Behavior Research Methods, Instruments, & Computers
1993, 25 (4), 425-433

METHODS & DESIGNS

A blueprint of ELI: A new method for eliciting subjective probability distributions

JELLE VAN LENTHE
University of Groningen, Groningen, The Netherlands

A blueprint of a new method for eliciting uncertain knowledge about continuous quantities is presented. The direct realization of a proper scoring rule in a graphically oriented interactive computer program is one of the central features of the new elicitation methodology. Uncertain knowledge is internally represented through subjective probability distributions. However, in its interaction with assessors, the elicitation method uses a score representation. A proper scoring rule is applied to transform probability density functions into score functions. In order to study its merits, central ideas for the new method were implemented in an experimental version of the elicitation technique ELI. The results were promising and encouraged further development of the technique.

Decision and risk analysts often need information about certain quantities to perform their analyses. When there are insufficient data to determine the quantities of interest objectively, they often call on human subjects as a source of information. Unfortunately, human knowledge normally is uncertain and more qualitative than quantitative in character. The analysts, on the other hand, usually prefer their input data in a quantitative mode. For this reason, the formalization of uncertain knowledge is an important topic in decision theory, in risk analysis, and in Bayesian statistics. Most research deals with uncertainty related to a particular event (e.g., the likelihood that X will be elected president). Less attention has been paid to the subject of the present paper, uncertainty related to a continuous quantity (e.g., the proportion of the electorate that will vote for Y as president). Probability distributions are one of the most commonly used representations of uncertain knowledge about a continuous quantity. The mental transformation of subjective knowledge in a probability distribution appears to be a difficult task, and several elicitation techniques that support the specification of subjective probability distributions have been proposed. From the discussion below, it will appear that the quality of the elicitation techniques is often disappointing.

The purpose of this paper is to develop a new method for eliciting uncertain knowledge that will meet the fundamental quality requirements to a more satisfactory degree. The next section of the paper provides a discussion of subjective probability distributions. It is followed by a review of the quality of elicitation techniques, and the conclusion will be that a new elicitation methodology is needed. The subsequent section is devoted to the development of a blueprint of a new elicitation method. Proper scoring rules play a central role in the blueprint. In the following section, the choice for a particular proper scoring rule is discussed. Next, it will be demonstrated how the blueprint is realized in a first experimental version of the elicitation technique ELI.

Subjective Probability Distributions

Uncertain knowledge about continuous quantities is usually represented through a probability distribution. The expression *subjective probability distribution* (SPD) is used to emphasize that the distribution reflects the subjective beliefs of a human subject. Figure 1 presents two SPD examples in a probability density mode: a discrete SPD A and a continuous SPD B. The top of the density function reflects the best guess, and the dispersion of the function corresponds with the uncertainty about the best guess. So, the steep SPD A represents knowledge that is relatively certain, whereas the flat SPD B represents knowledge that is much more uncertain. Throughout this paper, especially in the illustrative examples, a *proportion* or *percentage* will be the quantity of interest.

For decision or risk analysts, it is important to consider the quality of SPDs because it determines to a large extent the value of their analyses. One could argue that SPDs are merely formal expressions of what an assessor thinks or knows and that SPDs cannot be judged as right or wrong. Wallsten and Budescu (1983) demonstrated that

The research is supported by PSYCHON Grant 560-266-014 of the Netherlands Organization of Scientific Research (NWO). The author wishes to thank Ivo W. Molenaar, Willem K. B. Hofstee, and three anonymous reviewers for their extensive and helpful comments and suggestions for improving previous versions of the paper. Address requests for reprints to J. van Lenthe, Center for Energy and Environmental Studies (IVEM), University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands.

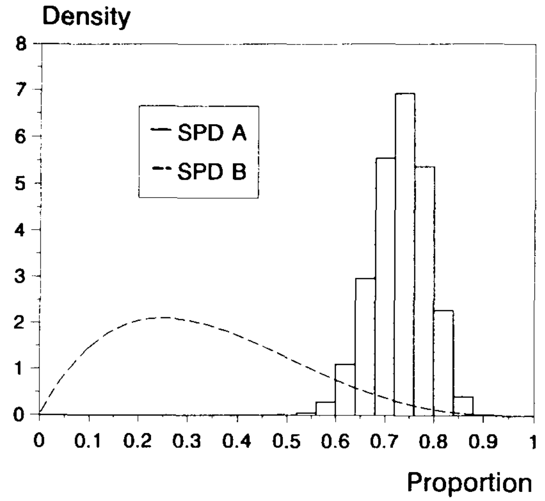


Figure 1. A steep SPD A representing subjective knowledge about a proportion that is quite certain and a flat SPD B representing knowledge that is much more uncertain.

