Talking probabilities: communicating probabilistic information with words and numbers

Silja Renooij

Dept. of Computer Science, Utrecht University
P.O. Box 80.089, 3508 TB Utrecht
The Netherlands
e-mail Silja.Renooij@cs.uu.nl

Cilia Witteman

Psychological Laboratory, Utrecht University
P.O. Box 80.140, 3508 TC Utrecht
The Netherlands
e-mail C.Witteman@fss.uu.nl

Abstract

The number of knowledge-based systems that build on Bayesian belief networks is increasing. The construction of such a network however requires a large number of probabilities in numerical form. This is often considered a major obstacle, one of the reasons being that experts are reluctant to provide numerical probabilities. The use of verbal probability expressions as an additional method of eliciting probabilistic information may to some extent remove this obstacle. In this paper, we review studies that address the communication of probabilities in words and/or numbers. We then describe our own experiments concerning the development of a probability scale that contains words as well as numbers. This scale appears to be an aid for researchers and domain experts during the elicitation phase of building a belief network and might help users understand the output of the network.

Key words: Communicating probability; Expert systems; Knowledge elicitation; Explanation

1 Introduction

Bayesian belief networks, also referred to as probabilistic networks, were first introduced in the late 1980s by Pearl [39]. Since then, an increasing num-

ber of successful applications of such networks in different problem domains have been developed, which demonstrates that they have established their position in Artificial Intelligence as valuable representations of reasoning with uncertainty [1,25,28,12,31] . A belief network consists of a qualitative and a quantitative part. The qualitative part is a directed graph, where the nodes represent the domain's variables (in a medical diagnostic application the variables could for example be a laryngitis and its symptoms such as a sore throat and fever) and the arcs their dependencies. The quantitative part encodes probabilities over these variables, such as the probability of some symptom given a diagnosis of laryngitis.

Constructing the qualitative part of a belief network, although elaborate, seems relatively straightforward and experts feel comfortable doing so. The quantitative part, with the probabilities over the variables, is more problematic. On the input side, it involves the elicitation from domain experts of (conditional) probabilities for all variables. This may be a prohibitive quantity of probabilities, even in restricted domains [19]. What's more, experts are required to express all these probabilities numerically, something they are often reluctant to do. They do not feel familiar enough with the concept of probability or they find it difficult to attach a number to their beliefs [23]. On the output side, explanations of the results of computations of the network in terms of variables with numerical probabilities may be uncomfortable. Researchers have recognised the importance of providing users with more easily understandable explanations of the results, for which numbers may not necessarily be the best option and verbal probability expressions are seen as good alternatives [16,13].

Except in situations where the odds are objectively measurable, most people feel more at ease with verbal probability expressions than with numbers. When people communicate probabilities, they frequently do so in words rather than in numbers. In the development of a computer system, viz. a belief network, that is intended to reason with probabilities and to communicate the results of that reasoning to users, the mode in which people normally represent probability needs to be taken into account (compare [29,52]).

However, it is not obvious which verbal probability expressions could then be used. We therefore investigate whether we can find a set of verbal probability expressionswhose meaning is agreed upon and which together cover the whole range from zero to a hundred percent probability. A combination of this scale with a numerical scale could then be an aid, both during the elicitation phase of building a belief network and in understanding the output of the network's computations. We are not the first to study this problem. Bonissone and Decker [3] analyzed the use of linguistic terms to establish a certain granularity to facilitate knowledge elicitation in possibilistic reasoning. In the next paragraph, we review other researchers' empirical studies on the

use of probability expressions. Results in favour of numbers as well as results in favour of words will be reported. Because no unequivocal conclusion can be drawn from these studies, we undertook some additional experiments. In the third paragraph we present this series of four experiments.

2 Previous studies

Numbers have a persuasive advantage over words in the sense that they are precise, they allow calculations and they have a fixed rank-order. A hundred percent is always more than seventy-five percent. Words are, in comparison, vaguer, they do not allow calculations and they are more variably interpretable [54]. This disadvantage of words is visible in the quite consistent finding that the interpretation of verbally expressed probabilities is context-dependent [5,56]. If winning a lottery is "possible", entering the lottery may generally be seen as a good thing to do, while if encountering a much disliked person at a party is "possible", going to that party is generally not judged to be wise. Moreover, personal opinions about the consequences of the events referred to result in individual variations in the meanings assigned to probability expressions [34]. Some people may not mind meeting a disliked person or even enjoy the confrontation, others may definitely wish to avoid it.

Studies of numerical versus verbal probabilities generally ask subjects to translate numerical expressions into words and vice versa. In such studies, researchers have found a great between subject variability in the numerical values assigned to verbal probabilities and great overlap between the words (see among others [2,53]). Much less within and between subject variability was found in the numerical probability expressions subjects used when asked to describe a graphical representation of chance than in the verbal expressions they used [8]. Subjects were found to be consistent with themselves in their interpretations of verbal expressions, but much less so with others [7].

Physicians are no exceptions. When they were asked to give the meanings of verbal probability expressions by marking them on a 0 – 100 percentage scale [15] or when they were asked to translate verbal expressions into numerical expressions ([6], see also [43]) physicians regularly gave different interpretations. When probability information was communicated by verbal expressions, interpretations were also found to be highly imprecise, presumably because they were influenced by the severity of the consequences associated with the information [32]. For example, "low probability of infection" was interpreted differently than "low risk of death". Most of the authors referred to above conclude that physicians should use numerical, not verbal, expressions of probability (see also [35]). Verbal probability expressions may lead to confusion, therefore numbers should be used [7].

However, ceteris paribus, verbal expressions of probability are perceived as more natural than numerical probabilities, easier to understand and communicate and better suited to convey the vagueness of one's opinions [54]. But this may be annotated. Erev and Cohen [17] detected an interesting phenomenon, which they dubbed the 'communication mode preference paradox' (CMP). The subjects in their study preferred to receive precise, i.e. numerical, information, but they knew that their own opinions were imprecise and therefore preferred to express them in, vaguer, verbal terms. Other researchers found that one person in three prefers numbers for both expressing and receiving information, the second prefers words for both, and the third indeed betrays Erev & Cohen's communication mode preference paradox [55]. But this preference was not firm, neither for Erev and Cohen's subjects nor for the subjects in a study by Rapoport and colleagues [41]. Subjects were found to be willing and able to use both modes of expression.

Brun and Teigen [5] found that physicians preferred to use words in communicating probabilities to their patients. Other researchers report similar results. Physicians and other health workers express and process probabilities in verbal rather than numerical form [32,30]. Physicians rarely reason using numerical probabilities, and if they do, they tend to make errors [14,49]. Kuipers and colleagues [30] conclude that to physicians, subjective probabilities are not numbers.

