This paper is accepted for and published in *The Knowledge Engineering Review*, Vol. 16:3, 255–269. ©2001, Cambridge University Press
DOI: 10.1017/S0269888901000145

Probability Elicitation for Belief Networks: Issues to Consider

Silja Renooij

Institute of Information and Computing Sciences, Utrecht University,
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
e-mail: silja@cs.uu.nl

Abstract

Upon assessing probabilities for Bayesian belief networks, the knowledge and practical experience of experts is often the only available source of probabilistic information. It is important to realise that issues concerning the human capabilities with respect to making judgements come into play when relying on experts for probability elicitation. A number of methods for the elicitation of probabilities are known from the field of decision analysis. These methods try, to some extent, to deal with those issues. We present an overview of the issues to consider when relying on expert judgements and describe the methods that are available for expert elicitation, along with their benefits and drawbacks.

1 Introduction

In the late 1980s, *Bayesian belief networks* were introduced for representing and reasoning with uncertainty [Pearl, 1988]. Belief networks consist of a qualitative part and a quantitative part. The qualitative part is a directed graph with nodes modelling statistical variables and arcs representing the probabilistic influences between these variables. The strengths of these influences are captured by probabilities specified for each value of a variable, conditioned on every possible combination of values for the parents (in the graph) of that variable. These probabilities constitute the quantitative part of the network.

As more and more Bayesian belief networks are being developed for complex real-life problem domains, it is becoming increasingly apparent that the construction of the qualitative part with the help of domain experts is feasible; the elicitation of the large number of probabilities required, however, is a far harder task. In fact, the elicitation of probabilities is often referred to as a major obstacle in building a Bayesian belief network [Druzdzel & Van der Gaag, 1995, Jensen, 1995]. In most problem domains, various sources of probabilistic information are available. Examples of such sources are

databases, for example containing patient records, and literature. Unfortunately, databases are often incomplete, biased, and not large and rich enough to allow for reliable assessment of the required probabilities. In research reports, the results published hardly ever match the conditional probabilities required for a belief network under construction [Druzdzel & Van der Gaag, 1995]. When the above mentioned sources do not contain the necessary probabilistic information, the single remaining source is the knowledge and experience of a domain expert.

Extensive psychological research has shown that people, even experts, tend to find it difficult to assess probabilities; to simplify this task they use heuristics, most often leading to poorly calibrated and biased assessments [Kahneman, Slovic & Tversky, 1982]. From the field of decision analysis, several methods are available for the elicitation of probabilities [Cooke, 1991, Winterfeldt & Edwards, 1986, Morgan & Henrion, 1990]. These methods are designed for the elicitation of probabilities in general and not tailored to probability elicitation for belief networks. Most of these methods have been designed to overcome, or at least suppress, the problems of bias and poor calibration. However, these methods tend to be so time-consuming that it is infeasible to apply them when hundreds or thousands of probabilities are to be assessed. Faster elicitation methods are available, but are prone to even more biased answers. Before undertaking a large elicitation task, it is therefore important to be aware of the advantages and drawbacks of these methods.

In the field of belief networks, it is well-known that probability elicitation is a problem. We feel, though, that the knowledge about why it is a problem is less wide-spread; it is also less known that there exist various methods designed especially for probability elicitation. Besides being aware of problems of bias, the builder of a network has to take into consideration not only the method to use, but also, for example, which expert to choose, how to motivate and train the expert, and how to perform the actual elicitation. In this paper we will give an overview of the entire elicitation process and the available methods, discussing issues to be aware of and to take into consideration when faced with the task of probability elicitation.

This paper is organised as follows. In Section 2 we will first discuss the process of probability elicitation, including motivating and training the expert, the actual elicitation phase, and the verification of the probabilities obtained. Then, in Section 3, we will consider the different ways an expert can be presented with the probabilities required and the representation formats that experts can use for indicating their assessment. In Section 4 we will discuss various elicitation methods found in, for example, the decision analysis literature, along with their benefits and their drawbacks. As our main concern is probability assessment for belief networks, we will only consider methods for eliciting *discrete* probability distributions. We are interested in point probabilities and will therefore not consider elicitation methods for interval probabilities. Finally, in Section 5 we will discuss some matters concerning elicitation methods in general, and draw some conclusions.

2 The Elicitation Process

Research in experimental psychology has shown that simply asking a person to provide a (numerical) probability results in biased probability judgements [Kahneman, Slovic & Tversky, 1982]. To overcome biases, it seems necessary to have a well-structured process for probability elicitation. Such a process is called an *elicitation process* [Fenton, 1998, Merkhofer, 1987, Edwards, 1995]; it can be roughly divided into five stages:

1. select and motivate the expert
2. train the expert
3. structure the questions
4. elicit and document the expert judgements
5. verify the results.

We will further detail these stages in the following subsections, after devoting a subsection to the biases that call for a well-structured elicitation process.

2.1 Heuristics and Biases

A *bias* is a systematic tendency to take into account factors that are irrelevant to the task at hand, or to ignore relevant facts, thereby failing to make an inference that any appropriate normative theory, for example probability theory, would classify as necessary [Evans & Over, 1996]. Two types of bias can be distinguished: motivational bias and cognitive bias [Skinner, 1999]. *Motivational biases* are caused by personal interests and circumstances of the expert. For example, an expert makes careful assessments if he[1] believes that his job depends on the success of the current project; he will be too confident about his assessments, because he, being an expert, feels he should not be uncertain about them. Motivational biases can often be overcome by explaining to the expert that an honest assessment is requested, not a promise. *Cognitive biases* arise during the processing of information by the expert and are typically the result of using heuristics [Kahneman, Slovic & Tversky, 1982]. Cognitive biases can, to some extent, be suppressed by informing the expert of their existence and by using different elicitation methods.