SPD assessment is much like other forms of psychological measurement, and they claimed that SPDs should be judged on the criteria of reliability and validity. The reliability of SPDs can be determined by assessing their stability in time and their consistency across methods. Strictly speaking, an SPD is valid if it accurately reflects the uncertain knowledge of a person. It is difficult to examine this *internal* validity. From a pragmatic point of view, it is also important that SPDs correspond to events in the external world. This *external* validity of SPDs can be determined afterwards, when actual values are available.

From previous research it is known that, in general, the quality of SPDs is poor. There has been relatively little research dealing with the problem of reliability. The few available data suggest that SPDs are only moderately reliable (Lourens, 1984; Terlouw, 1989; Wallsten & Budescu, 1983). Much more is known about the external validity of SPDs. Numerous studies have shown that assessors display a systematic *overconfidence* bias. Their SPDs tend to be too tight, and an unduly large percentage of actual values fall into the extreme tails. For an excellent review, see Lichtenstein, Fischhoff, and Phillips (1982). Often the surprise index—that is, the percentage of actual values falling outside the 98% credibility intervals of the SPDs—is used to examine overconfidence. This surprise index should be 2%, but frequently values as high as 30% or 40% are observed.

Elicitation Techniques

The assessment of SPDs appears to be a demanding task for both statistically naive assessors and statistically expert assessors. Therefore, several techniques that support the specification of SPDs have been suggested. These elicitation techniques range from methods that directly ask for certain distribution characteristics to indirect procedures with a less clear relationship between response and resulting distribution. Direct elicitation techniques typically ask for probabilities or values. The probability-oriented methods require assessors to assign probabilities to fixed values of the continuous quantity. For example, after subdividing the range of possible values, assessors are requested to assign their subjective probabilities to the intervals. This histogram method yields discrete density functions like, for example, that depicted for SPD A in Figure 1. Value-oriented methods require the subject to give values for fixed probabilities. The percentile method is such a value-oriented method. It asks for specific percentile points, usually the 1st, 5th, 25th, 50th, 75th, 95th, and 99th percentile. With indirect techniques, assessors are usually requested to consider alternatives. For example, they are asked to choose between bets or are required to make paired comparisons. More complete classifications, together with detailed descriptions of the elicitation techniques, are given by, among others, Schütt (1981), Spetzler and Staël von Holstein (1975), and Van Steen and Oortman Gerlings (1988).

Reviews of evaluation and comparison studies reveal that different techniques elicit different distributions (Lichtenstein et al., 1982; Ludke, Stauss, & Gustafson, 1977; Van Steen & Oortman Gerlings, 1988; Von Winterfeldt & Edwards, 1986). Given this state of the art, which technique and which distributions should be preferred? Schütt (1981) explored the practical usefulness of a large number of elicitation techniques. Generally, direct techniques seem to be more efficient than indirect techniques, because the procedures are rather straightforward and inexpensive and do not require much time. Unfortunately, the straightforward direct procedures do not automatically constitute assessment tasks that are easy for the assessors. On the contrary, direct techniques, which usually require values or probabilities as answers, seem to be difficult for statistically naive assessors. On the average, indirect techniques are less efficient but, at the same time, require hardly any statistical knowledge (usually, assessors are asked to consider alternatives).

The quality of resulting SPDs is another, and probably more important, requirement for elicitation techniques. From the recurrent observation of poor SPD quality, one might conclude that most elicitation techniques apparently are inadequate. Surprisingly little attention is paid to the possibility of method-induced bias—that is, that poor SPD quality originates from the particular elicitation technique used (Fischer, 1982; Hogarth, 1980; Lourens, 1984). More often, cognitive-induced biases are held to be responsible. For example, Koriat, Lichtenstein, and Fischhoff (1980) showed that overconfidence might originate from the tendency to selectively focus on evidence supporting a best guess and disregard evidence contradicting it. Others have demonstrated item-induced biases and shown, for example, that general knowledge questions frequently produce overconfidence (Ronis & Yates, 1987; Wright & Wisudha, 1982). In our opinion, however, the question of method-induced biases is basic. Only after a thorough examination of method-induced biases is it possible to draw unambiguous conclusions about cognitive- and item-induced biases.

