will be the KL comparing the revised distribution with the originally elicited, , for each individual j. The following model will be used

|  |  |
| --- | --- |
| E(j)= g-1(mu + b1\*y0j + b2\*I(Sj==2) + b3\*I(Oj==70) + b4\*I(Oj==80) +b5\*Xj + ej) | (Equation 5) |

where j is the subject, y0j is individual’s j KL for the initial distribution (the dependent variable in experiment 2), Sj is the scenario individual j was randomised to and I() the indicator function that assumes a value of 1 of the enclosed statement is true and 0 if false. The remaining are blocking variables retained from experiment 2: Oj the split individual j was randomised to in experiment 2 and Xj the method of elicitation used by subject j (0 if chips and bins is used or 1 if bisection is used). As with experiments 1, 2 and 3.1, a GLM framework will be used allowing different link functions (g) and distributions to be tested.

Based on this model, the hypotheses being tested are:

Ha1: b2≠0

Ha2: b1≠0

Ha3: b3≠0 or b4≠0

## Training

A bespoke training package was developed for this exercise. It drew on other training packages available to the research team, such as the SHELF training package [available at http://www.tonyohagan.co.uk/shelf/ecourse.html] and training sessions developed by the research team for other exercises. The training package aims to cover a number of aspects that are detailed in the next headings. A script was written for the training session to ensure that the same background material was presented to participants in a standardised way in all elicitation sessions. The training session was delivered face to face via power point presentation. Practice questions were delivered via the SHINY app.

### Motivation

Participants will be introduced to elicitation, and to why elicitation is needed in health research. Participants will be told that we are exploring different methods for elicitation, and while all examples in the exercise are hypothetical, the findings may be used in the future, when clinical participants are asked to express their uncertainty around real quantities in health research. We will emphasise that they are contributing to work funded by the MRC.

We will tell participants that it is their opinion that we are interested in so they shouldn’t communicate with others sitting near them or write down notes. They will not be identified personally in any reporting of results. They should try to concentrate on the task throughout. They will be informed about the remuneration; however they will not be given specific details about the incentive structure. We will also give them a brief overview of the tasks that they will undertake during the experiment.

*Training in probabilities, uncertainty and probability distributions*

Training will cover the following content:

* + What are probabilities?
  + What are probability distributions and how can they be used to express uncertainty?
  + Explain the difference between uncertainty (in mean values) and variations observed across patients. Introduce the factors that might affect uncertainty (e.g. sample size and heterogeneity).
  + Give pointers on avoiding biases relevant in this context (overconfidence, anchoring, availability heuristic, representativeness heuristic).
  + Give examples of how to express different levels of uncertainty

### Training in elicitation techniques

Both elicitation methods (bisection and chips and bins methods) will be explained thoroughly using examples. Visual aids will be used to explain concepts key to the methods, for example the median and upper and lower quartiles.

### Game training

Participants will be told the general rules of the game: there will be multiple actions with different consequences, they will be required to optimise the outcomes to the patients they treat (the better the outcome, the higher the reward).

There will be a practice session of both methods (without the game component). For this, individuals will here be asked about the proportions of UK household owning a pet and, after eliciting, they will be given the value for this quantity (but no uncertainty summary will be reported). Additionally, two practice games will also be included. The practice game will be similar to the ‘real’ game. The game training will include elicitation.

## 8. References

[1] Expert Elicitation Task Force. Expert Elicitation Task Force White Paper In: Agency USEP, ed. 2009.

[2] European Food Safety Authority. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. EFSA Journal. 2014;12(6).

[3] Gosling J. Methods for eliciting expert opinion to inform health technology assessment. 2014 [cited; Available from: https://mrc.ukri.org/documents/pdf/methods-for-eliciting-expert-opinion-gosling-2014/

[4] O'Hagan A. Uncertain judgements : eliciting experts' probabilities. London ; Hoboken, NJ: John Wiley & Sons 2006.

