Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING

ELSEVIER International Journal of Approximate Reasoning 33 (2003) 117–131

www.elsevier.com/locate/ijar

Evaluation of a verbal–numerical probability scale ☆

Cilia Witteman [1], Silja Renooij *

Institute of Information and Computing Sciences, Utrecht University, Padualaan 14, P.O. Box 80.089, 3508 TB, Utrecht 3584 CH, The Netherlands

Received 1 August 2002; received in revised form 1 October 2002

Abstract

For purposes such as the development of decision support systems, the probabilities that model the uncertainties in the domain of application are usually elicited from domain experts. A number of elicitation methods is available. While constructing a real-life system, we however found none of these methods to be quite usable: they turned out to be too time-consuming and difficult for experts. In an earlier paper we described a verbal–numerical response scale we developed to facilitate elicitation of a large number of probabilities. In this paper we describe a study that justifies our claim that use of this verbal–numerical scale generally facilitates the assessment process.

Keywords: Probability elicitation; Decision support systems; Accuracy of assessments; Response scale

1. Introduction

Expert models for decision support systems generally need to represent uncertainty, which is most often captured in the form of probabilities. Nowadays

☆ This research was partly supported by the Netherlands Organisation for Scientific Research.
* Corresponding author. Tel.: +31-30-2539266; fax: +31-30-2513791.
E-mail addresses: c.witteman@ped.kun.nl (C. Witteman), silja@cs.uu.nl (S. Renooij).
[1] Present address: Diagnostic Decision Making, Faculty of Social Sciences, University of Nijmegen, Nijmegen, The Netherlands.

0888-613X/02/$ - see front matter
doi:10.1016/S0888-613X(02)00151-2

such systems are often based on a probabilistic network, which is a mathematical model firmly rooted in probability theory [5]. The probabilistic information required for such a network may be available in textbooks or databases, but is almost always, at least partly, elicited from domain experts.

Different elicitation methods are available to ask experts for their probability judgments (for an overview, see e.g., [6,8]). Among these elicitation methods, the best-known direct method is the presentation of a horizontal or vertical numerical scale, with, for example, five anchors labelled with 0%, 25%, 50%, 75% and 100%. Judges are asked to mark a position on the scale, after which the indicated probability is determined by measuring the distance between this mark and 0%. Such a scale is easy to understand and use. Its drawback is that it is difficult for judges to mark fine distinctions; this is especially important at the endpoints, e.g. between probabilities of 0.01 and 0.001. Two indirect methods that are often used are gambles and probability wheels. Indirect methods infer a probability assessment from judges' choice behaviour in a controlled situation. With the gamble method, a judge is presented with a choice between two lotteries. In one of the lotteries the probability of winning is the probability of the event that is to be assessed, in the other lottery the probability of winning is set by the elicitor. The elicitor varies the latter probability until the judge is indifferent about the two lotteries; then the to-be-assessed probability may be determined. With the probability wheel method, the judge is asked to compare the probability that a pointer lands on, e.g., the red section of the wheel, with the probability of the event under consideration. The to-be-assessed probability is the proportion of the red section of the wheel when the judge is indifferent about the two chances. The assumption underlying these indirect methods is that they yield unbiased assessments, but major drawbacks show up in their application. The choices presented are often difficult to conceptualise, and the methods are difficult to learn, especially the lotteries, and very time-consuming in use.

We tried out the described methods in constructing a real-life probabilistic network for therapy selection in the domain of oesophageal cancer, for which we required the assessment of 4000 point probabilities [11]. We found that for such a large number of assessments, these standard methods were unsuitable. A numerical scale gave our experts too little to go by. The gambles proved to be too difficult and time-consuming, and the experts were aversive to, albeit hypothetically, gambling with their patients' health.

In order to elicit the required probabilities, we had to try a different approach, and we therefore designed our own elicitation method. Our new method is based on the use of a probability scale with both verbal and numerical labels. The design of the scale is described in [7] and briefly reviewed in Section 2. The method worked quite well with our experts and our domain, and we hope that it will be applicable in general, but of this we cannot be certain without further study. Section 3 introduces the study we undertook to ratify

our scale; Section 4 describes the results and analyses. Conclusions are presented in Section 5.