Brun and Teigen pursued their inquiry and discovered what could be called a variation of the communication mode preference paradox, with physicians and their patients. While physicians preferred to use words and thought their patients would understand words more easily, the patients, on the contrary, preferred to receive information in numbers and reported that numbers were easier to understand. But the patients did not understand the numbers in accordance with the physicians' intention. For example, a physician would state a 35% probability of having a disease and thereby intend to communicate a moderate probability. Some patients might then understand that they had a very real probability of indeed having the disease and be more alarmed than the physician meant them to be, while others would understand it as less than fifty percent chance and overestimate their well-being (compare also [10]). Numbers may, in short, wrongly suggest a precision of opinion [7]. Brun and Teigen [5] conclude that numbers should not self-evidently be preferred to words in a medical context. O'Brien [37] is of the same opinion. He concludes that the two modes of communicating probability, numerical and verbal, can both be used. The argument that verbal expressions are too vague in meaning to be used in medicine is counter-balanced by indications from survey results that numbers have very little meaning for the average member of the public.

Another argument that words would be at least equally suitable to express probabilities as numbers is that it appears that whether people receive infor-

mation in verbal or in numerical form, does not influence either the subsequent thought processes or the actions based on the information. The overall quality of decisions in the two communication modes as well the judgement processes were found to be similar [54]. The two did differ, in that with numbers, judges used the 50% category much more often than the tossup category was used in the verbal mode; and overconfidence was systematically greater given verbal than numerical probability expressions. However, neither the numerical nor the verbal mode was found to result in uniformly better processing than the other. The conclusion seems to be that there are no grounds to prefer either numerical or verbal probability expressions as the better medium (cf. also [8,17,21,46,38]).

It has also been suggested that in some situations clearly interpretable verbal expressions are preferable to numbers. For example, Wallsten et al. [55] advise probability judgements to be elicited in verbal form whenever possible, except when there is a large amount of specific data at hand and numerical probabilities may be justified, because the use of verbal expressions seems to be more comfortable to people than the use of numbers. For medical diagnostic networks it has been suggested that, even if verbal expressions would be less precise than numerical expressions, imprecision of probabilities does not entail a deterioration in the average performance of a belief network [23,40,36]. This suggestion awaits further corroboration.

Limiting the number of verbal expressions in a scale might, however, be advisable. The use of some verbal expressions should be avoided, to wit those for which the variation in interpretation is found to be high. This is especially the case with expressions in the central range, such as possible or likely. Expressions for the extremes of the range, that is: impossible and certain, are much less variably interpreted and could be used [26]. A small number of carefully selected expressions seems best [7,42], or a table could be presented codifying the numerical meaning of the verbal phrases (compare among others [22,18]). Physicians could then continue to use verbal expressions if they prefer, but with more consistency of terminology [32].

A scale with a small number of expressions is recommended because it would be easier on people's cognitive capacities. It is easier to distinguish some seven information categories or expressions clearly than it is to demarcate the meanings of a long list of expressions [33,50]. Other studies have used lists of 18 expressions [8] or 19 [7,22] or 34 in a long list and 14 in a shorter version [5], as few as two [21] or as many as 52 [34]. These lists were compiled by the researchers, which does not guarantee that people would actually use them. Indeed, Zimmer [57] found that when subjects were asked directly for verbal descriptions of probability, the mean number of expressions used was 5.44.

Differences in interpretation of the verbal expressions may be prevented when

the expressions are offered with a pre-defined rank-order, in a scale. Indeed, Hamm [22] found that subjects were less variable in assigning (numerical) meanings to expressions in an ordered list than to expressions in a random list and that ordered lists produced more accurate responses than random lists. A pre-defined rank-order would not be artificial, because an encouraging between subject consistency was found in the rank ordering of verbal probability expressions by general practitioners [37], and individuals were found to have a relatively stable rank ordering of verbal probability phrases over time [7,27].

Our suggestion is that when probabilities are elicited from experts, for example when constructing a belief network, the experts be shown a scale, depicted graphically as a vertical line with numbers on the one side and words on the other. When experts are more comfortable with numbers, they may refer to the number side of the scale and when they prefer to express their opinions in words, they may refer to the verbal expressions. This same double scale might then be offered as reference when the output of the belief network is given.

Summing up the above (compare [24]), there is sufficient justification for an attempt to construct a scale which includes both numerical and verbal expressions of probability. Merz et al. [32] proposed a six-expression verbal scale to be used as a standard by physicians, which covers the whole probability range: extremely high probability, very high, high, low, very low and extremely low probability. Although this scale does have the elegant properties of being very simple and symmetrical, the use of qualifiers such as 'very' has been found to introduce additional vagueness [45] and we will therefore try to avoid it.

3 Our study

We undertook four successive studies to develop a scale of verbal probability expressions, usable in combination with a numerical probability scale. We took into account the possibility that the context in which these expressions are elicited and presented influences their interpretation. And because decision support systems such as belief networks are often used in the domain of medical diagnosis, we included medical subjects. In study 1 we asked subjects which verbal probability expressions they commonly use. In study 2 we asked (other) subjects to rank order the expressions from study 1. In study 3 we asked subjects to make pairwise comparisons between each pair of expressions, to determine how the words should be projected on a numerical probability scale. Studies 2 and 3 are unlike the experiments most other researchers have done. We never asked subjects to directly translate words into numbers or vice versa. In study 4 we tested whether decisions were influenced by the mode, verbal or numerical, in which probability information was pre-

sented, using the scale that had resulted from the previous three studies for the verbal expressions.

FIRST EXPERIMENT

In this first study we aimed at a list of comonly used probability expressions. Most researchers use a dictionary or published articles to draw up a list of probability expressions. Since we had no a priori reason to assume that such sources contain only the expressions actually used and think they are more likely to list all linguistic possibilities, we designed a short questionnaire with which we approached the subjects.

Procedure

The participants received a short questionnaire. In the first paragraph they were asked for their co-operation by generating a list of commonly used verbal terms expressing (im)probability. Examples were given, such as "it is unlikely that I will pass my exam" or "I will probably go to Amsterdam this weekend". Instructions were given, in the second paragraph, to write down a list of terms judged suitable in situations where one wishes to express a degree of (im)probability, for example about the chance of rain tomorrow. Subjects were reminded to only list expressions they thought were common, and to try them out for themselves in different virtual situations. At the end of the page they were asked for their gender, year of birth and profession or study.

Subjects

There were 53 participants, 47 students (computer science, psychology and artificial intelligence), and 6 faculty members, 23 female and 30 male, whose age ranged between 18 and 54 with an average of 23 (SD = 8.7).

Results

The 53 participants together generated 144 different expressions. They wrote down a mean of 8.2 expressions per participant (SD = 4.1). Of these 144 expressions, 108 (75%) were composed of a probability term plus a modifier such as 'very' or 'reasonably'. Some modifiers seemed synonymous, but we counted the phrases containing such modifiers, e.g. 'almost possible' and 'nearly possible', as different phrases. Ninety-five expressions (66%) were used by only one participant and another seventeen (11%) by only two participants. Table 1 lists the seven expressions that were used by fifteen or more, or almost thirty percent, of the participants[1] The next often used term was written down by

[1] Translations of the original Dutch phrases 'mogelijk, waarschijnlijk, onwaarschijnlijk, zeker, onzeker, te verwachten, onmogelijk'.

eleven subjects, a couple of expressions were given by nine and eight subjects respectively, the rest had a frequency near one. Looking at the composite expressions as a check of common use, the difference in frequency between our list of seven and the rest was even greater.