[5] Wang HQ, Dash D, Druzdzel MJ. A method for evaluating elicitation schemes for probabilistic models. Ieee T Syst Man Cy B. 2002 Feb;32(1):38-43.

[6] Cokely ET, Galesic M, Schulz E, Ghazal S, Garcia-Retamero R. Measuring Risk Literacy: The Berlin Numeracy Test. Judgm Decis Mak. 2012 Jan;7(1):25-47.

[7] Risk Literacy: science for informed decision making,. [cited 2018 05/09/2018]; Available from:

[8] Scott SG, Bruce RA. Decision-Making Style - the Development and Assessment of a New Measure. Educ Psychol Meas. 1995 Oct;55(5):818-31.

[9] Read D. Monetary incentives, what are they good for? Journal of Economic Methodology. 2005 2005/06/01;12(2):265-76.

[10] Soares MO, Sharples L, Morton A, Claxton K, Bojke L. Experiences of Structured Elicitation for Model-Based Cost-Effectiveness Analyses. Value Health. 2018 Jun;21(6):715-23.

[11] Garthwaite PH, Kadane JB, O'Hagan A. Statistical methods for eliciting probability distributions. J Am Stat Assoc. 2005 Jun;100(470):680-700.

[12] Seaver DA, Winterfeldt DV, Edwards W. Eliciting Subjective-Probability Distributions on Continuous Variables. Organizational Behavior and Human Performance. 1978;21(3):379-91.

[13] Murphy AH, Winkler RL. Credible Interval Temperature Forecasting - Some Experimental Results. Mon Weather Rev. 1974;102(11):784-94.

[14] Grigore B, Peters J, Hyde C, Stein K. A comparison of two methods for expert elicitation in health technology assessments. BMC Med Res Methodol. 2016 Jul 26;16:85.

[15] Rowe G, Wright G. The Delphi technique as a forecasting tool: issues and analysis. Int J Forecasting. 1999 Oct;15(4):353-75.

## Appendix 1: Berlin Numeracy Test and Scott and Bruce's Decision Style inventory

*Berlin Numeracy Test*

1. Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir? Please indicate the probability in percent. \_\_\_25%\_\_\_

2a. Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)? \_\_30\_\_ out of 50 throws.

2b. Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6? \_\_\_20\_\_\_out of 70 throws.

3. In a forest 20% of mushrooms are red, 50% brown and 30% white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red? \_\_\_50%\_\_\_

*Scott and Bruce's General Decision-making Style questionnaire*

Listed below are statements describing how individuals for about making *important decisions*. Please indicate whether you agree or disagree with each statement.

(responses are made on a 5-point scale from strongly agree to strongly disagree)

1. I double-check my information sources to be sure I have the right facts before making decisions. (*Rational decision-making style*)
2. I make decisions in a logical and systematic way. (*Rational decision-making style*)
3. My decision making requires careful thought. (*Rational decision-making style*)
4. When making a decision, I consider various options in terms of a specified goal. (*Rational decision-making style*)
5. When making decisions, I rely upon my instincts. (*Intuitive decision-making style*)
6. When I make decisions, I tend to rely on my intuition. (*Intuitive decision-making style*)
7. I generally make decisions that feel right to me. (*Intuitive decision-making style*)
8. When I make a decision, it is more important for me to feel the decision is right than to have a rational reason for it. (*Intuitive decision-making style*)
9. When making a decision, I trust my inner feelings and reactions. (*Intuitive decision-making style*)
10. I often need the assistance of other people when making important decisions. (*Dependent decision-making style*)
11. I rarely make important decisions without consulting other people. (*Dependent decision-making style*)
12. If I have the support of others, it is easier for me to make important decisions. (*Dependent decision-making style*)
13. I use the advice of other people in making my important decisions. (*Dependent decision-making style*)
14. I like to have someone steer me in the right direction when I am faced with important decisions. (*Dependent decision-making style*)
15. I avoid making important decisions until the pressure is on. (*Avoidant decision-making style*)
16. I postpone decision making whenever possible. (*Avoidant decision-making style*)
17. I often put off making important decisions. (*Avoidant decision-making style*)
18. I generally make important decisions at the last minute. (*Avoidant decision-making style*)
19. I put off making decisions because thinking about them makes me uneasy. (*Avoidant decision-making style*)
20. When making decisions I do what feels natural at the moment. (*Spontaneous* *decision-making style*)
21. I generally make snap decisions. (*Spontaneous* *decision-making style*)
22. I often make impulsive decisions. (*Spontaneous* *decision-making style*)
23. I often make decisions on the spur of the moment. (*Spontaneous* *decision-making style*)
24. I make quick decisions. (*Spontaneous* *decision-making style*)
25. I explore all my options before making a decision (*Rational decision-making style*)