When people are asked to make complicated judgements such as probability assessments, they often *subconsciously* use *heuristics*, or rules of thumb, to simplify the task. Four heuristics, among others, are commonly found: availability, anchoring, representativeness, and control [Kahneman, Slovic & Tversky, 1982]. *Availability* is a heuristic with which an expert assesses the probability of an event by the ease with which occurrences of the event are brought to mind. The idea behind the heuristic is that frequent events

[1] For any occurrence of a masculine pronoun, the feminine form is understood to be included.

are more available, and therefore an event that is easily brought to mind will have a high probability. Often this heuristic works quite well, but it can become a misleading indicator of the frequency with which certain events occur. If, for example, plane crashes are headline news more often than car crashes, people will *overestimate* the probability of a plane crash and *underestimate* the probability of being involved in a car crash. The process of assessing a probability by choosing an initial value, termed the anchor, and then adjusting up or down from this value, is called the heuristic of *anchoring and adjustment*. Assessments acquired this way are typically biased towards the starting value, due to insufficient adjustment. The resulting bias is termed *anchoring bias*.

The *representativeness* heuristic describes the process where people use the similarity of two events to estimate the degree to which one event is representative of the other. Consider the following well-known example from a study by Tversky and Kahneman [Kahneman, Slovic & Tversky, 1982]:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Please check off the most likely alternative:
>
> ☐ Linda is a bank teller.
>
> ☐ Linda is a bank teller and is active in the feminist movement.

From this description, most people find it likely that Linda is a feminist and therefore conclude that it is more likely that Linda is a feminist bank teller, than just a bank teller. The example illustrates how a description representative of a feminist can trick people into choosing the less likely event. For this example, the cognitive bias introduced by the representativeness heuristic is called the *conjunction fallacy*. A more detailed description seems to be more representative, though the conjunction of two events can never be more likely than the probability of either event alone. Other well-known biases introduced by the representative heuristic are the *gambler's fallacy* and *base-rate neglect*. The gambler's fallacy is the belief that when a series of trials all have the same outcome then soon an opposite outcome will follow. This belief originates from the idea that random sequences of outcomes seem more representative of a sample space. Base-rate neglect is neglecting the relative frequency with which an event occurs. This is again illustrated by an example from Tversky and Kahneman, where a group of subjects is presented with the following description of a person who they know stems from a population of 30 engineers and 70 lawyers:

> Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.

This description is entirely uninformative with respect to Dick's profession. However, when subjects were asked to indicate the probability of Dick being an engineer, the subjects

gave a median probability estimate of 50%, whereas the correct answer would have been 30%. The subjects ignored the base-rate and simply judged the description as equally representative of an engineer or a lawyer.

The *control* heuristic represents the tendency of people to act as if they can influence a situation over which they actually have no control. For example, people value a lottery ticket they selected themselves more highly than an arbitrary one given to them, even though the probability of winning a prize is the same for both tickets [Anderson, 1998]. This *illusion of control* can cause overestimation of probabilities.

We have seen that the use of heuristics can introduce cognitive biases in probability assessments. The most prevalent biases are said to be overconfidence and base-rate neglect [Baecher, 1998]. Overconfidence is especially a problem with extreme probabilities, that is probabilities close to 0% or 100%. People find extreme probabilities hard to assess; they are less likely to be overconfident about probability judgements that lie more in the centre of the $0\% - 100\%$ range [Winterfeldt & Edwards, 1986].

2.2 Selection and Motivation

Ideally, for probability elicitation, an expert should be selected who has the necessary domain knowledge and who is familiar with assessing probabilities. However, due to the nature of expertise (it is by definition a scarce commodity), there is often not a very large pool of experts to choose from. When eliciting probabilities for belief networks, it is best to select an expert who has also been involved in building the structure of the network, to prevent errors due to the possible existence of different definitions for certain variables. It is also better to have more than one expert involved [Clemen & Winkler, 1999, Winterfeldt & Edwards, 1986], since different experts have different kinds of knowledge, all of which should be incorporated in the assessment. Assessments by more than one expert can be handled in two ways: collect the assessment of each expert and combine the assessments into a single one, or have the experts come to a consensus. The first approach has the mathematical advantage of enlarging the sample space, but assumes that nothing is gained from sharing knowledge and thought among the experts. With the second approach, group interaction problems, such as dominance of one expert over the other or pressure for conformity, can influence the assessment. Research on the subject of group assessment suggests that an optimal number of experts is around three [Clemen & Winkler, 1999].

Once the experts have been selected, the elicitation task is introduced and its purpose is explained. The elicitation task will often be part of a larger process of step-wise refinement [Coupé, Van der Gaag & Habbema, 2000], where the experts are first asked to provide only initial assessments. With these assessments, a sensitivity analysis of the belief network is performed, revealing the most sensitive parts of the network; the most sensitive probabilities can then be refined, and so on. Refinement of the most sensitive probabilities is done by using additional information obtained from other sources than the experts involved, such as research reports or other experts. It has been observed that experts may feel that the assessments they are asked to provide are not subjective opinions, but numbers that can be checked in every-day practice [Van der Gaag *et al.*, 1999]. They then

have the uncomfortable feeling that the assessments they provide should be "correct" and this makes them less willing to cooperate. It is therefore important to convince the experts that their assessments need only be accurate in the sense that they should represent the knowledge and judgement of the expert: there are no right and wrong answers. Also, it may reassure the experts when it is explained to them that their initial assessments will be subjected to a sensitivity analysis and that they can thereupon refine their assessments.

Experts should also be informed about the biases discussed in the previous subsection; knowledge of their existence might help in counteracting them.

2.3 Training

Once an expert has been selected and is willing to cooperate, he has to learn the art of probability assessment. To this end, the expert should first of all become familiar with the concept of probability and should learn to express his knowledge in the format required by the elicitation method used. Part of the training is done with probabilities for events whose frequencies can be checked. This allows for exposing biases in the expert's assessments and to practice the elicitation method. Several elicitation methods and representation formats can be tried to see which best fit the task, the experience and preferences of the expert.

Feedback of the true frequencies of the events for which probabilities are assessed will help experts calibrate their responses [Baecher, 1998], that is, teaches them to make assessments as close as possible the true frequencies. However, care should be taken to not discourage the experts by the confrontation with their frequent mistakes. The events for which probabilities need to be assessed in a belief network are often unobservable, making feedback impossible. The expert must, however, also become an expert at making probability judgements in this domain and part of the training should therefore be done with probabilities from the domain of the belief network [Edwards, 1995].

The amount of time spent on training depends on available time and other constraints. At the end of the training period, however, the expert should fully understand and feel comfortable with the methods to be used.