2. Design and initial use of the verbal–numerical scale

In an earlier paper [7] we describe a series of four studies that resulted in the design of a scale with both verbal and numerical probability labels, to be used in a probability elicitation method. We briefly summarise these four studies here; the interested reader is referred to [7] for details and also for an extensive and comprehensive review of literature on the subject of the use of verbal probability expressions.

In the first study we asked subjects which verbal probability expressions they commonly use. This resulted in a list of seven expressions (see [7] for a discussion about the appropriateness of these seven expressions). In the second study we asked (other) subjects to rank order these expressions. This revealed a quite stable ordering. Distances between the expressions were determined in the third study, where we asked (yet other) subjects to compare the (dis)similarity of all pairs of expressions. The distances were used to project the seven expressions onto a numerical probability scale. We established the following projections: certain 100%, probable 85%, expected 75%, fifty–fifty 50%, uncertain 25%, improbable 15% and impossible 0%. [2] The pair-wise comparisons had however artificially enlarged the distances. For example, 'probable' and 'improbable' in isolation were often rated as completely dissimilar, while obviously they are not the end-points of the whole probability scale. In the final version of our scale we corrected for this trend.

We note that these studies were unlike the experiments most other researchers have done: we did not compose our own list of expressions or use other researchers' lists, because we did not want to force a possibly unnatural vocabulary upon our subjects. In addition, we never asked subjects to directly translate words into numbers or vice versa, because although words and numbers are both overt expressions of an internal construct of uncertainty, solicitation of a numerical expression can have importantly different consequences than solicitation of a verbal expression [15].

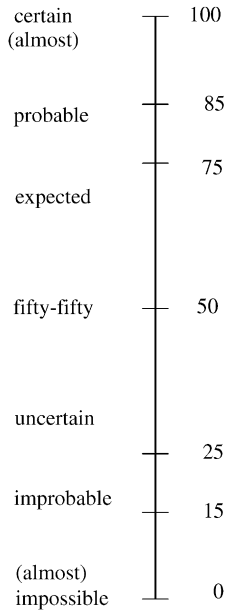
In the fourth study we tested the above projections of the expressions onto the scale, by examining whether subjects' decisions were influenced by the mode, verbal or numerical, in which probability information was presented. Subjects were for example asked whether they would or would not prescribe drug X for a patient with a certain disease, when the probability that the

[2] In Dutch: zeker, waarschijnlijk, te verwachten, fifty–fifty, onzeker, onwaarschijnlijk, onmogelijk.

patient is allergic to that drug is ..., with either a verbal or a numerical expression from the above described set of seven on the dots. Our analyses showed that the decision subjects made depended on the probability used in the description of the decision situations and on its context, but that the decisions were not influenced by the mode in which the probability was presented. We thus found differences between the decision contexts, but not per context between the verbal and the numerical mode, indicating that context-effects influence both the interpretation of the verbal and of the numerical expressions [14].

We concluded that a double scale, with both numerical and verbal labels, could possibly be helpful in situations where judges prefer to communicate probabilities in verbal rather than numerical form but where the elicitor wants numbers as output. Based on the results of these studies we constructed the scale shown in Exhibit 1 to be used with our new elicitation method. It is a continuous scale, to allow subjects to indicate any degree of probability. Since, as argued above, the third study had artificially enlarged differences between verbal expressions, we decided to position the verbal labels closer to the centre of the scale and not right beside the numerical labels. This had the additional benefit that the verbal probability labels would not, incorrectly, be taken to be

Exhibit 1
The response scale with both verbal and numerical labels



exact translations of precise numbers but as a set of labels with a stable rank-ordering, covering the whole probability continuum.

The verbal–numerical probability scale from Exhibit 1 was used in a probability elicitation method for the assessment of the 4000 probabilities required for the construction of a real-life probabilistic network in the domain of oesophageal cancer (for details on the network and its construction we refer to [11]). The events in this domain are precise, and therefore assessment of point probabilities is possible in principle (cf. [13]). Probabilities were elicited from two domain experts. For each probability that had to be assessed, the experts were shown the double scale, together with a transcription of that probability as a fragment of text. By providing a scale for each probability we avoided a spacing effect, that is: people's tendency to evenly or aesthetically distribute different assessments on one scale [12]. Transcriptions and scales pertaining to the probabilities of a single distribution, that is, those that should sum to 100%, were grouped on a single page or on two consecutive pages. By asking the experts to assess a complete distribution at a time we avoided a centering effect of the separate probabilities [12]. We told our experts that, initially, it sufficed to give only rough assessments of these probabilities. Sensitivity analysis methods could subsequently identify those probabilities that highly influenced the output of the network [3], and these probabilities could, if necessary, be refined at a later stage.