## Appendix 2: Randomisation sequences

Key

|  |  |  |
| --- | --- | --- |
|  | M (method) | P (precision) |
| A | 1 | 1 |
| B | 1 | 2 |
| C | 2 | 1 |
| D | 2 | 2 |
| M= 1 for bisection and M=2 for chips and bins | | |
|  |  |  |
|  | P (probability parameter) |  |
| a | 0.3 |  |
| b | 0.4 |  |
| c | 0.6 |  |
| d | 0.7 |  |
|  |  |  |
|  | (P7) probability parameter for game 7 |  |
| 1 | 0.6 |  |
| 2 | 0.7 |  |
| 3 | 0.8 |  |

**Randomisation sequences**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | MxP | MxP | MxP | MxP | G5\_M | G6\_M | P7\_M |
| **ID** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| 1 | Ab | Cd | Da | Bc | 1\_0 | 1\_0 | 2\_1 |
| 2 | Da | Cd | Bc | Ab | 2\_1 | 2\_0 | 3\_1 |
| 3 | Ca | Bb | Dd | Ac | 2\_1 | 1\_1 | 2\_1 |
| 4 | Ab | Bc | Cd | Da | 3\_0 | 2\_0 | 3\_1 |
| 5 | Dd | Ac | Bb | Ca | 1\_0 | 2\_0 | 1\_1 |
| 6 | Dc | Cb | Ba | Ad | 2\_0 | 1\_1 | 2\_0 |
| 7 | Ac | Dd | Bb | Ca | 1\_1 | 2\_0 | 1\_0 |
| 8 | Ba | Ad | Dc | Cb | 1\_1 | 1\_0 | 2\_1 |
| 9 | Cd | Ab | Bc | Da | 2\_0 | 1\_1 | 1\_0 |
| 10 | Ba | Ad | Cb | Dc | 2\_1 | 2\_0 | 2\_1 |
| 11 | Ac | Dd | Ca | Bb | 3\_1 | 2\_1 | 1\_0 |
| 12 | Cb | Dc | Ad | Ba | 2\_0 | 2\_0 | 2\_0 |
| 13 | Ad | Dc | Ba | Cb | 1\_0 | 2\_0 | 1\_0 |
| 14 | Cd | Da | Ab | Bc | 3\_1 | 2\_1 | 2\_0 |
| 15 | Cd | Bc | Ab | Da | 2\_0 | 1\_0 | 1\_0 |
| 16 | Db | Cc | Bd | Aa | 2\_1 | 1\_0 | 2\_1 |
| 17 | Da | Ab | Cd | Bc | 3\_1 | 1\_1 | 1\_1 |
| 18 | Bc | Cd | Da | Ab | 1\_1 | 1\_0 | 2\_0 |
| 19 | Bd | Cc | Aa | Db | 1\_1 | 2\_1 | 2\_1 |
| 20 | Ad | Cb | Ba | Dc | 1\_0 | 1\_1 | 1\_1 |
| 21 | Bb | Ac | Ca | Dd | 3\_0 | 1\_0 | 1\_0 |
| 22 | Dd | Ca | Ac | Bb | 1\_0 | 1\_1 | 3\_1 |
| 23 | Ca | Bb | Ac | Dd | 1\_0 | 2\_1 | 2\_0 |
| 24 | Dc | Ad | Ba | Cb | 2\_1 | 1\_0 | 2\_1 |
| 25 | Ac | Bb | Ca | Dd | 1\_1 | 1\_1 | 1\_0 |
| 26 | Ad | Ba | Cb | Dc | 2\_0 | 2\_0 | 1\_1 |
| 27 | Cd | Da | Bc | Ab | 3\_0 | 2\_1 | 3\_1 |
| 28 | Dd | Ca | Bb | Ac | 1\_1 | 2\_1 | 1\_0 |