2.4 Structuring

Before the actual elicitation takes place, several issues need to be addressed. The definitions of the variables and values for which probabilities are to be assessed should be documented so that this information can be easily and promptly conveyed to the expert during the elicitation. For belief networks this documentation will be already available from the construction of the graphical part of the network. Since probability elicitation is often done with the expert who was also involved in the construction of the graphical part, the expert will already be familiar with these definitions; it is, however, always a good idea to keep the documentation of definitions of variables and their values at hand during the elicitation interviews.

After the important variables and values are determined, the conditioning circumstances that influence a variable's uncertainty need to be determined. For belief networks,

the conditioning contexts for a variable follow directly from the structure of the network: the conditioning contexts are all possible combinations of values of the parents of the given variable in the network's directed graph. The number of probabilities to be assessed for each variable in a belief network is therefore, in general, exponential in the number of parents of the variable. Often, however, some values of a variable are independent of the values of some of its parents. If the belief network's graph is constructed using *fragments* conditioned on a variable's values, these independencies are made explicit and can be exploited during probability elicitation [Laskey & Mahoney, 1997, Mahoney & Laskey, 1996]. Another way of reducing the number of probabilities to be assessed for a variable is by assuming a simple interaction model between a variable and its parents, such as a *disjunctive interaction* [Pearl, 1988]. A disjunctive interaction model specifies that the values of a variable are (almost) a logical-or of the values of its parents. Using such a model, the number of probabilities to be assessed for a variable is linear in the number of its parents. After identifying the conditioning contexts, for each probability to be assessed a question describing this probability should be prepared. To suppress overconfidence and overestimation, questions should be prepared for assessment of an event's probability as well as for its complement(s).

In addition to the choice of elicitation method, the elicitor is faced with the choice of how to present the expert with the questions describing the probabilities that need to be assessed and what format to use for the expert's answers. Whatever representation is used to describe the probabilities to be assessed, the associated questions should be clear and structured in such a way that there is no doubt about the variable a probability pertains to and the conditioning contexts. An attractive format should be prepared for the questions and, if possible, a graphical format for the answers. Experience shows that experts dislike writing numbers for subjective probabilities [Cooke, 1991], since numbers suggest an accuracy that experts feel they cannot provide. The experts prefer to check scales, or place a '$\times$' in a box, etc. We will address the issue of presentation of both questions and answers in further detail in Section 3. The preparation of the questions and answering format may require a large amount of time on the elicitor's part, but it is considered time well-spent [Van der Gaag *et al.*, 1999].

2.5 Elicitation and Documentation

Various people will be present during the actual elicitation interviews. First of all, there will be one or more experts involved, interacting during elicitation [Winterfeldt & Edwards, 1986]. There should be at least one, but preferably two, elicitors present during the elicitation, not in the least to show the experts that the task has sufficient priority for the expert to take it seriously. The elicitor has several tasks:

- He has to clarify the inevitable problems of the experts with the interpretation of questions, definitions of variables and values, etc.
- He has to record all information stated by the experts that cannot be expressed in the answering format, but may still be of use. For example, if an expert is allowed

to express *trends* between conditioning contexts [Van der Gaag *et al.*, 1999], such as "the conditional probabilities for this variable given this context are 10% higher than for that context", it should be carefully recorded what is meant by this trend. Also, if the expert has overestimated the probabilities that pertain to a single conditional probability distribution, such that their sum exceeds 100%, possible information he has stated on the range within which the probabilities should lie, can serve to adjust them.

- It may turn out that certain conditioning contexts necessary to estimate the probabilities of certain variables are incomplete, or that certain contexts turn out to be unnecessary. For belief networks, this indicates that changes have to be made to the structure of the network; it is important to carefully record this information.

- For some probability assessments, the elicitor may expect that certain biases are easily introduced; he should then once more make the experts aware of the biases.

- The elicitor should watch the clock: the elicitation is more taxing for the expert than for the elicitor and therefore sessions should not exceed one hour [Cooke, 1991].

Despite the mentioned tasks, the elicitor should avoid coaching the expert and taking too much control [Winterfeldt & Edwards, 1986, Cooke, 1991]; the expert should feel relaxed, not challenged, for he is the expert and the elicitor is not. The elicitation method that is used should be straightforward, easy to handle, and not difficult to learn [Van Lenthe, 1993]. The various elicitation methods commonly used will be discussed in some detail in Section 4.

2.6 Verification

When all required probabilities have been assessed, the elicitor should verify them. Verification is the process of checking whether the probabilities provided by the expert are well-calibrated (conform to observed frequencies), obey the laws of probability (are coherent) and are reliable [Fenton, 1998]. Checking whether the assessments conform to "reality", is often impossible, since the events for which the probabilities are assessed are often unobservable. Regarding coherence, we can check whether all probabilities that should sum to 100% indeed do so. It is convenient to do this check during the elicitation.

Test-retest reliability [Edwards, 1995] tests whether the expert agrees with his own assessments, that is, whether the expert would provide the same estimates when asked for the same probabilities again. However, when dealing with belief networks, the number of probabilities to be assessed is so large, that it is infeasible to assess them more than once. Instead of testing the reliability of separate assessments, entire probability distributions can be considered. As most probabilities are conditional, the expert can be shown the assessed probability distributions for a certain variable given different conditioning contexts and be asked to check whether the relationships for these different contexts are as he would expect. If not, the expert can adjust some of the assessments. Edwards [Edwards, 1998]

calls this an *antecedent conditions check* and he experienced that when his expert took these relationships into account during elicitation, the probabilities had a high test-retest reliability. Van der Gaag *et al.* observed that their experts spontaneously mentioned these relationships, or trends, during elicitation [Van der Gaag *et al.*, 1999].

An indication of the validity of the assessments can also be obtained by entering observations into the belief network and computing the effect of the observations on the probabilities for certain variables of interest. The outcomes for these variables can then be checked against available data or presented to the expert.

3 Presentation

The presentation issues to be addressed for probability elicitation concern the representation format of the required probabilities, the description format of the questions to be asked, and the answering format. Although we are interested in probabilities, the probability format is not necessarily required for the communication with the expert. The experts can, for example, be asked to provide odds or log-odds, or the most familiar competitor of numerical probability, verbal communication of uncertainty, can be used. When dealing with relatively probable events, probabilities or percentages may be intuitively convenient to experts, but in dealing with rare events, odds or log-odds may be easier because they avoid very small numbers.