As far as the relative merits of elicitation techniques are concerned, results have been contradictory (Lichtenstein et al., 1982; Seaver, Von Winterfeldt, & Edwards, 1978; Van Steen & Oortman Gerlings, 1988; Von Winterfeldt & Edwards, 1986). Each technique has problems of its own, and there is apparently no elicitation technique that meets to a sufficient degree the fundamental requirements of reliability and validity of assessments and practical usefulness of the technique. There exists overwhelming evidence that, in particular, the elicitation technique most commonly used with continuous quantities, the fractile method, yields overconfident assessments (Alpert & Raiffa, 1982; Lichtenstein et al., 1982; Pickhardt & Wallace, 1974). The SPDs over the continuous quantities are far too tight.

For the present, the question of which elicitation technique should be used remains unanswered. There apparently exists no elicitation technique that results in sufficiently reliable and valid SPDs. In our opinion, poor SPD quality is at least in part an artifact of the particular techniques used; equipped with a more appropriate elicitation tool, assessors might prove to be more capable probability estimators than has been suggested by research thus far. For this reason, a new method for eliciting uncertain knowledge will be proposed.

Blueprint For A New Elicitation Method

The research reported here was motivated by the desire to devise an innovative elicitation method that would contribute to the assessment of reliable and valid SPDs and that would meet the requirement of practical usefulness. Several recommendations concerning appropriate techniques guided the search for a new elicitation methodology (Hogarth, 1975, 1980; Huber, 1974; Lourens, 1984; Staël von Holstein, 1970a; Terlouw, 1989). These recommendations can be summarized in two important guidelines. First, the new method should be efficient and acceptable for assessors with different (statistical) backgrounds. Second, proper scoring rules should play a regular and central role.

Cognitive studies suggest that the human being is a selective and stepwise information-processing system with limited capacity. An appropriate elicitation method therefore should place a minimum of information-processing demands on assessors' cognitive resources (Hogarth, 1975). It should allow these resources to be directed to the task of estimating the uncertain quantity. A new elicitation method ideally should combine the positive aspects of both direct and indirect elicitation procedures. It should be straightforward, inexpensive, and fast and, at the same time, it should be easy to handle and not difficult to learn. So, preferably the new technique should require assessors to consider alternatives rather than to specify probabilities or values.

In the early stages of the development, it was considered to provide assessors with several probability distributions and ask them to select the distribution corresponding most closely with their subjective knowledge. To keep the assessment task as simple as possible, graphical displays of probability distributions were used. A computerized procedure was considered an effective means for displaying the graphs and for providing assessors with an appropriate selection mechanism. As in Bayesian statistics, a particular *natural conjugate* family of probability distributions (Novick & Jackson, 1974) was chosen for representing uncertain knowledge about a particular uncertain quantity. The beta distribution, for example, is the natural conjugate distribution for proportions. Empirical research revealed that this family is sufficiently rich and flexible for representing uncertain knowledge about a proportion (Terlouw, 1989).

A disadvantage of the method described thus far is that it still uses the concept of probability distributions. Psychological studies of judgmental processes reveal that human subjects have several shortcomings in acting as *intuitive statisticians* (Hogarth, 1980; Kahneman, Slovic, & Tversky, 1982). So, in its interaction with assessors, the new elicitation method preferably should not use statistical concepts or methods. With the discussion of the second guideline, the central role of proper scoring rules, it will become clear how this suggestion was realized.

Scoring rules involve the computation of a score based on the relation between the stated SPD and the value that actually occurs. Scoring rules are important SPD assessment tools for several reasons (Murphy & Winkler, 1970; Staël von Holstein, 1970a, 1970b; Van Naerssen, 1962; Winkler, 1971, 1986). First, after an SPD is assessed and the uncertain quantity of interest is observed, scoring rules can be used to evaluate the accuracy of the SPD in terms of an association between the stated SPD and the value that actually occurs. Second, these accuracy scores can be useful in a training situation for giving accuracy feedback. Thus far, however, empirical evidence concerning the effects of scoring-rule feedback is scarce and incomplete (Fischer, 1982; Staël von Holstein, 1971, 1972). Third, announcing the use of the class of *proper* (also called *reproducing*) scoring rules might encourage assessors to be honest and careful during the specification of an SPD. During the assessment task—that is, when actual values are unknown—the expected rather than the actual scores are of primary interest. With proper scoring rules, assessors can maximize their expected score if their stated SPD corresponds with their subjective beliefs.