| 29 | Ad | Dc | Cb | Ba | 3\_1 | 2\_0 | 2\_0 |
| 30 | Cb | Ba | Dc | Ad | 1\_1 | 2\_1 | 2\_0 |
| 31 | Dc | Cb | Ad | Ba | 3\_0 | 2\_0 | 1\_1 |
| 32 | Db | Bd | Cc | Aa | 3\_0 | 2\_1 | 1\_0 |
| 33 | Db | Aa | Cc | Bd | 3\_0 | 1\_1 | 3\_1 |
| 34 | Bd | Db | Aa | Cc | 1\_1 | 1\_0 | 3\_1 |
| 35 | Db | Aa | Bd | Cc | 3\_1 | 1\_0 | 3\_0 |
| 36 | Ba | Dc | Cb | Ad | 3\_0 | 2\_0 | 2\_0 |
| 37 | Bb | Dd | Ca | Ac | 3\_0 | 1\_1 | 2\_0 |
| 38 | Cc | Db | Aa | Bd | 3\_1 | 2\_0 | 1\_1 |
| 39 | Bc | Ab | Da | Cd | 2\_1 | 1\_0 | 3\_0 |
| 40 | Ca | Dd | Ac | Bb | 3\_1 | 1\_1 | 1\_0 |
| 41 | Da | Bc | Cd | Ab | 2\_1 | 2\_0 | 3\_1 |
| 42 | Dd | Bb | Ac | Ca | 1\_0 | 2\_1 | 1\_1 |
| 43 | Ba | Cb | Dc | Ad | 3\_1 | 2\_1 | 3\_0 |
| 44 | Bd | Cc | Db | Aa | 2\_0 | 2\_1 | 1\_0 |
| 45 | Aa | Bd | Cc | Db | 3\_0 | 1\_1 | 3\_1 |
| 46 | Bc | Cd | Ab | Da | 1\_0 | 2\_1 | 3\_0 |
| 47 | Ab | Da | Cd | Bc | 3\_1 | 1\_1 | 2\_0 |
| 48 | Ab | Bc | Da | Cd | 3\_0 | 1\_0 | 2\_1 |
| 49 | Da | Ab | Bc | Cd | 1\_1 | 2\_1 | 2\_1 |
| 50 | Ca | Ac | Dd | Bb | 1\_0 | 2\_1 | 3\_0 |
| 51 | Cb | Ad | Dc | Ba | 2\_0 | 1\_0 | 2\_0 |
| 52 | Aa | Db | Cc | Bd | 2\_1 | 2\_0 | 3\_1 |
| 53 | Cc | Aa | Bd | Db | 3\_1 | 1\_1 | 1\_1 |
| 54 | Ac | Ca | Bb | Dd | 3\_0 | 2\_1 | 2\_0 |
| 55 | Bb | Ca | Dd | Ac | 1\_1 | 1\_0 | 1\_1 |
| 56 | Cc | Db | Bd | Aa | 3\_1 | 2\_1 | 3\_0 |
| 57 | Dc | Ba | Ad | Cb | 1\_0 | 1\_1 | 3\_0 |
| 58 | Cb | Ba | Ad | Dc | 1\_0 | 1\_1 | 2\_1 |
| 59 | Aa | Cc | Db | Bd | 1\_1 | 2\_1 | 3\_1 |
| 60 | Bc | Da | Ab | Cd | 2\_0 | 1\_0 | 3\_0 |
| 61 | Aa | Bd | Db | Cc | 2\_0 | 2\_0 | 1\_0 |
| 62 | Bb | Ac | Dd | Ca | 2\_1 | 2\_0 | 3\_0 |
| 63 | Bd | Aa | Db | Cc | 3\_1 | 2\_0 | 3\_0 |
| 64 | Cc | Bd | Aa | Db | 2\_1 | 1\_0 | 3\_1 |
| 65 | Ab | Bc | Da | Cd | 2\_0 | 1\_0 | 1\_1 |
| 66 | Ab | Cd | Da | Bc | 2\_1 | 2\_0 | 3\_0 |
| 67 | Da | Ab | Cd | Bc | 2\_0 | 1\_0 | 2\_1 |
| 68 | Ad | Dc | Ba | Cb | 2\_1 | 1\_1 | 3\_0 |
| 69 | Aa | Db | Cc | Bd | 1\_1 | 1\_0 | 1\_1 |