Regardless of the format used for uncertainty, the required assessment can be described to the expert in various different ways. The description format used should be conceptually simple and compatible with the expert's abilities. When probabilities are chosen as the format for uncertainty representation, the required probabilities can be described, for example, in mathematical notation.

Example 3.1 Consider the domain of oesophageal carcinoma. We focus on the probabilities concerning the length of the tumour in the oesophagus of an arbitrary patient presented with oesophageal carcinoma. In mathematical notation, the probability that an arbitrary patient with oesophageal carcinoma has a tumour longer than 10 *cm* would be presented as:

$$\Pr(Length > 10).$$

□

However, only experts who are very much familiar with this notation will be able to completely understand it, especially when considering conditional probabilities.

Example 3.2 Again, consider the domain of oesophageal carcinoma. We now focus on the probabilities concerning the passage of food through the patient's oesophagus, which depends on the length of the carcinoma, its shape, and whether or not it is circular. In mathematical notation, the probability that an arbitrary patient with oesophageal carcinoma can swallow only liquid food, given that he has a polypoid, circular oesophageal

carcinoma of more than 10 *cm* would be presented as:

$$\Pr(Passage = \text{liquid} \mid Circ = \text{circular} \wedge Shape = \text{polypoid} \wedge Length > 10).$$

□

People unfamiliar with the notation of conditional probability can easily get confused about the meaning of what is represented on either side of the vertical bar.

Another way of describing the required probability to an expert is to use the frequency format [Gigerenzer & Hoffrage, 1995]. This format builds on the observation that representing occurrences of events is a fairly automatic cognitive process requiring little conscious effort. The basic idea is to describe probabilities in terms of frequencies, thereby converting abstract mathematics into simple manipulations on sets that are easy to recall and visualise.

Example 3.3 The probability presented in the example above using mathematical notation is described using the frequency format in the following way:

> Imagine 100 patients with a circular, polypoid oesophageal carcinoma of more than 10 *cm*. How many of these patients will be able to swallow only liquid food?

□

Gigerenzer *et al.* argue that cognitive biases are merely artifacts of the presentation format and that the frequency format serves to suppress biases such as base-rate neglect, overconfidence, and the conjunction fallacy [Gigerenzer & Hoffrage, 1995]. Overestimation of probabilities is reduced by assessing them as frequencies, because then people are more likely to be aware whether the sum of their assessments exceeds 100. The conjunction fallacy tends to disappear, because the frequency format appears to help people avoid choosing the most plausible description. For example, when asked "Out of 100 people like Linda, how many are bank tellers?" and "Out of 100 people like Linda, how many are bank tellers and active in the feminist movement?" (see also Subsection 2.1), most people correctly answered the latter with a smaller number.

Although the frequency format is easier for people to understand and apparently less liable to lead to mistakes, it is not always intuitively appealing. This is, for example, the case in domains where experts find it impossible to imagine 100 occurrences of a rare event. The domain of oesophageal carcinoma is such a problem domain. The probabilities used in the examples above all originate from a belief network built to support therapy selection for patients with oesophageal carcinoma [Van der Gaag *et al.*, 1999]. The probabilities for this network were assessed with the help of two domain experts from the Netherlands Cancer Institute. Since oesophageal carcinoma is a low incidence disease in the Netherlands, the experts consulted often found it impossible to imagine 100 patients having the same characteristics.

Although the frequency method cannot always be applied, the idea of transcribing probabilities in words can be exploited in various other ways [Van der Gaag *et al.*, 1999].

Example 3.4 The probability presented in the example above using the frequency format, is transcribed without frequencies in the following way:

> Consider a patient with a circular, polypoid oesophageal carcinoma of more than 10 *cm*. How likely is it that this patient will be able to swallow only liquid food?

□

The final presentation issue concerns the format in which experts are required to give their answer. This format not only depends on the choice of uncertainty representation, but also on the choice of elicitation method. As we will see in the next section, some methods will require a verbal response, whereas others require an expert to, for example, mark a scale.

4 Methods

With the term *probability elicitation method*, we denote any aid that is used to acquire a probability from an expert. Generally, a distinction is made between *direct* and *indirect* methods. With direct methods, experts are asked to directly express their degree of belief as a number, be it a probability, a frequency or an odds ratio. For expressing probabilities, however, people find words more appealing than numbers. This is probably because the vagueness of words captures the uncertainty they feel about their probability assessment; the use of numerical probabilities can produce considerable discomfort and resistance among those not used to it [Winterfeldt & Edwards, 1986]. Since, in addition, directly assessed numbers tend to be biased, various indirect elicitation methods have been developed. With these methods an expert is asked not for a direct assessment but for a decision from which his degree of belief is inferred; the use of an indirect method avoids having to explicitly mention probabilities for those who do not have clear intuitions about them [Morgan & Henrion, 1990]. For most methods, visual aids have been developed to make the elicitation easier on the experts.

In this section, we review the most commonly used methods for the elicitation of probabilities. These methods can be roughly divided into three categories:

- probability-scale methods;
- gamble-like methods;
- probability-wheel methods.

A probability-scale method is a direct method, where the expert is asked to indicate his degree of belief on a scale. The probability-wheel and gamble-like methods are indirect methods, since they require a decision instead of a number from the expert. We will devote a subsection to each of these categories and another subsection to some less known methods for probability elicitation we have encountered in the literature.

4.1 Probability scales

A well-known direct method of elicitation is the use of a numerical probability scale such as the one shown in Figure 1. A probability scale can be a horizontal or vertical line with several anchors. In Figure 1, we have anchors denoting 0%, 25%, 50%, 75%, and 100% probability.

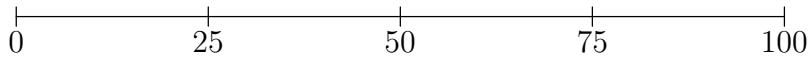

Figure 1: A numerical probability scale

For each probability that is to be assessed the expert is asked to mark the "correct" position on the scale. A separate scale is used for each probability. The indicated probability can be determined by measuring the distance between the mark and 0% on the scale. The expert should mark the scale in such a way that it is clear what position on the scale he is indicating, for example by using a small line or a carefully centred '×', instead of circling the scale. The basic idea of the scale is to support experts in their assessment task by allowing them to think in terms of visual proportions rather than in precise numbers.