Table 1
Verbal probability expressions generated by participants in the first experiment ($n = 53$), with their frequencies

expressions	frequency
possible	38
probable	30
improbable	28
certain	25
uncertain	21
expected	18
impossible	15

Discussion

Most other researchers use lists of expressions they have compiled themselves, by scanning literature or borrowing from others. Clark [9] proposed the method we used, to ask subjects to generate lists of commonly used expressions. He had fewer subjects (20), who generated more expressions each, with a mean of 12.9, than our 53 subjects, with a mean of 8.2. His most frequently used expressions were 'certain', 'possible', 'likely', 'definite', 'probable', 'unlikely' and 'impossible'. He thus also found seven expressions, quite comparable to ours. He also found more expressions on the positive side; that is: from fifty-fifty towards certain, than on the negative side of the range, that is: from fifty-fifty towards impossible.

In their attempt at codification of probability expressions, Mosteller and Youtz [34] advised 'impossible' and 'certain' for the two extremes, and 'even chance' for the mid-point. To cover the rest of the range, they advised 'probable' with modifiers. However, we think expressions with modifiers may give more rise to ambiguity than one-word expressions, therefore we decided for our list of 'possible', 'probable' and 'certain' plus their negations, plus 'expected' because that was used relatively often by our subjects.

We used our list of seven frequently generated expressions for the next experiments, adding one term, 'undecided' [2], to express fifty-fifty probability. This list of eight expressions neatly kept us within Miller's range of seven plus or

[2] In Dutch 'onbeslist'.

minus 2. We expected to resolve the asymmetry between the number of positive versus negative expressions in our next, ranking and scaling experiments.

SECOND EXPERIMENT

The second study was set up to determine if a single, stable rank order existed for the eight probability expressions from the first study. In this study we only looked for a rank order. Distances between the expressions are established in the third experiment.

Design

Subjects were asked to rank order the eight probability expressions. We had a context and a no context condition. In the no context condition the expressions were offered in isolation. In the context condition the probability expressions were embedded in a (Dutch) sentence describing a medical situation (for example: It is certain that young people do not get varicose veins). In both conditions we had medical students and other (social sciences) students (see Table 2).

Table 2
Design of the ranking experiment, with numbers of subjects in each group

	medical subjects	other subjects
no context	group 1, $n = 26$	group 2, $n = 26$
context	group 3, $n = 21$	group 4, $n = 22$

Procedure

Subjects received a one-page questionnaire. At the top of the page the task was introduced and instructions were given. The instructions were the same in both conditions, that is to order the eight expressions, be they presented in isolation or embedded in a sentence, by assigning a ranking number to each. The number 1 was to be given to the expression denoting the highest level of probability and subsequent numbers to expressions denoting subsequently less probability. Assignment of the same rank to more than one expression was allowed (compare [9]). Then the eight expressions or sentences were presented, listed vertically, indented and double-spaced. Presentation order was arbitrarily set to possible, impossible, uncertain, certain, probable, improbable, expected and undecided in the no context condition and to probable, improbable, possible, undecided, impossible, uncertain, expected and certain in the context condition. At the bottom of the page subjects were asked for their gender, year of birth and profession or study and thanked for their cooperation.

Subjects

Of the no context groups, group 1 consisted of 15 female and 11 male medical students. Their ages ranged from 19 to 45 years, with an average of 21 (SD = 5). Group 2 consisted of 19 female and 7 male social sciences students. Their ages ranged from 18 to 29, with an average of 21 (SD = 2.5). Of the context groups, group 3 consisted of 13 female and 8 male medical students, with ages ranging from 19 to 32 with an average of 22.5 (SD = 5). Group 4 consisted of 19 female and 3 male social sciences students, whose ages ranged from 19 to 26 with an average of 21 (SD = 1.6).

Data analysis

We analysed the between-group variance of the mean ranks given by the subjects to the eight probability expressions with a one-way ANOVA (ANalysis Of VAriance). Because we found some significant differences, we then analysed the data with a non-linear principal components technique developed by the University of Leiden PRINCALS, an acronym for principal components analysis by alternating least squares [20]. It may be used for ordinal data. Important for our study, PRINCALS can compute solutions that reduce the orderings of all subjects together to one or more dimensions and indicate the quality (eigenvalue, max. 1) of the solution on each dimension. We assumed that the subjects all did their ordering along the one dimension of level of probability. To test this assumption, we had PRINCALS compute a solution in two dimensions. If our assumption was correct, then the solution on one dimension would have a high quality and on the other dimension the quality would be low enough to be able to discard it.

Results

– ANOVA

Four subjects in group 1 and one subject in group 4 had to be excluded from the ANOVA analyses because they had given an incomplete ordering. The ANOVA analyses revealed that the four groups of subjects had assigned significantly different mean ranks to five of the eight terms: possible, impossible, improbable, expected and certain [3]. Post Hoc tests using Tukey's HSD-procedure showed that for these five expressions it were only the with context and without context group means that differed significantly at $\alpha = .05$, while there were no significant differences between medical subjects and other subjects for any of the expressions. Since the only factor that influenced differences in mean rankings was context, we present in Table 3 the mean rankings of the two no context groups together (groups 1 and 2) and the two context groups together (groups 3 and 4).

[3] With $F(3, 86)$ for possible = 4.018, $p = .01$, for impossible = 4.605, $p = .005$, for improbable = 5.684, $p = .001$, for expected = 4.481, $p = .006$ and for certain = 4.390, $p = .006$

Table 3
Mean ranks (and standard deviations) of the eight probability expressions, by the subjects in the no context condition (groups 1 and 2) and the subjects in the context condition (groups 3 and 4)

	no context ($n = 48$)	context ($n = 42$)
certain	1.15 (0.94)	2.26 (2.46)
probable	2.70 (0.78)	3.29 (1.89)
expected	2.65 (0.85)	3.57 (1.49)
possible	3.81 (0.52)	4.56 (1.76)
undecided	5.69 (1.22)	5.40 (1.56)
uncertain	5.96 (0.87)	5.50 (1.46)
improbable	6.44 (0.84)	5.13 (2.16)
impossible	7.61 (0.99)	6.29 (2.33)

Discussion

We concluded that context did indeed appear to influence the ranking of the expressions but that the medical subjects and the others did not differ in their rankings. We present our results next to those of other researchers. These other studies were all ranking experiments in which the expressions were presented without a context, so we include only our no context results. To facilitate comparison, we rescaled the means of the other four studies onto a 1 to 8 scale (see Table 4). Where there were expressions in common, the correspondence between our results and the other four is satisfactory.