Suppose, uncertain knowledge about a continuous quantity is formalized by asking assessors to assign their subjective probabilities to N mutually exclusive and collectively exhaustive outcomes (the histogram method); r_i denotes the assigned probability, and p_i denotes the true subjective judgment for outcome i. Let $\boldsymbol{r} = (r_1, r_2, \ldots, r_N)$, and let $\boldsymbol{p} = (p_1, p_2, \ldots, p_N)$. The expected score is given by $ES(\boldsymbol{r},\boldsymbol{p}) = \Sigma_i p_i S(\boldsymbol{r},i)$, in which $S(\boldsymbol{r},t)$ denotes the score for a stated $\boldsymbol{r}$ and the actual outcome t. A scoring rule is proper if $ES(\boldsymbol{p},\boldsymbol{p}) \geq ES(\boldsymbol{r},\boldsymbol{p})$ for $\boldsymbol{r} \neq \boldsymbol{p}$. It is said to be strictly proper when the expected score is maximized if and only if $\boldsymbol{r}=\boldsymbol{p}$, that is, $ES(\boldsymbol{p},\boldsymbol{p}) > ES(\boldsymbol{r},\boldsymbol{p})$ for $\boldsymbol{r} \neq \boldsymbol{p}$ (Murphy & Winkler, 1970; Van Naerssen, 1962). In this way, internal validity of assessments might be enhanced because assessors are encouraged to state SPDs that correspond with their subjective judgment.

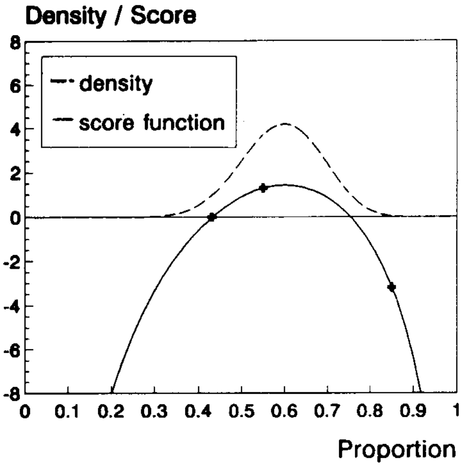


Figure 2. The beta density $f_{p,q}(x) = x^{p-1}(1-x)^{q-1}/B(p,q)$ with $0 \le x \le 1$, $p > 0$, $q > 0$, and beta function $B(p,q)$ for $p = 16$ and $q = 11$, and its corresponding score function $S_L(f_{p,q},t) = Ln(f_{p,q}(t))$.

Studies of proper scoring rules have been mainly theoretical. In the rare occasion of a practical application, they function merely as extra help and mainly for training purposes. In the new method, proper scoring rules should play a more regular and central role. In particular, the elicitation method should exploit the properness characteristic of proper scoring rules. So, the next problem was to give proper scoring rules a central role in the graphically oriented interactive computer program mentioned before. Of course, one could try to instruct subjects to consider a particular scoring rule while selecting an SPD, but then things become very complicated. Let us assume that an assessor is asked to estimate a particular proportion and that he/she answers by selecting the beta density function $f_{p,q}(x)$ of Figure 2. Let us also assume that assessors are asked to consider the strictly proper logarithmic scoring rule $S_L(f_{p,q},t) = ln(f_{p,q}(t))$, in which ln denotes the natural logarithm, for possible actual values t. To appreciate the implications of the SPD in terms of scores, the assessor should (1) grasp the statistics of probability density functions, (2) understand the mathematics of the scoring rule, and (3) apply the scoring rule to the selected SPD to calculate in advance several scores for assumed actual values t. Figure 2 shows logarithmic scores (0.0, 1.3, −3.2) for three assumed actual values (0.43, 0.55, 0.85).