With this elicitation method, our experts were able to give their assessments at a rate of 150–175 per hour [10]. To our satisfaction, we were thus able to elicit the 4000 probabilities in reasonable time and without asking too much effort from our experts. Evaluation of the network's diagnostic accuracy against patient data showed that, based on the initial rough assessments, for 85% of the patients the network gave the correct diagnosis. For only 20% of the assessments the experts had marked the anchors on the scale, that is, for 80% of the assessments they had exploited the flexibility provided by the continuous scale.

In an interview set up to evaluate our experts' use of the described elicitation method, they indicated that they had found it most effective and quite easy to use. More specifically, they said that they had found the presence of both numerical and verbal labels next to the scale quite helpful. They had used words as well as numbers when thinking about their assessments, depending on how familiar they had felt with the situation to be assessed. The more uncertain they had felt, the more they had been inclined to think in verbal terms. The experts also said that they felt comfortable with the specific expressions presented. They did indicate that the expression 'impossible' is hardly ever used in oncology, especially when communicating with patients; they preferred to use 'improbable' to refer to almost impossible events. The fact that the experts' interpretation of the word improbable was lower than the interpretation suggested by our scale did not hamper the scale's usability: the experts would

assess an event as improbable, subsequently indicate that a probability of around 15% was too high, and place a mark lower on the scale. We observed that they almost never used extreme probabilities in their assessments and wondered whether this was an artefact of the scale. However, when we put this observation to our experts, they said they had never felt the need to indicate such extreme assessments.

These initial experiences with the use of our verbal–numerical scale as part of a probability elicitation method were very encouraging, but they only involved a single domain and two experts. To assess the general applicability of the scale, we decided that further study was required.

3. Preliminary considerations

The aim of the study described in this paper is to assess the general usability of our scale. Research has shown that it hardly makes any differences in the assessments which method is used to elicit probability judgements [4] and we had found that adding verbal labels to a numerical scale facilitated the assessment process for our experts in our domain. Would this be so for other judges and in other domains as well?

In a review of human probability processing, based on an abundance of literature on the subject, Budescu and Wallsten [1,13] address issues to consider in the communication of uncertainty. They observe that one of the solutions to potential communication problems caused by the use of verbal expressions of uncertainty is to standardise the language by using a verbal scale with a small subset of rank-ordered, frequently used terms and associating a range of probabilities with each term. Such a standardised scale would, they continue, be feasible if people can suspend or suppress the meanings they normally associate with these terms, and if the interpretation of the terms is independent of context.

The solution proposed by Budescu and Wallsten not only standardises the verbal expressions used, but also the interpretation of these expressions in terms of a fixed range of probabilities. They therefore actually do not propose using a *scale* but rather a list of non-overlapping categories. The verbal–numerical scale we developed does offer a small subset of rank-ordered, frequently used terms. It is, moreover, an actual (continuous) scale that only *suggests* an interpretation for each verbal expression, but does not enforce the interpretation to be within a predefined range. It is therefore no problem if the interpretation of an expression depends on context, and people do not have to suspend or suppress the meaning they associate with the terms. In addition we claim that our scale, with both words and numbers, minimises the between-subject variability found in the interpretation of both verbal and numerical probabilities (see e.g., [2,9]).

We found that a verbal–numerical probability response scale allowed our experts to assess reliable probabilities efficiently and comfortably. Informal support for our scale also came from students in a course on medical decision making, who compared the elicitation of probabilities with our scale and with lotteries. They found that the lotteries were complicated, cumbersome and difficult to understand, while the double scale was judged quite easy to use and enabled faster assessments than the time-consuming lotteries. Artificial Intelligence students also tested our scale, this time on expert chess players. They compared our scale to three other versions: with numbers only, words only and with no labels. They found no differences between the experts' assessments of their chance of winning given a certain chess position, no differences in the confidence in their assessments, nor in the time it had taken them or in reported ease of use.

These initial experiences were quite promising, but obviously provided insufficient guarantee that the scale would indeed facilitate the process of probability assessment in other domains and for other judges. To justify the more general promotion of our scale, we needed to answer some further questions: Do people in general find the scale comfortable? Is it a good alternative to a numerical scale, the most widely used response scale? Does it elicit accurate assessments? We addressed these questions in the study reported in this paper.