Although other researchers generally only present means, we feel uncomfortable calculating mean ranks. The subjects only assigned rank numbers, not distances between the expressions, so the data are ordinal. If an expression is ranked fourth, that does not necessarily mean that it refers to twice as much improbability as an expression ranked second. Also, the high standard deviations for the mean ranks of the context group (compare Table 3) are difficult to explain by just calculating means. Possibly the subjects did not rank the expressions on the one dimension of no to complete probability, but on another dimension as well. To check this, we performed additional PRINCALS analyses.

– PRINCALS

No subjects had to be excluded for the PRINCALS analyses, as they had been from the ANOVA analyses, because PRINCALS can be performed with missing data. PRINCALS reveals the order underlying the rankings of the expressions by the different subjects.

Table 4
Mean ranks (and standard deviations) of the eight probability expressions by the subjects in the no context condition (groups 1 and 2) and as reported by Tavana et al. [48], by Budescu & Wallsten [7]and by Clark [9], studies 5.4 and 5.2

	no context (this study) ($n = 48$)	Tavana ($n = 30$)	Budescu ($n = 32$)	Clark study 5.4 ($n = 16$)	Clark study 5.2 ($n = 16$)
certain	1.15 (0.94)	1.05	–	–	1.36
probable	2.70 (0.78)	–	2.80	2.83	2.56
expected	2.65 (0.85)	–	–	2.62	–
possible	3.81 (0.52)	5.29	4.71	3.61	3.62
undecided	5.69 (1.22)	–	–	–	–
uncertain	5.96 (0.87)	–	4.94	5.38	5.94
improbable	6.44 (0.84)	–	6.22	6.45	6.83
impossible	7.61 (0.99)	8.00	–	7.87	7.90

For both groups in the no context condition a solution was found in one dimension for most subjects, with an eigenvalue of 0.9175 for group 1 and 0.9500 for group 2. For two medical subjects and one other subject a high quality solution was found on the second dimension. Because on inspection of their answers there seemed to be no logical explanation to their orderings, we presumed that these three subjects had misunderstood the task and we excluded their data. The preference orderings for the eight probability expressions of the rest of the subjects, 24 in group 1 and 25 in group 2, were quite the same. Since the ANOVA analyses had shown that there were no differences between medical subjects and others, we took these two together as the no context group. A high quality solution with an eigenvalue of 0.9504 was found in one dimension, with the expressions in the order presented in the first column in Table 5.

For the two groups in the context condition, the second dimension was important. For the medical subjects (group 3), a high quality solution with an eigenvalue of 0.9550 was found in one dimension for twelve of the subjects, with comparable preference orderings, while nine subjects scored high on the second dimension. For the other subjects (group 4), a high quality solution with an eigenvalue of 0.9515 was found for eleven of the twenty-two subjects on the first dimension, with the same preference orderings, while the other eleven subjects scored high on the second dimension.

The nine medical subjects and eleven others who scored high on the second dimension appeared to have judged the probability that the sentences in which

the expressions were embedded were truthful statements instead of judging the expressions themselves. As an illustration, one of these subjects judged 'improbable' in the sentence "It is improbable that someone with tonsillitis does not have a sore throat" to express the highest level of probability and 'possible' in the sentence "It is possible that someone faints from the heat" to express the lowest level of probability. But taking together these apparently sentence-ranking subjects did not reveal an understandable pattern. We speculate that another factor had influenced these rankings, possibly familiarity with the complaint for the medical students and everyday beliefs about such complaints for the other students. Again taking the medical subjects and the others together, and excluding the subjects who seemed not to have followed our instructions to rank order the expressions, we found a high quality solution with an eigenvalue of 0.9652 for both groups in the context condition together with the expressions ordered as shown in the second column of Table 5.

Table 5
Rank order of the eight expressions of probability for the medical subjects and the other subjects in the no context condition (groups 1 and 2) and the medical and other subjects in the context condition (groups 3 and 4)

	no context condition groups 1 and 2 ($n = 49$)	context condition groups 3 and 4 ($n = 23$)
certain	1	1
probable	3	2.5
possible	3	4
expected	3	2.5
improbable	5.5	6
uncertain	5.5	6
undecided	8	6
impossible	7	8

Discussion

Surprisingly, the term we had introduced as the mid-point ('undecided') was ranked last by the subjects in the no context condition. Looking back to the calculated means of this expression in Table 3, we also see a high standard deviation. Clearly, the interpretation of 'undecided' is not unambiguous. An explanation may possibly be the order of presentation: 'undecided' was the last expression in the list.

Our PRINCALS analyses showed that not all subjects in the context condition rank-ordered the expressions. Almost half of them gave rankings on a second

dimension. This was not true for the subjects in the no context condition. We therefore conclude that context does not influence the rank ordering of the expressions themselves, but context does seem to distract subjects from the actual task.

We performed a final analysis over all four groups in the two conditions who had ranked the expressions on one dimension ($n = 72$). Their rankings could be summarised in one dimension with an eigenvalue of 0.9658, representing the order: certain, probable, expected, possible, (uncertain, improbable, undecided), impossible. Because in our opinion the PRINCALS analyses are more appropriate than the mean rank orderings, we summarise the results of this second experiment as revealing the following rank order of our eight expressions: certain and impossible at the extremes, with probable, expected and possible, in that order, expressing less probability from the certain-side down, and uncertain, improbable and undecided toward the impossible-side.

THIRD EXPERIMENT

We were not satisfied with an order of expressions only, but we also wished to establish whether two (or more) expressions were taken to mean almost the same or were quite distinguishable in meaning. In other words, we wanted to know the 'distances' between the expressions. We therefore set up a third study, in which we asked subjects to rate sameness of or difference between expressions. We expected to find that 'certain' and 'impossible' would be judged as extremely different, while 'possible', 'probable' and 'expected' might be rather similar.

Procedure
In the third experiment we asked subjects for pairwise comparison, that is: for similarity judgements among all pairs of verbal probability expressions, to uncover the underlying structure of relationships among them [44,4]. For the eight expressions, there were 28 pairs to compare. A similarity judgement was made by scoring each pair of expressions on a 10 cm. line, using as anchors the expressions 'exact same' and 'completely different'. Each judgement was made on a separate sheet of paper. The order of presentation was random across subjects and across stimulus pairs, and, for a pair AB, half the subjects received A first while the other half received B first. Subjects performed four practice runs before starting the real experimental judgements.

Subjects
We had two groups of subjects, again one group with a medical background and one comparable group with no medical background. Subjects in group 1 were 28 medical biology students, 12 female and 16 male. Their age range was

between 19 and 25, with an average of 20 (SD = 1.5). Subjects in group 2 were 56 computer science students, 13 female and 43 male, with ages ranging from 20 to 53, an average of 24 (SD = 4).

Data analysis

The judgement of (dis)similarity between each pair of expressions for each subject was scored in millimetres, read from a ruler placed against the 10 cm. line. For each subject a data matrix was drawn up, in lower triangular form with zero's on the diagonals. The matrices were analysed with a Multi-Dimensional Scaling (MDS) technique. MDS takes a set of distances between objects and creates the 'map' by computing the positions (co-ordinates) of the objects.