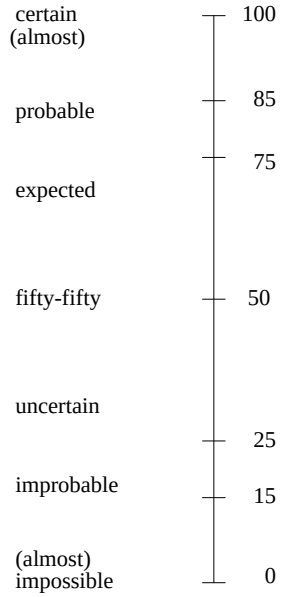

Figure 2: A probability scale with both verbal and numerical anchors

In addition to this horizontal probability scale, there exist variants with a different number of anchors and also vertical variants; in addition, there exists a scale with both verbal and numerical anchors [Renooij & Witteman, 1999]. This scale is depicted in Figure 2, with on the right-hand side seven numerical anchors representing 0%, 15%, 25%, 50%, 75%, 85% and 100% probability, and on the left-hand side the seven verbal anchors impossible, improbable, uncertain, fifty-fifty, expected, probable, and certain. The scale was developed

to be part of an elicitation method tailored to the fast elicitation of a large number of probabilities for a belief network [Van der Gaag *et al.*, 1999]. For each probability to be assessed, the expert is presented with a figure containing a description of the required probability and the scale; figures are grouped in such a way that those pertaining to probabilities that should sum to 100% are on the same sheet, or on consecutive sheets. Depending on how familiar experts are with the probability to be assessed, they can use either the verbal or the numerical expressions, or both.

Advantages of using a probability scale are that it is easy to understand and use and provides a fast method of elicitation, thereby allowing for elicitation of large numbers of probabilities. However, assessments made using a probability scale tend to be inaccurate and prone to scaling biases such as centering and spacing effects [Winterfeldt & Edwards, 1986]. The *centering effect* describes the tendency of people to use the middle of the probability scale; if people aesthetically divide their responses over the scale, this is termed *the spacing effect*. Note that the spacing effect cannot occur if a different scale is used for each separate assessment. Also note that the probability scales discussed are linear scales and therefore do not allow for elicitation of very large or very small probabilities. The use of a logarithmic scale would solve this problem. It should be kept in mind, however, that experts' subjective scales are naturally equal-interval linear scales, not logarithmic scales [Wright & Bolger, 1992].

4.2 Gamble-like methods

When people find it hard to express their degree of belief about some event as a number, their judgemental probability can be inferred from their behaviour in a controlled situation [Baecher, 1998]. Indirect methods of probability elicitation such as, for example, the gamble-like methods are designed to represent such a controlled situation. The gamble-like methods for eliciting probabilities originate from the *Standard Gamble* introduced by Von Neumann and Morgenstern [Von Neumann & Morgenstern, 1953] as an indirect method for utility elicitation. The basic idea behind a gamble-like method is that the expert is presented with a choice between two lotteries. For one of the lotteries, the probability of winning corresponds to the probability of the event to be assessed; the probability of winning in the other lottery is set by the elicitor. The latter probability, or the associated price, is varied until the expert is indifferent about the two lotteries, whereupon the probability of the event to be assessed can be determined.

With a gamble-like method an expert is not required to give a probability assessment, but may instead compare a complicated concept with an event that does have meaning such as winning a lottery or a bet. We can distinguish two types of gamble. In the *certain-equivalent* gamble, a *sure thing*, that is, a 100% chance of winning, is compared to a lottery; the *lottery-equivalent* gamble consists of comparing two lotteries. From the choices made by the expert, the subjective probability for the associated event is inferred. Gamble-like methods can be presented to the expert graphically with the help of a decision tree depicting the possible alternatives, probabilities, and outcomes. The concept of decision trees, along with the symbols used, will have to be explained to the expert. When the

expert fully understands the drawings, the elicitation process can proceed.

We will give an example of both variants of the gamble-like method. For each example we will explain what choices the expert has and how to determine the desired probability from his answers. In the first example we will also briefly explain the decision tree.

Example 4.1 Again consider the domain of oesophageal carcinoma. We focus on the probabilities to be elicited from the domain expert concerning the length of the tumour in the oesophagus of an arbitrary patient presented with oesophageal carcinoma. For ease of exposition we take the variable *Length* to be a binary variable with values ≤ 6 *cm* and > 6 *cm*. The probabilities required are the probability of a patient having a tumour with a length of 6 *cm* or less, and the complementary probability of the patient having a tumour longer than 6 *cm*.

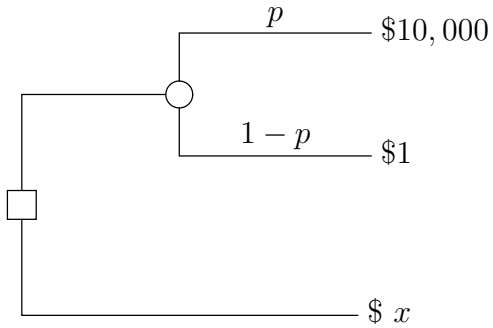

Figure 3: A certain-equivalent gamble

We will first consider a gamble with a certain equivalent, as depicted in Figure 3. Here the domain expert is presented with the following choice, indicated by a box (the decision node) in the figure:

- either enter a lottery where the pay-off ($\$10,000$, resp. $\$1$) depends on the "true" probability p of an arbitrary patient having a tumour of more than 6 *cm*,
- *or* accept a certain amount of money x set by the elicitor, instead.