4. Ratifying our scale

This section describes the study we set up to put our verbal–numerical probability assessment scale to test. Subjects were given a set of questions describing well-defined probability events, such as at least two people in a group of 20 having their birthday on the same day. Although the task was intended to be an estimation task rather than a calculation task, we used this type of question to be able to determine the accuracy of subjects' assessments by comparing them to the correct answers. See Appendix A for the list of questions used.

We checked the accuracy of the probability judgements elicited with two different scales. The first scale was our double scale and the second scale had numerical labels only, which were the same as the numerical labels on our double scale. We predicted that the judgments given on the double scale would be as correct as the judgements on the single scale. We included two groups of subjects, arts students and mathematics students, which allowed us to control for a possible bias towards verbal or numerical expressions by language versus number oriented subjects. By making the questions quite difficult, and by setting a time limit, we practically excluded the possibility of calculation. We also compared the subjects' certainty, which they stated with each assessment

in answer to the question how certain they were that that assessment was correct. We expected subjects who used the double scale to be more certain about their assessments, because they had been allowed to choose their preferred mode of expression. We also looked at how comfortable subjects felt using the two scales. We expected that the double scale would be at least as easy, and generally easier, to use than the numerical scale, because it gave subjects more support. This was determined by comparing answers to pertinent questions (see below).

4.1. Subjects

There were 29 arts students, 10 male and 19 female, who filled in the questionnaire with the double scale (group A1) and 29 arts students, 9 male and 20 female, who used the scale with only numerical labels (group A2). There were 22 mathematics students, 18 male and 4 female, who used the double scale (group B1) and 27 mathematics students, 23 male and 4 female, who used the numerical scale (group B2). The mean age of all 107 subjects together was 22 years, ranging from 17 to 43.

4.2. Procedure

We used the eight probability questions from Appendix A. Next to each question the vertical response scale was depicted; for half the subjects (groups A1 and B1) this response scale was our double scale with words to the left and numbers to the right of the anchors, for the other half (groups A2 and B2) the scale had only numbers positioned to the right. Underneath each problem was the question: How certain are you that your answer is correct? to be answered on a five-point scale labelled 'certain' and 'uncertain' at the extremes. To control for order-effects, we had two versions of the list of questions, one with the questions in the order as given in Appendix A, and one with the questions in reverse order.

On the cover page, subjects were instructed to use their numerical intuition to estimate the answers, not to think too long about their answers and certainly not to calculate, and to finish all questions within 5 min. Subjects were also asked for their name (optional), age, gender and discipline. On the last page, we included four questions to get our subjects' opinion about the scale: How difficult did you find the questions? (with answers on a five-point scale from 'very difficult' = 1 to 'very easy' = 5); Did you calculate at all? (with answers from 'not at all' = 1 to 'all the time' = 5); To what extent was the scale a support in giving your estimations? ('not at all' = 1 to 'perfectly' = 5) and To what extent did the scale enable you to express your assessment correctly? ('not at all' = 1 to 'perfectly' = 5).

4.3. Data analysis

We scored the probability and certainty assessments of all questions separately. When, for their probability assessments, subjects had marked the scale itself, we rounded off our measurement of their assessment to the nearest multiple of 5, thereby taking into account that the scale only allows for rough assessments. When subjects had circled a word, we took the closest multiple of 5 of the corresponding (virtual) number on the scale (e.g., 'uncertain' = 30). For each subject and each question we computed the difference between her/his answer and the correct answer rounded off to a whole percent. We then performed a multivariate analysis of variance (MANOVA) to establish whether background (arts versus mathematics) and type of scale influenced the assessments for the eight questions.

For each subject we computed the mean certainty about her/his assessment, which ranged from 1 = 'uncertain' to 5 = 'certain'. For the four evaluation questions we also established the mean values. With MANOVA we tested the influence of background and type of scale on the different scores.

4.4. Results

We found no effects of the different order of the questions on the assessments or any of the scores, so we report our results over the four groups without making a distinction between subjects who answered the questions in the different orders.