As stated before, an elicitation technique should place a minimum of information-processing demands on the assessors' cognitive resources. So, the new method should take care of necessary calculations. Hofstee (1987) suggested that a computer program would be an ideal means for calculating, in advance, scores, not only for a few assumed actual values, but for all possible actual outcomes t. The computerized technique could display these scores graphically with a curve $S_L(f_{p,q},t)$ (see Figure 2). Unfortunately, now assessors have to consider *two* graphical displays. They have to select a density function that matches their uncertain knowledge and, at the same time, they have to consider a corresponding score function to examine possible consequences of their assessments. On reflection, however, providing the score function might be sufficient because it is merely another representation of uncertain knowledge. Considering the problems that assessors might have interpreting the density functions, it is probably preferable to use the score representation that might be more compatible with their capacities than the usual probability representation. Another advantage of using score functions instead of probability density functions is that the score functions provide *feedforward* information about possible consequences of the assessment. One of the recommendations found in the literature is that assessors should reconsider their assessments (Hogarth, 1975; Lourens, 1984; Terlouw, 1989). The feedforward feature of the score function might stimulate assessors to reflect on the implications of their assessments before making a final choice.

In short, in the new elicitation method, proper scoring rules play the central role of generating score curves from underlying probability density functions. In its interaction with assessors, the new method uses these score curves instead of the probability density functions. Assessors are asked to consider several score curves, and they are provided with a simple selection mechanism to choose the curve that matches their subjective knowledge most closely.

Which Proper Scoring Rule?

Another issue was to choose a proper scoring rule for generating the score curves. Strictly proper scoring rules that are well known and appealing because of their relatively simple structure are the logarithmic, the quadratic, and the spherical scoring rule, which are, respectively,

$$S_L(\mathbf{r},t) = \log(r_t), \qquad (1)$$

$$S_Q(\mathbf{r},t) = \left(1+2r_t-\textstyle\sum_i r_i^2\right)/2, \qquad (2)$$

and

$$S_S(\mathbf{r},t) = r_t/\left(\textstyle\sum_i r_i^2\right)^{0.5}. \qquad (3)$$

If t is the actual value of the interest and $f(x)$ represents the stated SPD, continuous analogues of the logarithmic, quadratic, and spherical scoring rules are (Matheson & Winkler, 1976), respectively,

$$S_L(f,t) = \log f(t), \qquad (4)$$

$$S_Q(f,t) = \left(1+2f(t) - \int_{-\infty}^{\infty} f^2(x)dx\right)/2, \qquad (5)$$

and

$$S_S(f,t) = f(t)/\left(\int_{-\infty}^{\infty} f^2(x)dx\right)^{0.5}. \qquad (6)$$

Results of studies considering the relative merits of proper scoring rules have been inconclusive (e.g., Jensen & Peterson, 1973; Murphy & Winkler, 1970). Meteorologists seem to favor the quadratic scoring rule—or the *Brier* score, as they call it—which can be decomposed into several useful components (Yates, 1982, 1988). Others have argued on both theoretical and empirical grounds that the logarithmic scoring rule is superior (for a short review see, e.g., Staël von Holstein, 1970a, 1970b). Staël von Holstein (1970b, 1977) also introduced the requirement of *sensitivity to distance*—that is, a scoring rule should reward putting density mass near the actual value. The ranked probability score

$$S_R(f,t) = \int_{-\infty}^{t}(F(x))^2dx + \int_{t}^{-\infty}(1-F(x))^2dx, \quad (7)$$

with $F(x) = \int_{-\infty}^{x} f(y)dy$, appears to meet this requirement. Jensen and Peterson (1973) observed that linear transformations of proper scoring rules—the properness characteristic remains with a linear transformation with a positive multiplicative constant—had a larger effect than the particular type of proper scoring rule. For example, scoring rules containing both positive and negative scores appeared to induce suboptimal strategies. Unfortunately, research concerning the relative merits of different scoring rules is almost completely restricted to the case of two possible events, and it appears to be difficult to generalize the results beyond the two-state case (Murphy & Winkler, 1970).

Figures 3A and 3B present a flat and a steep beta density function, respectively. Both SPDs are transformed according to four proper scoring rules: the logarithmic, the quadratic, the spherical, and the ranked probability scoring rule. The first three rules were transformed linearly in a manner that would (1) yield zero scores for a uniform distribution and (2) allow the score curves to be considered on about the same score scale. Inspection of the score curves reveals that none of the scoring rules is sensitive to small deviations of the actual values from the best guess. Important differences arise for the tails of the distributions. The logarithmic scoring rule appears to be especially sensitive in the tail areas. For example, an actual value slightly smaller than the lower bound can result in a highly negative score.