The circle in the above figure indicates an uncertain event: with probability p the expert will earn $\$10,000$, and with a probability of $1 - p$ only $\$1$. The idea is that the elicitor varies the amount of money x in the certain equivalent until, for some value x' the expert is indifferent about the two alternative choices. In that case it is assumed that the expected value for both alternatives is the same. We can then compute the probability p that the patient has a tumour of more than 6 *cm* from

$$x' = 10,000 \cdot p + 1 \cdot (1 - p)$$

□

A major drawback of this version of the gamble-like method is that elicited probabilities tend to be highly influenced by the *risk-attitude* of the expert. Some people are risk-seeking in the sense that they tend to choose a less probable alternative if it has a potentially more favourable outcome; other people tend to be risk-averse and will, for example, be more inclined to go for the certain outcome. Always going for the "sure thing" is known as the *certainty effect* [Law, Pathak & McCord, 1998].

A gamble with a lottery equivalent is less influenced by risk-attitudes. With this version of the gamble-like method, the expert is asked to choose between two lotteries; the price received upon winning (or losing) is equivalent for both lotteries.

Example 4.2 Consider again the example from the domain of oesophageal carcinoma, dealing with the elicitation of probabilities concerning the length of the tumour in an arbitrary patient with oesophageal carcinoma.

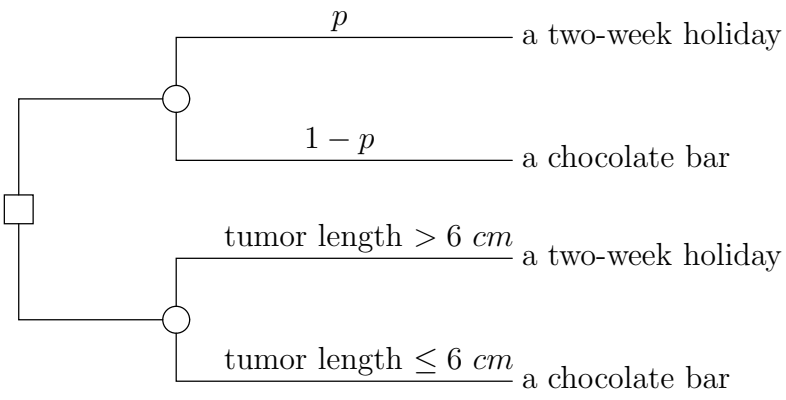

Figure 4: A lottery-equivalent gamble

When presented with a lottery equivalent gamble, the domain expert has the following choice:

- either enter the lottery where the outcome depends on some probability p set by the elicitor,
- *or* enter the lottery where the outcome depends on the probability of an arbitrary patient having a tumour of more than 6 cm.

In this lottery-equivalent gamble the probability p is varied until the expert is indifferent about the two alternatives. Again assuming that in that case the expected value of both alternatives is the same, we compute the probability p that the patient has a tumour of more than 6 cm from

$$\begin{aligned} & p \cdot value(\text{a two} - \text{week holiday}) + (1-p) \cdot value(\text{a chocolate bar}) \\ & = \\ & \Pr(\text{tumour length} > 6\ cm) \cdot value(\text{a two} - \text{week holiday}) + \\ & \Pr(\text{tumour length} \leq 6\ cm) \cdot value(\text{a chocolate bar}) \end{aligned}$$

where *value* is a subjective measure of how valuable the outcome is to the expert. When the expert is indifferent, p directly represents the probability of a patient having a tumour longer than 6 *cm*, that is, $p = \Pr(\text{tumour length} > 6\ cm)$. □

An advantage of this latter gamble over the former is that it directly presents the probability of interest and is less bothered by risk-attitudes. In addition, rewards can be expressed in terms other than money. As the gamble-like method does not require an expert to provide a probability assessment, it is considered to suppress some of the cognitive biases described in Subsection 2.1. However, the certain-equivalent gamble is easily influenced by risk-attitudes, which causes the probability derived from this method to be unequal to the expert's subjective probability, thus introducing a bias.

Gamble-like methods are not very expert-friendly methods. The methods are complicated to learn. Also, experts may feel confronted with lotteries that are hard to conceive because of the rare and unethical situations they represent, like, for example, winning a two-week holiday if a patient dies [Van der Gaag *et al.*, 1999]. Another drawback is that these methods are very time-consuming; they tend to take a lot of time per probability which makes them less suitable for assessing the thousands of probabilities required for belief networks.

Studies that have used the discussed elicitation methods for utility elicitation, report the consistent finding that numbers elicited with a probability scale are significantly lower than those elicited with the Standard Gamble [Stavem, 1998, Rutten-van Molken *et al.*, 1995, Ubel *et al.*, 1996]. Also, values obtained with the certain-equivalent gamble are consistently lower than for the lottery-equivalent gamble [Law, Pathak & McCord, 1998]. We are unaware of similar studies using the elicitation methods for probability elicitation.

4.3 Probability-wheel methods

An indirect method that is not influenced by risk-attitudes is the probability-wheel method. A probability-wheel is a wheel-of-fortune-like wheel with two differently coloured sections. The sizes of these sections are adjustable and there is a pointer attached to the center of the wheel. An example of a probability-wheel is shown in Figure 5.

Example 4.3 Using the same example as before, the expert is now asked which of the following events he considers most likely:

- the length of the tumour of an arbitrary patient with oesophageal carcinoma is more than 6 *cm*,
- *or*, after spinning the pointer, it will land in the red section.

The size of the red section of the probability wheel is adjusted by the elicitor until the expert considers the two events to have equal probability. The probability of an arbitrary patient having a tumour longer than 6 *cm* now equals the proportion of the probability wheel that is coloured red. □

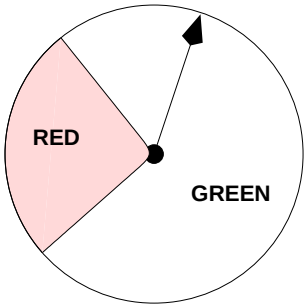

Figure 5: A probability wheel

The probability-wheel method has several drawbacks. The method tends to be very time-consuming and even infeasible when hundreds or thousands of probabilities are needed, as for belief networks. Also, the method is quite close to direct estimation as the expert may recognise that the judgements he is asked to make are disguised assessments of the proportion of red showing on the wheel [Winterfeldt & Edwards, 1986]; the advantage of suppressing judgemental biases, therefore, may disappear. The method is not suitable for assessing very large or very small probabilities, for it will be difficult for an expert to distinguish between a very small red section and an even smaller red section. The advantage of probability wheels could be that they help experts visualise probabilities, but definitive conclusions from research on this are lacking [Baecher, 1998].