We used the SPSS module ALSCAL (Alternating Least-Squares Scaling) as our MDS procedure [47]. The type of MDS we used was Replicated MDS, which computes a single MDS solution for all matrices together and uses the Euclidean distance model (an n-dimensional version of the Pythagorean theorem) as scaling model. The data were treated as continuous ordinal data (i.e. ties within the matrices were untied) and as matrix conditional (i.e. the meaning of the numbers in a matrix is conditional on the subject). Since all probability expressions seemed to be comparable, we did an analysis in only one dimension.

ALSCAL produces a list of co-ordinates for the eight probability expressions. These co-ordinates are such that their fit with the distances between the expressions given by the different subjects is as good as possible. We mapped the co-ordinates of the expressions onto a probability scale, by setting as anchors the extreme expressions (certain and impossible) representing 100% and 0% probability respectively, and then using a linear function to calculate the probabilities of the other expressions (compare [48]).

Results

Medical students

An initial ALSCAL analysis of the matrices of the medical students, group 1, showed that the matrices of two subjects didn't fit the calculated co-ordinates. Upon inspection of the matrices of these two subjects, this bad fit seemed to be the result of their judgement of certain and impossible as very similar (a distance of 1 millimetre on the 10 cm. line, where one would expect the full 10 cm.).

We removed the data from these two subjects and did another analysis with the remaining 26 matrices. The co-ordinates of the eight expressions are given in the left half of the leftmost double column of Table 6 below. The right half of this double column presents a mapping of these co-ordinates onto a

probability scale from one to zero, calculated with the function

$$\text{probability} \;=\; (\text{coordinate} \;+ 1.4572) \div 2.6522.$$

Other students

An initial analysis of the matrices of the other students, group 2, showed that the matrices of four subjects had to be removed because of their poor fit. The analysis with the remaining 52 matrices gave the co-ordinates of the eight expressions as given in the left half of the middle double column of Table 6 below. The right half of this column presents a mapping of these co-ordinates onto a probability scale, calculated with the function

$$\text{probability} \;= (\text{coordinate} \;+ 1.5394) \div 2.8346.$$

Table 6
Co-ordinates and calculated probability points for the eight expressions of group 1, medical students ($n = 26$), group 2, other students ($n = 52$) and all subjects together ($n = 78$)

	group 1		group 2		all subjects	
expression	co-ord.	prob.	co-ord.	prob.	co-ord.	prob.
certain	1.1950	1.00	1.2952	1.00	1.2738	1.00
possible	1.0897	0.96	0.8284	0.84	0.9105	0.86
probable	0.8409	0.87	0.9252	0.87	0.9043	0.86
expected	0.7239	0.82	0.7211	0.80	0.7133	0.79
undecided	−0.5972	0.32	−0.3730	0.41	−0.4394	0.38
uncertain	−0.7210	0.28	−0.8139	0.26	−0.7939	0.25
improbable	−1.0741	0.14	−1.0435	0.17	−1.0610	0.16
impossible	−1.4572	0.00	−1.5394	0.00	−1.5075	0.00

All subjects

We performed a final analysis over the two groups, thus over the 78 matrices. The right double column of Table 6 presents the results of this analysis, again with the calculated probabilities, with the function

$$\text{probability} \;= (\text{coordinate} \;+ 1.5075) \div 2.7813.$$

Discussion

The calculated probabilities of 'probable' and 'possible' are close together in the final analysis, and different (and inverted) with the medical students and

the other students, suggesting they could be taken to refer to the same range on the scale and that one of them can be removed. We will leave out 'possible', because its calculated probabilities differ most in the two groups. Note that the calculated probabilities for 'undecided' are very different for the two groups as well. Since it was again (compare experiment 2) not interpreted as intended, that is for the mid-point, we decided to leave out this term as well. This leaves us with a scale without a mid-point, so we add one term which can hardly be misunderstood, 'fifty-fifty', although one can argue whether this really is a verbal probability expression.

Upon closer inspection of the matrices, we saw that the positive-negative pairs certain-uncertain and possible-impossible were judged by most subjects as 100% dissimilar. Taking all expressions into consideration, 'uncertain' and 'possible' may be expected to be at some distance from the extremes 'impossible' and 'certain', but our method of eliciting pair-wise dissimilarity judgements artificially forced interpretation of the expressions toward the endpoints of the scale. We thus feel justified to slightly reinterpret the calculated probabilities toward the mid-point, resulting in the scale with seven categories presented in Table 7, which we will use in our next and final experiment.

Table 7
Final scale with seven categories of probability expressions plus their calculated probability points

	expression	probability
I	certain	100%
II	probable	85%
III	expected	75%
IV	fifty-fifty	50%
V	uncertain	25%
VI	improbable	15%
VII	impossible	0%

FOURTH EXPERIMENT

Our fourth experiment was designed to test the translations of the verbal probability expressions in our scale into the calculated numerical probabilities. In contrast to what most other researchers have done, we didn't ask subjects to translate words into numbers or vice versa. We think that having to give such a translation is an artificial task, not true to actual cognitive processes (compare [9]). It may not capture how people actually use words and numbers for probability. However, we did need to validate our translations.

We did so by comparing the decisions subjects made when they were offered probability information in verbal form to their decisions when the information was presented numerically. If the calculated probability points indeed have the same meaning as the verbal probability expressions, decision makers will make similar decisions with the probability information presented verbally and numerically. This would be even more convincing if the decisions were also made with comparable confidence. For example, we expect that when subjects decide, with high confidence, to cancel an appointment when they are informed that rail-workers will 'probably' continue their strike, they also decide, highly confidently, to cancel their appointment when rail-workers continue their strike with 85% probability.

Procedure

Subjects received a two-page questionnaire, with an introduction followed by six decision situations. Each decision situation was described in two or three lines. To give an example:

> Ms. T. has a non-serious physical complaint, which does however need to be treated. The probability that Ms. T. is allergic to the usually prescribed drug H. is .... There are alternative drugs for her complaint, but these are less effective. Do you prescribe drug H?

Each of the six descriptions was followed by a table that contained either the seven verbal probability expressions or the seven numerical probabilities (see Table 7), with each of these seven followed by "decision: yes/no" (to be circled) and by a 2-cm line on which subjects were to check their measure of confidence in their decision (from complete to none). The subjects were instructed to mentally write, on the dots in the description, each of the expressions in turn, to make a yes/no decision for that hypothetical situation and to check their confidence. Each subject thus made seven decisions plus confidence checks, for each of the six situations.

We had four versions of the questionnaire. Version one started with three situations (A, B and C) with verbal expressions, followed by three situations (D, E and F) with the numerical probabilities. Version two contained the same six situations in the same order, but now with situations A, B and C with numerical probabilities and situations D, E and F with verbal expressions. In versions three and four the six situations were given in the order D, E, F followed by A, B, C, with version three starting with verbal expressions and version four with the numerical probabilities. The tables of expressions and probabilities, each of which occurred three times in a questionnaire, werefirst given in the order we had determined (as in Table 7), then twice in a different random order.