4.4.1. Probability assessments

The correct answers and the mean assessments of the four groups are presented in Exhibit 2. The mean errors, that is: differences between the probability assessments and the correct answers, were −2.9 ($SD = 8.16$) for group A1, −6.8 ($SD = 9.34$) for group A2, −4.1 ($SD = 9.52$) for group B1 and −6.8

Exhibit 2
Correct probabilities and mean answers (plus SD) for each of the eight questions and for each group

Question	Correct answer	Group A1	Group A2	Group B1	Group B2
1	33	45 (21)	47 (23)	46 (17)	41 (18)
2	14	21 (23)	17 (18)	28 (18)	24 (22)
3	92	73 (16)	69 (18)	74 (19)	71 (17)
4	39	48 (15)	44 (18)	50 (16)	48 (18)
5	80	82 (12)	73 (15)	71 (16)	70 (17)
6	4	21 (18)	19 (17)	8 (7)	9 (12)
7	92	61 (20)	54 (22)	61 (27)	54 (26)
8	41	18 (19)	16 (16)	26 (21)	25 (26)

($SD = 6.82$) for group B2. These negative mean errors indicate that subjects on average assessed lower probabilities than was correct. From Exhibit 2 we see that subjects strongly underestimated high probabilities and slightly overestimated low probabilities.

The MANOVA showed that over all questions there was a significant main effect of background ($F(8,93) = 4.160$, $p = 0.000$), but not of the type of scale ($F(8,93) = 0.678$). An interaction effect of background and scale was also absent ($F(8,93) = 0.222$). Univariate analyses showed that the overall main effect of background was caused by significant differences on questions 5 ($F(1,100) = 6.800$, $p = 0.011$), 6 ($F(1,100) = 19.417$, $p = 0.000$) and 8 ($F(1,100) = 4.969$, $p = 0.028$).

4.4.2. Certainty

The mean certainty about the assessments was 3.45 ($SD = 0.54$) for group A1, 2.92 ($SD = 0.67$) for group A2, 3.69 ($SD = 0.61$) for group B1 and 3.33 ($SD = 0.81$) for group B2. The MANOVA showed that over all questions there was a significant main effect of background ($F(8,91) = 3.500$, $p = 0.001$), with the mathematics students being significantly more certain, and of scale ($F(8,91) = 2.098$, $p = 0.044$), with the groups using the double scale being significantly more certain. We found no interaction effect of background and scale ($F(8,91) = 0.846$). The certainties differed only slightly per question, and were not related to the correctness of the probability assessments.

4.4.3. Evaluation of the scale

In Exhibit 3 we present the means of the scores, per group, on the four evaluation questions. The MANOVA revealed significant differences between the four groups on the first question, about the ease of the problems ($F(3,102) = 7.232$, $p = 0.000$), and on the third question, whether the scale had helped assessment ($F(3,102) = 6.890$, $p = 0.000$).

We found that both groups A1 and B1, who had been presented with the double scale, together found the problems much easier ($F(1,104) = 7.125$, $p = 0.009$) and had appreciated the support given by the scale more ($F(1,104) = 5.055$, $p = 0.027$). Looking at the arts groups together (A1 and

Exhibit 3
Means (and SD) of answers to evaluation questions for all groups, on a 1–5 scale with 1 = 'not at all' and 5 = 'very much'

Group	Easy?	Calculated?	Support?	Enables expression?
A1	2.5 (0.2)	2.2 (0.2)	2.7 (0.2)	2.9 (0.2)
A2	1.8 (0.2)	2.1 (0.2)	2.2 (0.2)	2.9 (0.2)
B1	2.8 (0.2)	2.5 (0.2)	1.8 (0.2)	2.7 (0.3)
B2	2.6 (0.2)	2.8 (0.2)	1.4 (0.2)	2.6 (0.2)

A2) versus the mathematics groups (B1 and B2) we saw that the arts students had found the questions significantly more difficult ($F(1,104) = 10.529$, $p = 0.002$) and had found the scale significantly more of a support ($F(1,104) = 15.131$, $p = 0.000$), and also that they had calculated significantly less ($F(1,104) = 6.906$, $p = 0.010$). The question whether the scale allowed correct assessments was not answered significantly differently by the four groups of subjects.

4.5. Conclusion

We had expected that probability assessments on the double scale would be as correct as assessments on the scale with only numerical expressions. Our results confirmed this hypothesis; the probability assessments were not influenced by the type of scale subjects had been presented with to indicate their assessments. In Exhibit 2 we do observe a centering effect. This effect is caused by the fact that in this study we only asked for single probabilities and not, as with our two experts, for complete distributions.