4.4 Other Methods

In this subsection we will briefly describe two other, very different and less-known methods for probability elicitation encountered in the literature. With the first method, experts are allowed to express their knowledge about uncertainties in any form they prefer and not necessarily in numbers. The second method requires experts to make pair-wise comparisons between events.

Druzdzel and Van der Gaag [Druzdzel & Van der Gaag, 1995] have presented a method for probability elicitation where experts are allowed to provide both qualitative and quantitative information, whichever they are most comfortable with. The assumption underlying this method is that in the hyperspace of all possible probability distributions over the set of variables under consideration, one of these distributions is the "true" one. The information provided by the experts can be looked upon as a set of constraints used to diminish the hyperspace of possible distributions. These constraints are put in a canonical form resulting in a system of (in)equalities with constituent probabilities as unknowns. From the inequalities an upper and lower bound can be computed for any probability of interest. For the interval between upper and lower bound a second-order distribution is computed to determine the point within the interval that is most likely to be the actual probability. This second-order distribution is found by sampling from the distribution hyperspace and checking for each selected distribution whether it is a solution for the system

of (in)equalities.

Another method, originally designed for utility elicitation, is the *analytical hierarchy process* [Saati, 1980]. With this method an expert is presented with all possible combinations of pairs of events for which utilities are to be assessed. When the method is used for probability elicitation, the expert is asked to compare, for each pair, the two events and to indicate the relative likelihood of events A and B using the scores shown in Table 1. This

score	*relative likelihood*
1	A and B are equally likely
2	undecided between 1 and 3
3	A is weakly more likely than B
4	undecided between 3 and 5
5	A is strongly more likely than B
6	undecided between 5 and 7
7	A is very strongly more likely than B
8	undecided between 8 and 9
9	A is absolutely more likely than B

Table 1: The scale for pair-wise comparisons

method has the advantage that experts are not required to explicitly state probabilities. Another advantage is that consistency of the expert's statements can be easily checked, for the result from the comparisons should be a transitive ordering of events. However, using this method for probability elicitation for belief networks poses two problems:

- The number of comparisons to be made exceeds, by far, the number of probabilities to be assessed. For example, the assessment of a mere 100 probabilities would require an expert to make $\binom{100}{2} = 4950$ pairwise comparisons of events.
- A lot of the events will differ so much that they are hard to compare for an expert.

Besides the problem of the great number of comparisons to be made, rather uninsightful statistical methods are required to compute the probabilities from the results of the comparisons.

5 Discussion

We have discussed various issues that are to be taken into consideration when faced with the task of probability elicitation. We have seen that probability judgements are prone to bias and that several elicitation methods have been developed to aid an expert in assessing probabilities, thereby suppressing, to some extent, these biases. It is clear that an important motivation for choosing a particular probability elicitation method is the ease with which both elicitor and expert understand and use the method. Moreover, the

time an expert has available can limit the choice of methods. There will often be a trade-off between available time and the precision required, since the methods that are said to provide the most precise results are also the most time-consuming. Some people doubt however, that this trade-off really exists [Kadane & Winkler, 1988]: the use of gambles might not result in assessments that are as good as is believed, and faster methods such as the probability-scale methods might produce results that are better than believed.

While some of the phenomena reported in the heuristics and biases literature are real, reliable and reproducible, they may not be relevant, that is, they may not apply to the situation in which thousands of probabilities need to be assessed for a belief network. For example, some biases, such as the conjunction fallacy, cannot arise during elicitation of probabilities for belief networks [Anderson, 1998]. Edwards [Edwards, 1995] gives another three arguments why some of the phenomena may be irrelevant to probability elicitation for belief networks. The first is domain expertise: for the elicitation of probabilities experts are used, who presumably know all there is to know about the subject matter of the probabilities being judged. The studies concluding that humans are typically overconfident when providing probability estimates, arrive at that conclusion using general knowledge (almanac) questions and student subjects that are often not trained in estimating probabilities. It is not at all clear that these results can be generalised to experts making assessments pertaining to their expert knowledge. Weather forecasters, for example, turn out to be very well calibrated indeed [Edwards, 1990]. Another argument is probability judgement expertise: judging probabilities is something that can be learned. An expert who has done some training in estimating probabilities will find it easier to translate his knowledge and experience into probability judgements. The third reason is the possibility of consistency checks such as the sum checks and antecedent checks discussed in Subsection 2.6. These checks can be used during elicitation to provide the expert with information based on which he can, if necessary, reconsider his judgements.

When probability elicitation is seen as part of a stepwise refinement procedure, fast elicitation methods can be used to get initial rough estimates of the desired probabilities; sensitivity analysis methods [Coupé, Van der Gaag & Habbema, 2000] are then used to determine to which variables in the network the outcome is very sensitive. The focus of precise probability elicitation can then be put on the most sensitive parts of the network. Another important issue to keep in mind is that the networks are used to support a decision maker. They should at least improve the situation in which they are to be used, which means they do not always have to be 100% correct [Edwards, 1995].

We are unaware of any systematic experimental evaluation studies of the different elicitation methods, especially in view of belief networks; the results of the considerable number of empirical comparisons of methods do not show great consistency [Morgan & Henrion, 1990]. It is clear that a lot of research necessary to be able to decide on the best elicitation method, still has to be done. What is lacking are large multi-method studies where experts are asked to assess a large number of probabilities with every single method. It is important to get ecologically valid results, that is, results based on behaviour that is relevant to a real-world situation. Such results can provide for insight into when to use which method and what methods not to use at all.

Acknowledgments

I would very much like to thank Linda van der Gaag and Cilia Witteman for their extensive and useful comments on earlier drafts of this paper.

References

[Anderson, 1998] J.L. Anderson (1998). Embracing uncertainty: The interface of Bayesian statistics and cognitive psychology. *Conservation Ecology[online]*, vol. 2(1):2. `http://www.consecol.org/vol2/iss1/art2`

[Baecher, 1998] G.B. Baecher (1998). Judgemental probability in geotechnical risk assessment. Technical report, prepared for The Office of the Chief, U.S. Army Corps of Engineers. `http://www.ence.umd.edu/~gbaecher/papers.d/judge_prob.d/judge_prob.html`

[Clemen & Winkler, 1999] R.T. Clemen, R.L. Winkler (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, vol. 19, pp. 187 – 204.