Subjects
There were 123 participants (students and faculty members of Computer Science, Psychology, Artificial Intelligence, and Medicine), 59 female and 64 male, whose ages ranged from 18 to 61 (mean of 28, SD = 9.7).

Data analysis
We had to remove the answers of 12 of the 123 subjects, because they had misunderstood the assignment and had made a decision for only one category in each situation. Of the 110 subjects left, 52 answered version one or three of the questionnaire and 58 subjects answered version two or four.

For each of the six situations we have a verbal and a numerical answering mode. In each mode a yes or no decision was made for seven probability categories. With these three variables (mode, decision and category) we constructed a three-way table for each situation. In the cells of the tables we have the total number of subjects who made a certain decision in a certain mode for a certain category. For example: 34 subjects decided 'n' (no) in mode V (verbal) for category IV ('fifty-fifty'). The same three-way tables, with mode-decision-category, were drawn up with the cells containing the subjects' confidence. We measured the confidence subjects had in their decisions by scoring their checks on the confidence line. Complete confidence was counted as 1.0, no confidence as 0.0. We looked both at the confidence of all subjects together for each mode-decision-category tuple, and the mean confidence for each tuple. This gave us another two sets of six three-way tables with in each cell the total and the mean confidence, respectively.

We analysed our three-way tables using a log-linear analysis (the log of the expected cell frequencies is a linear function of the log of the observed frequencies of the variables). We tried to find a log-linear model that describes all important associations between the variables, in our case mode, with values N(umeric) and V(erbal), decision, with values y(es) and n(o), and category, with values I through VII.

A log-linear model describes the associations between variables. For example: the model "Category × Decision × Mode" describes three-way association, between all three variables; the model "Category × Decision + Mode" indicates that only Category and Decision are associated and that Mode is unrelated to either.

Fitting a log linear model is a process of deciding which associations are significantly different from zero. A model that includes these significant associations fits the data, that is: explains the observed frequencies. It is known that the tests used to determine the fit are somewhat too liberal, especially if some expected cell frequencies are small. Therefore, Darlington [11] suggests correcting the observed frequencies before computing the fit. This continuity correction

consists of adjusting each value of the observed frequencies by 0.25 toward its own expected frequency. We used this correction in situations where no simple model would fit the data and we thought this could be caused by small cell frequencies. Note that especially the situations with the extreme probability categories (I and VII) will probably have small cell frequencies for one of the decisions.

We performed separate analyses for the decisions, over the tables with the total number of subjects per cell, and for the confidence, over the tables with total confidence per cell and those with mean confidence per cell.

Results

Decisions

For five of the six situations we could directly choose "Category × Decision + Mode" as the best model [4]. In other words: the decisions were related to the category of the probability expression, and not related to the mode, verbal or numerical, in which the probabilities were given. In the sixth situation the only model that fit initially was "Category × Decision × Mode", which will always fit. We therefore performed a continuity correction on our data. Then the same model as for the other five situations had the best fit, but it was not quite convincing [5]. Examining the cells, we retraced the sub optimality to one category: the proportion of $yes : no$ decisions in category V differed a factor 4 between the numerical mode (25%) and the verbal mode (uncertain).

Confidence

For the tables with the total confidence per cell we could, again, directly see that "Category × Decision + Mode" was the best model for four out of the six situations [6]. This means that the confidence subjects had in their decisions was related to the category of the probability expression, and not related to the mode (verbal or numerical) in which the probabilities were given. In one of the two remaining situations the model "Category × Decision + Mode × Decision" was initially significantly better. However, after we performed a continuity correction on our data the same "Category × Decision + Mode" model was again the best model [7]. In the other remaining situation the only model that fit was "Category × Decision × Mode". Performing a continuity correction did not change this. In this situation we again had a big difference in the proportion of $yes : no$ decisions for the two modes of category V. Deleting

[4] with, respectively, $\chi^2 = 2.876$, $p = 0.998$; $\chi^2 = 15.468$, $p = 0.279$; $\chi^2 = 16.085$, $p = 0.245$; $\chi^2 = 9.438$, $p = 0.739$ and $\chi^2 = 9.261$, $p = 0.753$; $df = 13$ for all tables

[5] $\chi^2 = 20.342$, $p = 0.087$, $df = 13$

[6] with, respectively, $\chi^2 = 2.574$, $p = 0.999$; $\chi^2 = 12.541$, $p = 0.484$; $\chi^2 = 7.520$, $p = 0.873$ and $\chi^2 = 9.208$, $p = 0.757$; $df = 13$ for all tables

[7] $\chi^2 = 13.138$, $p = 0.437$, $df = 13$

category V and a continuity correction again showed "Category × Decision + Mode" to be the best model[8].

Analyses of the tables with mean confidence per cell showed that in the six situations there was no difference in the subjects' mean confidence in their decisions and no difference between the subjects' mean confidence for decisions in the verbal mode and in the numerical mode. Subjects were consistent in their confidence judgements over situations as well as in the two modes.

Discussion
Our analyses showed that the decisions subjects made depended only on the probability category used in the description and that the decisions were not influenced by the mode in which the probability information was presented.

We did find that category V caused some problems in some situations. This could indicate that to some subjects 'uncertain' and '25%' does not mean the same. Indeed, to some people 'uncertain' may mean anything less than a 100% certain, others could interpret 'uncertain' to mean the same as 'fifty-fifty'. However, this problem only occurred in the situations where the probability categories were presented in random order. It did not occur when the probabilities were presented in order. In fact, the best model values were those for the situations in which the ordered lists were presented[9]. We conclude that when the probability expressions are presented in an ordered list, they will be interpreted as intended.

We may conclude that context influences the decisions people make, but because we only found differences in decisions between the situations and not per situation between the verbal or numerical mode, the two modes are interchangeable and neither is better or worse. Our results suggest that the agreement between the calculated probability points and the verbal probability expressions, given in Table 7, is reliable.

CONCLUSIONS

In some situations people prefer to express and process probabilities in verbal rather than numerical form. Knowledge-based systems such as belief networks on the other hand always internally represent and compute probabilities numerically. This means that experts are required to state their probabilities numerically and to understand explanations containing numbers. We suggest that this communication problem would be reduced if there were to exist an

[8] $\chi^2 = 15.287$, $p = 0.170$, $df = 13$
[9] $\chi^2 = 2.876$, $p = 0.998$ for the decisions and $\chi^2 = 2.547$, $p = 0.999$ for the confidence

acceptable representation of a probability scale that contained mutually exchangeable verbal and numerical expressions.

Our first three experiments provided us with an ordered list of seven commonly used verbal probability expressions, which together span the whole scale. Unlike other researchers we did not use a pre-set list but we worked with the expressions people themselves said they most commonly used. Our experiments differed from others in another important aspect. We did not ask people to translate numerical expressions into numbers or vice versa. In our opinion, asking for such a translation forces subjects to usetwo different mental representations of probability at the same time and to look for a mapping between the two. We addressed only one representation, thereby avoiding possible confusion. We tried to construct a scale for the verbal representation of probability, the numerical scale being quite straightforward. In order to present the two together along one scale, we needed numerical equivalents for the verbal expressions. We used the dissimilarities of the third experiment to determine this mapping of verbal expressions onto a numerical scale.