Subjects who had been presented with the double scale were, as expected, more certain of their assessments. These subjects had found the problems easier to answer than the subjects who had only been given numbers, and they had appreciated the scale as significantly more supportive. The arts students differed from the mathematics students; they had, as was to be expected, found the questions more difficult. Some said that they would not be able to give correct answers because they had not been trained in mathematics. Understandably then, they had been less certain, had made fewer calculations and they had appreciated the support given by the scale more than the mathematics students.

Taken together, this study supports our hypothesis that the double scale does not hamper probability assessment, on the contrary: it leads to accurate assessments and it facilitates the process for the numerically less literate.

4.6. Supplement

We performed a small additional experiment to test whether the assessments we elicited with the double scale were stable, that is, remained constant over time. We presented 21 Information Science students with the questionnaire that had been used before, but this time only the version with the verbal–numerical response scale was used. The same group of students answered the same questions twice with a two-week interval. At the end of the first session they were told that we would come back "with a similar questionnaire" in two weeks; in fact, it was the exact same questionnaire. It was unlikely that the subjects would remember their answers after such an interval, and they were not given feedback about the correctness of their answers.

Spearman's correlations between the assessments for the two sessions were computed for each subject individually and ranged between 0.361 and 0.994. They were significant, two-tailed, for 14 of our 21 subjects (at $p = 0.01$ for 5 and at $p = 0.05$ for 9 subjects). The correlation over all subjects was significant: 0.752, $p = 0.00$. The probability assessments of subjects thus remained constant over time when the verbal–numerical response scale was used.

5. General discussion

Our study confirmed our initial experiences and showed that presenting subjects with a response scale that includes both verbal and numerical labels for their probability assessments facilitates the assessment process. The accuracy of the assessments with the double scale is comparable to that of assessments with a numerical scale and people find the double scale more comfortable to use. Results indicate that assessments with the double scale remain stable over time, implying that the verbal labels do not cause random variation in the assessments. In assessing a probability for the same event twice, people apparently use the same label as anchor. We thus think that this scale is of great help to elicitors and experts who are co-operating in specifying probabilities, especially if large numbers of probabilities are to be assessed and inaccuracy is not a big issue. Although our verbal–numerical scale makes it difficult to express fine-grained probabilities, for most purposes a coarser assessment suffices, and differences between a probability of for example 0.15 and 0.17 are irrelevant.

Stating probabilistic information may be a daunting task for experts when the questions are presented in a format that makes great demands on their cognitive processes. When their response mode preferences are taken into account, as we did by presenting them with the opportunity to choose whether to state their probabilities verbally or numerically, the task becomes feasible, and the possibility of building real-life decision support systems based on probabilistic networks becomes more realistic. We therefore feel justified in advocating the more widespread use of our verbal–numerical scale.

We would like to note that we might have found more pronounced differences between a numbers only and a double scale had we used different questions, in which no numbers figured. We could for example have asked: 'Recognising that it is nowadays to be expected that trains are delayed, what is your chance of arriving in time at your meeting in another city if you take the regular 8.15 a.m. train?' We had to decide not to use such questions, because we would have been unable to establish the correct answer to compare our subjects' assessments to. However, now that we have seen that the assessments are equally correct with both types of scale, we may set up a follow-up study in which such questions are used.

Finally, we do not pretend to have developed *the* verbal–numerical response scale. Indeed, different groups of subjects, from different professional domains, may prefer other words than the ones we have used. If so, the words that are used in a specific domain can replace the words we selected for the scale. The actual numerical interpretation of the words, which may vary per context, is of less importance than familiarity with terms, as the continuous scale allows the assessor to correct for effects of variable interpretation. Further research is foreseen to investigate this claim.

We conclude that we have shown that a continuous response scale that combines verbal and numerical labels close by or next to the anchors is not only possible, but is indeed very helpful to assessors who face the task of assessing a large number of probabilities.

Acknowledgements

We would very much like to thank Karl Harvor Teigen for his extensive and useful comments on earlier drafts of this paper, Pieter Koele for his statistical advice, and our students for gathering data.