[Cooke, 1991] R.M. Cooke (1991). *Experts in Uncertainty. Opinion and Subjective Probability in Science*, Oxford University Press.

[Coupé, Van der Gaag & Habbema, 2000] V.M.H. Coupé, L.C. van der Gaag, and J.D.F. Habbema (2000). Sensitivity analysis: An aid for belief-network quantification. *Knowledge Engineering Review*, vol. 15, pp. 1 – 18.

[Druzdzel & Van der Gaag, 1995] M.J. Druzdzel and L.C. van der Gaag (1995). Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In Ph. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 141 – 148. Morgan Kaufmann Publishers, San Francisco.

[Edwards, 1990] W. Edwards (1990). Unfinished tasks: A research agenda for behavioral decision theory. In R.M. Hogarth, editor, *Insights in Decision Making*, pp. 44 – 65. The University of Chicago Press, Chicago.

[Edwards, 1995] W. Edwards (1995). How to estimate thousands of reliable, valid probabilities. In M.J. Druzdzel, L.C. van der Gaag, M. Henrion, and F.V. Jensen, editors, *Working Notes of the IJCAI Workshop on Building Probabilistic Networks: Where Do the Numbers Come From ?* AAAI Press. `http://www.sis.pitt.edu/~dsl/hailfinder/probms2.html`

[Edwards, 1998] W. Edwards (1998). Hailfinder. Tools for and experiences with Bayesian normative modeling. *American Psychologist*, vol. 53 (4), pp. 416 – 428.

[Evans & Over, 1996] J.St.B.T. Evans, D.E. Over (1996). *Rationality and Reasoning*. Psychology Press, Hove, United Kingdom.

[Fenton, 1998] N. Fenton (1998). Probability elicitation and bias. `http://www.csr.city.ac.uk/people/norman.fenton/bbns/frconten.html#ProbabilityElicitation`

[Gigerenzer & Hoffrage, 1995] G. Gigerenzer and U. Hoffrage (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, vol. 102, pp. 684 – 704.

[Jensen, 1995] A.L. Jensen (1995). Quantification experience of a DSS for mildew management in winter wheat. In M.J. Druzdzel, L.C. van der Gaag, M. Henrion, and F.V. Jensen, editors, *Working Notes of the IJCAI Workshop on Building Probabilistic Networks: Where Do the Numbers Come From ?*, pp. 23 – 31. AAAI Press.

[Kadane & Winkler, 1988] J.B. Kadane, R.L. Winkler (1988). Separating probability elicitation from utilities. *Journal of the American Statistical Association*, vol. 83, pp. 357 – 363.

[Kahneman, Slovic & Tversky, 1982] D. Kahneman, P. Slovic, A. Tversky (Editors) (1982) *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.

[Laskey & Mahoney, 1997] K.B. Laskey and S.M. Mahoney (1995). Network fragments: Representing knowledge for constructing probabilistic models. In D. Geiger and P. Shenoy, editors, *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pp. 334 – 341. Morgan Kaufmann Publishers, San Francisco.

[Law, Pathak & McCord, 1998] A.V. Law, D.S. Pathak, M.R. McCord (1998). Health status utility assessment by standard gamble: a comparison of the probability equivalence and the lottery equivalence approaches. *Pharmaceutical Research*, vol. 15(1), pp. 105 – 109.

[Mahoney & Laskey, 1996] S.M. Mahoney and K.B. Laskey (1996). Representing and combining partially specified CPTs. In K.B. Laskey and H. Prade, editors, *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp. 391– 400. Morgan Kaufmann Publishers, San Francisco.

[Merkhofer, 1987] M.W. Merkhofer (1987). Quantifying judgmental uncertainty: Methodology, experiences, and insights. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 17 (5), pp. 741 – 752.

[Morgan & Henrion, 1990] M.G. Morgan, M. Henrion (1990). *Uncertainty, a Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge.

[Pearl, 1988] J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Palo Alto.

[Renooij & Witteman, 1999] S. Renooij and C.L.M. Witteman (1999). Talking probabilities: Communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, vol. 22, pp. 169 – 194. Elsevier Science Publishers.

[Rutten-van Molken *et al.*, 1995] M.P. Rutten-van Molken, C.H. Bakker, E.K. van Doorslaer, S. van der Linden (1995). Methodological issues of patient utility measurement. Experience from two clinical trials. *Medical Care*, vol. 33(9), pp. 922– 937.

[Saati, 1980] T.L. Saati (1980). *The Analytic Hierarchy Process.* McGraw-Hill, Inc.

[Skinner, 1999] D.C. Skinner (1999). *Introduction to Decision Analysis.* Probabilistic Publishing, Gainesville, Florida.

[Stavem, 1998] K. Stavem (1998). Quality of life in epilepsy: Comparison of four preference measures. *Epilepsy Research*, vol. 29, pp. 201 – 209. Elsevier Science.

[Ubel *et al.*, 1996] P.A. Ubel, G. Loewenstein, D. Scanlon, M. Kamlet (1996). Individual utilities are inconsistent with rationing choices: A partial explanation of why Oregon's cost-effectiveness list failed. *Medical Decision Making*, vol. 16(2), pp. 108 – 116.

[Van der Gaag *et al.*, 1999] L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman and B.G. Taal (1999). How to elicit many probabilities. In K.B. Laskey and H. Prade, editors, *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp. 647 – 654. Morgan Kaufmann Publishers, San Francisco.

[Van Lenthe, 1993] J. van Lenthe (1993). ELI: An interactive elicitation technique for subjective probability distributions. *Organizational Behavior and Human Decision Processes*, vol. 55(3), pp. 379 – 413.

[Von Neumann & Morgenstern, 1953] J. Von Neumann, D. Morgenstern (1953). *The Theory of Games and Economic Behavior.* Wiley, New York, third edition.

[Winterfeldt & Edwards, 1986] D. von Winterfeldt, W. Edwards (1986). *Decision Analysis and Behavioral Research*, Cambridge University Press, Cambridge.

[Wright & Bolger, 1992] G. Wright, F. Bolger (1992). *Expertise and Decision Support*, Plenum Press, New York.