It is often said that numbers are better than words, because words are more variably interpretable, the meaning being influenced by, among other things, context and personal opinions. The assumption then is that numerical probabilities are always interpreted in the same way and that, since a verbal expression is translated into different numerical expressions, the verbal expression is too vague. Obviously, uncertainty is always dealt with within a context. This context can be either explicit, or people will implicitly think of one. We assume that context not only influences the interpretation of verbal probability expressions, but that it influences the interpretation of a numerical probability expression as well. For example, if 'low probability of infection' and 'low probability of death' are interpreted differently, then so will 'a 23% chance of infection' and 'a 23% chance of death'. For expressing uncertainty, numbers may be just as vague as words. Therefore, we should not ask people whether they think that a low probability equals 25%, but we should test if, in a certain context, they interpret 'low probability' the same as '25%'; that is: we should test if people react the same to the verbal and the numerical expression, if they take the same actions, make the same decisions. This is what we have done in our last experiment.

This fourth experiment was designed to test the validity of our translations. The finding that subjects made the same decisions, with the same confidence, irrespective of whether information was communicated to them in terms of the verbal expressions or in the corresponding numerical form, justifies the tentative conclusion that this scale containing both is usable. Further study is now called for to check the scale's usability in the applied context of belief networks. We are currently interviewing medical experts on cancer to get their probability estimates for a belief network, the qualitative part of which they

had already constructed. Previous interviews, in which these experts where asked to state their assessments numerically or to mark them on a horizontal line, were quite unsuccessful. Now, with our double scale, elicitation proceeds much more effectively, to a much greater satisfaction of the experts [51]. Some probabilities they easily give as a number, for others they use the verbal expressions and then check the scale at or near the expression that best fits their estimate. We will continue this study, and also set up a more systematic investigation into the benefits of the use of the double scale. A more systematic study is also called for into the explanations generated by the system, and the user's ability to understand these when they are offered the double scale as a reference.

There are some shortcomings to our study. We used Dutch subjects and consequently Dutch words, which we translated into English for this paper. Although we made no choices for the English words because dictionaries only give the one translation for each term, we cannot be sure that the connotations of the Dutch and the English words are similar. A replication with English speaking subjects could verify this point. In some situations, our scale may seem too coarse, containing too few verbal expressions. However, the expressions are not meant to be presented as a list, but next to a graphical representation of a 0 to 100 scale. Users may check this scale right next to a word, or at any point between two words. We propose a representation as shown in Figure 1. We found this scale to be quite usable to experts.

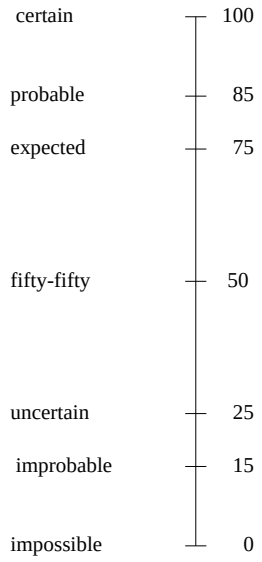

Fig. 1. Final Verbal and Numerical Probability Scale

We are not quite happy with the way the expression for the midpoint of the scale was determined. First, because our subjects did not write down such a term, we introduced 'undecided', after much thought. Its literal meaning may be fifty-fifty, but in the experiments subjects did not appear to interpret it this way. We then replaced it by 'fifty-fifty', which seems cheating on what

may count as verbal. Moreover, because we introduced this term later, the assumed distance to the other terms was not established as it had been for the rest of the terms.

In spite of these shortcomings, we think we have shown how people's preferences for verbal probability expressions may be accommodated. This may prove helpful in the construction of for example belief networks or other systems that represent and process probability information.

Acknowledgements: We would like to thank Linda van der Gaag and several reviewers for their useful comments on earlier versions of this paper. We are also grateful to Herbert Hoijtink and Pieter Koele for their advice about and support with the statistical procedures.

References

[1] S. Andreassen, M. Woldbye, B. Falck, and S.K. Andersen. MUNIN. A causal probabilistic network for interpretation of electromyographic findings. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 366 – 372, 1987.

[2] R. Beyth-Marom. How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting*, 1:257 – 269, 1982.

[3] P.P. Bonissone and K.S. Decker. Selecting uncertainty calculi and granularity: An experiment in trading-off precision and complexity. In L.N. Kanal and J.F. Lemmer, editors, *Uncertaintyy in Artificial Intelligence*, pages 217 – 247, North-Holland, 1986. Elsevier Science Publishers B.V.

[4] I. Borg and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York, 1997.

[5] W. Brun and K.H. Teigen. Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, 41:390 – 404, 1988.

[6] G.D. Bryant and G.R. Norman. Expressions of probability: Words and numbers. *The New England Journal of Medicine*, 302(7):411, 1980.

[7] D.V. Budescu and T.S. Wallsten. Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes*, 36:391 – 405, 1985.

[8] D.V. Budescu, S. Weinberg, and T.S. Wallsten. Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance*, 14(2):281 – 294, 1988.

[9] D.A. Clark. *Psychological Aspects of Uncertainty and their Implications for Artificial Intelligence*. PhD thesis, Department of Applied Psychology, University of Wales Institute of Science and Technology, 1988.

[10] L.D. Cohn, M. Schydlower, J. Foley, and R.L. Copeland. Adolescents' misinterpretation of health risk probability expressions. *Pediatrics*, 95(5):713 – 716, 1995.

[11] R.B. Darlington. *Regression and Linear Models*, chapter 19. McGraw-Hill, Singapore, 1990.

[12] F.J. Dietz, J. Mira, E. Iturralde, and Z. Zubillage. DIAVAL, a Bayesian expert system for echocardiography. *Artificial Intelligence in Medicine*, 10(1):59 – 73, 1997.

[13] M.J. Druzdzel. Qualitative verbal explanations in Bayesian belief networks. *Artificial Intelligence and Simulation of Behaviour Quarterly, special issue on Bayesian belief networks*, 94:43 – 54, 1996.

[14] M.J. Druzdzel and L.C. van der Gaag. Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI – 95)*, pages 141 – 148, 1995.

[15] J.A.H. Eekhof, S.S.L. Mol, and J.C. Pielage. Is doorgaans vaker dan dikwijls; of hoe vaak is soms? *Nederlands Tijdschrift voor Geneeskunde*, 136(1):41 – 42, 1992.

[16] C. Elsaesser. Explanation of probabilistic inference. In L.N. Kanal, T.S. Levitt, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence 3*, pages 387 – 400. Elsevier Science Publishers, North-Holland, 1989.

[17] I. Erev and B.L. Cohen. Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes*, 45:1 – 18, 1990.

[18] K. Fischer and H. Jungermann. Rarely occurring headaches and rarely occurring blindness: Is rarely = rarely? The meaning of verbal frequentistic labels in specific medical contexts. *Journal of Behavioral Decision Making*, 9:153 – 172, 1996.