G5 represents the group allocation in game 5 (experiment 3.1) and G6 the group allocation in game 6 (experiment 3.2)

**Prior information given to participants**

All participants received the same prior information. This was independent of the random allocation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Game 1 | Game 2 | Game 3 | Game 4 | Game 5 | Game 6 |
| r[1] | 3 | 2 | 3 | 2 | 3 | 3 |
| r[2] | 4 | 3 | 2 | 1 | 2 | 2 |
| n | 6 | 4 | 5 | 3 | 5 | 4 |

priord<-c(12,8,10,6,10,8)

## Appendix 3: Calculating Kullback-Leibler divergence

We wish to assess elicited approximations to a true posterior. We may not know the true posterior analytically but through an MCMC sample. The following R function calculates the Kullback-Leibler discrepancy of an elicited distribution, given by a fitted Beta(mu\*m, (1-mu)\*m) , to a true distribution given by an MCMC sample, y . It uses a kernel density estimate of the true distribution, and numerical integration. The following R function will be used to calculate KL:

kl <- function(y, mu, m){

d <- density(y, from=0, to=1)

dfn <- with(d, approxfun(x = x, y = y))

fn <- function(x){

p <- dfn(x)

q <- dbeta(x, mu\*m, (1-mu)\*m)

p\*(log(p) - log(q))

}

integrate(fn, 0, 1)

}

1. Throughout, the term ‘belief’ or ‘degree of belief’ refers to the construct, whilst the term ‘subjective probability’ refers to quantifications of the degree of belief [↑](#footnote-ref-1)
2. Two general types of uncertainty exist: aleatory and epistemic. Epistemic uncertainty is subjective in nature and arises primarily from limited or imperfect knowledge. It is, in principle, reducible by obtaining more, or better, information. Uncertainty in parameters of a statistical model is generally epistemic. [4] O'Hagan A. Uncertain judgements : eliciting experts' probabilities. London ; Hoboken, NJ: John Wiley & Sons 2006. [↑](#footnote-ref-2)
3. O’Hagan [4] ibid. identifies that this stream of elicitation research cannot distinguish ‘Inaccurate knowledge from poor elicitation’. We here avoid using the term “knowledge” as is often taken to be “justified true belief” and hence cannot be inaccurate (though this is controversial, see http://www.iep.utm.edu/gettier/#H2). [↑](#footnote-ref-3)
4. The target quantity for the elicitation is conditional on treatment, hence is independent of the other pill. [↑](#footnote-ref-4)
5. [↑](#footnote-ref-5)
6. We considered an alternative scenario of treatment vs no treatment, but considered that subjects may presume treatment is effective [↑](#footnote-ref-6)
7. The median of the Beta distribution does not present a closed form solution, hence to find values for the alpha and beta parameters using the median and variance (a system of two equations) numerical methods were used, specifically those implemented within the nleqslv package in R (Broyden or a full Newton method). The absolute differences in the median and variance were used as targets for the optimisation. [↑](#footnote-ref-7)