Appendix A. Probability questions

1. Consider a vase with fifty balls, thirty of which are colored red and the remaining twenty green. You randomly draw five balls, without replacement. What is your estimation of the chance of drawing more than three red balls?
2. In a group of three hundred students, 136 study French and 122 study Spanish; 65 students study both French and Spanish. A student is randomly selected from this group. What is your estimation of the chance that this student studies neither Spanish nor French?
3. In a court of law, a suspect is convicted when found guilty by the judge. If a suspect is guilty, then the judge will indeed convict the suspect in 95% of the cases; if the suspect is innocent, the judge will find him indeed innocent in 80% of the cases. Suppose that 70% of all suspects in a court of law are guilty. What is your estimation of the chance that a convicted suspect is indeed guilty?
4. Consider a bag with twelve coins. Five of the coins are fair, four coins have been manipulated such that the chance of tossing heads is only 30%, and 3 coins are two-headed. You randomly draw a coin from the bag and toss it twice. What is your estimation of the chance that both tosses result in heads?
5. Three students Wout, Piet and Bas are the only participants in a swimming contest. Wout and Piet have equal chances of winning and both are twice as

likely to win as Bas. What is your estimation of the chance that Wout or Piet wins?

6. A factory has three production belts on which micro-chips are produced. Of the chips produced on belt A, 4% is defective; for belt B this percentage is 5% and for belt C 1%. Half of all chips manufactured by the factory are produced on belt A, 30% on belt B and 20% on belt C. A random chip is selected from the factory's production. What is your estimation of the chance that this chip is defective?
7. Due to delays, the train from Groningen will arrive sometime between 8:00 and 8:30 (on a whole minute) in Utrecht; the train from Maastricht will arrive sometime between 8:15 and 9:00 (on a whole minute). What is your estimation of the chance that the train from Groningen will arrive before the train from Maastricht?
8. What is your estimation of the chance that in a class of twenty students, two of them have their birthdays on the same day?

References

[1] D.V. Budescu, T.S. Wallsten, Processing linguistic probabilities: general principles and empirical evidence, The Psychology of Learning and Motivation 32 (1995) 275–318.
[2] W. Brun, K.H. Teigen, Verbal probabilities: ambiguous, context-dependent, or both?, Organizational Behavior and Human Decision Processes 41 (1988) 390–404.
[3] V.M.H. Coupé, L.C. Van der Gaag, J.D.F. Habbema, Sensitivity analysis: an aid for belief-network quantification, Knowledge Engineering Review 15 (2000) 1–18.
[4] P. Koele, Subjectieve waarschijnlijkheid (Subjective probability), in: P. Koele, J. van der Pligt (Eds.), Beslissen en Beoordelen (Decision Making and Judgment), 1993, pp. 25–47.
[5] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers, Palo Alto, 1988.
[6] S. Renooij, Probability elicitation for belief networks: issues to consider, Knowledge Engineering Review 16 (3) (2001) 255–269.
[7] S. Renooij, C.L.M. Witteman, Talking probabilities: communicating probabilistic information with words and numbers, International Journal of Approximate Reasoning 22 (1999) 169–194.
[8] C.S. Spetzler, C.A.S. Stael von Holstein, Probability encoding in decision analysis, Management Science 22 (1975) 340–358.
[9] K.H. Teigen, W. Brun, Ambiguous probabilities: when does $p = 0.3$ reflect a possibility, and when does it express a doubt?, Journal of Behavioral Decision Making 13 (2000) 345–362.
[10] L.C. Van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, B.G. Taal, How to elicit many probabilities, in: K.B. Laskey, H. Prade (Eds.), Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, 1999, pp. 647–654.
[11] L.C. Van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, B.G. Taal, Probabilities for a probabilistic network: a case study in oesophageal cancer, Artificial Intelligence in Medicine 25 (2) (2002) 123–148.
[12] D. Von Winterfeldt, W. Edwards, Decision Analysis and Behavioral Research, Cambridge University Press, Cambridge, UK, 1986.
[13] T.S. Wallsten, D.V. Budescu, A review of human linguistic probability processing: general principles and empirical evidence, The Knowledge Engineering Review 10 (1) (1995) 43–62.

[14] P.D. Windschitl, E.U. Weber, The interpretation of "likely" depends on the context, but "70%" is 70%—right? The influence of associative processes on perceived certainty, Journal of Experimental Psychology: Learning, Memory and Cognition 25 (6) (1999) 1514–1533.
[15] P.D. Windschitl, G.L. Wells, Measuring psychological uncertainty: verbal versus numeric methods, Journal of Experimental Psychology: Applied 2 (4) (1996) 343–364.