[19] J. Fox, D. Barber, and K.D. Bardhan. Alternatives to Bayes? A quantitative comparison with rule-based diagnostic inference. *Methods of Information in Medicine*, 19:210 – 215, 1980.

[20] A. Gifi. *Nonlinear Multivariate Analysis*. Wiley, Chichester, 1991.

[21] M. Gonzalez and C. Frenck-Mestre. Determinants of numerical versus verbal probabilities. *Acta Psychologica*, 83:33 – 51, 1993.

[22] R.M. Hamm. Selection of verbal probabilities: A solution for some problems of verbal probability expression. *Organizational Behavior and Human Decision Processes*, 48:193 – 223, 1991.

[23] M. Henrion, M. Pradhan, B. Del Favero, K. Huang, G. Provan, and P. O'Rorke. Why is diagnosis using belief networks insensitive to imprecision in probabilities? In *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence (UAI – 96)*, pages 307 – 314, 1996.

[24] E.K.R.E. Huizingh and H.C.J. Vrolijk. A comparison of verbal and numerical judgments in the analytic hierarchy process. *Organizational Behavior and Human Decision Processes*, 70:237 – 247, 1997.

[25] F.V. Jensen, J. Nielsen, and H.I. Christensen. Use of causal probabilistic networks as high level models in computer vision. Technical Report R-90-39, University of Aalborg, 1990.

[26] R.M. Kenney. Between never and always. *The New England Journal of Medicine*, 305(18):1097 – 1098, 1981.

[27] A. Kong, G.O. Barnett, F. Mosteller, and C. Youtz. How medical professionals evaluate expressions of probability. *The New England Journal of Medicine*, 315:740 – 744, 1986.

[28] M. Korver and P.J.F. Lucas. Converting a rule-based expert system into a belief network. *Medical Informatics*, 18(3):219 – 241, 1993.

[29] P. Krause and D.A. Clark. *Representing Uncertain Knowledge: An Artificial Intelligence Approach*. Intellect, Oxford, 1993.

[30] B. Kuipers, A.J. Moskowitz, and J.P. Kassirer. Critical decisions under uncertainty: Representation and structure. *Cognitive Science*, 12:177 – 210, 1988.

[31] P.J.F. Lucas, H. Boot, and B.G. Taal. Computer-based decision-support in the management of primary gastric non-hodgkin lymphoma. *Methods of Information in Medicine*, 37:206 – 219, 1998.

[32] J.F. Merz, M.J. Druzdzel, and D.J. Mazur. Verbal expressions of probability in informed consent litigation. *Medical Decision Making*, 11:273 – 281, 1991.

[33] G.A. Miller. The magical number seven plus or minus two: Some limits on our capacity to process information. *Psychological Review*, 63:81 – 87, 1956.

[34] F. Mosteller and C. Youtz. Quantifying probabilistic expressions. *Statistical Science*, 5(1):2 – 12, 1990.

[35] M.A. Nakao and S. Axelrod. Numbers are better than words. Verbal specifications of frequency have no place in medicine. *The American Journal of Medicine*, 74:1061 – 1065, 1983.

[36] K. Ng and B. Abramson. A sensitivity analysis of pathfinder: A follow-up study. In *Proceedings of the Seventh Annual Conference on Uncertainty in Artificial Intelligence*, pages 242 – 248. Morgan Kaufmann Publishers, 1991.

[37] B.J. O'Brien. Words or numbers? The evaluation of probability expressions in general practice. *Journal of the Royal College of General Practitioners*, 39:98 – 100, 1989.

[38] M.J. Olson and D.V. Budescu. Patterns of preference for numerical and verbal probabilities. *Journal of Behavioral Decision Making*, 10:117 – 131, 1997.

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

[40] M. Pradhan, M. Henrion, G. Provan, B. Del Favero, and K. Huang. The sensitivity of belief networks to imprecise probabilities: An experimental investigation. *Artificial Intelligence*, 85:363 – 397, 1996.

[41] A. Rapoport, T.S. Wallsten, I. Erev, and B.L. Cohen. Revision of opinion with verbally and numerically expressed uncertainties. *Acta Psychologica*, 74:61 – 79, 1990.

[42] E. Reiss. In quest of certainty. *The American Journal of Medicine*, 77(6):969 – 971, 1984.

[43] W.O. Robertson. Quantifying the meaning of words. *Journal of the American Medical Association*, 249(19):2631 – 2632, 1983.

[44] S.S. Schiffman, M.L. Reynolds, and F.W. Young. *Introduction to Multidimensional Scaling. Theory, Methods, and Applications*. Academic Press, New York, 1981.

[45] S.E. Stheeman, P.A. Mileman, M.A. van 't Hof, and P.F. van der Stelt. Blind chance? An investigation into the perceived probabilities of phrases used in oral radiology for expressing chance. *Dentomaxillofac. Radiol.*, 22(2):135 – 139, 1993.

[46] D.N. Stone and D.A. Schkade. Numeric and linguistic information representation in multiattribute choice. *Organizational Behavior and Human Decision Processes*, 49:42 – 59, 1991.

[47] Y. Takane, F.W. Young, and J. DeLeeuw. Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42(1):7 – 67, 1977.

[48] M. Tavana, D.T. Kennedy, and B. Mohebbi. An applied study using the Analytic Hierarchy Process to translate common verbal phrases to numerical probabilities. *Journal of Behavioral Decision Making*, 10:133 – 150, 1997.

[49] D. Timmermans, J. Kievit, and H. van Bockel. How do surgeons' probability estimates of operative mortality compare with a decision analytic model? *Acta Psychologica*, 93:107 – 120, 1996.

[50] D.R.M. Timmermans and P.A. Mileman. Lost for words: Using verbal terms to express probabilities in oral radiology. *Dentomaxillofac. Radiol.*, 22:171 – 172, 1993.

[51] L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, and B.G. Taal. How to elicit many probabilities. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI – 99)*, 1999.

[52] Y. Waern, S. Hägglund, R. Ramberg, L. Rankin, and J. Harrius. Computational advice and explanations – behavioral and computational aspects. In K. Norby, P. Helmersen, D.J. Gilmore, and S.A. Arnesen, editors, *Human–Computer Interaction*, pages 203 –206. Chapman and Hall, London, 1995.

[53] T.S. Wallsten, D.V. Budescu, A. Rapoport, R. Zwick, and B. Forsyth. Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, 115(4):348 – 365, 1986.

[54] T.S. Wallsten, D.V. Budescu, and R. Zwick. Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, 39(2):176 – 190, 1993.

[55] T.S. Wallsten, D.V. Budescu, R. Zwick, and S.M. Kemp. Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society*, 31(2):135 – 138, 1993.

[56] E.U. Weber and D.J. Hilton. Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance*, 16:781 – 789, 1990.

[57] A.C. Zimmer. Verbal versus numerical processing of subjective probabilities. In R.W. Scholz, editor, *Decision Making under Uncertainty*, pages 159 – 182. North-Holland, Amsterdam, 1983.