At this stage, it was decided to use the logarithmic rule for generating score curves because it is conceptually as well as computationally the most simple scoring rule. The score for an actual value *t* depends only on the density associated with the actual outcome and, unlike the other above-mentioned proper scoring rule, the logarithmic rule does not require integral calculus. Generating curves with other than logarithmic scoring rules is more laborious, which might be a problem for on-line graphical presentation of the score functions. Besides, the logarithmic transformation of beta density functions provides a convenient operationalization for the concepts of lower and upper bound. The ignorance assessment of the uniform beta density always results in a zero score. So, the zero score points of the score curve, which correspond with density 1, define quite naturally a lower and an upper bound. Actual values outside the interval between lower and upper bound result in scores that are worse than the zero scores of the ignorance strategy of assessing a uniform distribution.

There are also arguments against the use of the logarithmic scoring rule. First, one might argue that its simplicity is also a disadvantage and that an appropriate scoring rule should consider the entire distribution. Second, it might be that the logarithmic scoring rule stimulates assessors to make more flat assessments only because they want to avoid negative scores, in which case, the method would operate like the Alpert and Raiffa (1982) instruction to "spread out those distributions." Fischer (1982) observed that logarithmic scoring rule payoffs reduced only the tendency to use zero probabilities, which in turn led to an improvement on only one dependent criterium variable, the logarithmic score itself. He found no improvement on any other quality criteria (e.g., the spherical scoring

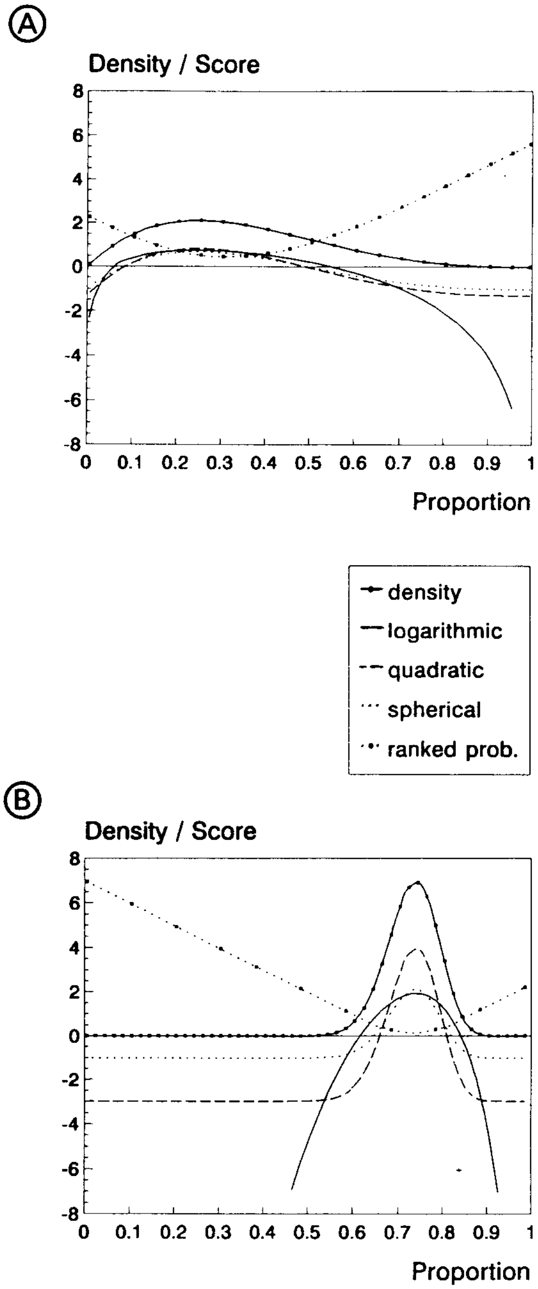


Figure 3. (A) A flat beta density and its transformations according to the ranked probability, the logarithmic, the quadratic, and the spherical scoring rule. (B) A steep beta density and its transformations according to the scoring rules of Figure 3A.

rule). Jensen and Peterson (1973) concluded that scoring rules containing both positive and negative scores might result in suboptimal strategies. They recommended restricting the scoring to all-positive or all-negative scores.

Recent empirical results do not support the ideas concerning potential artificial effects of the logarithmic scoring rule in ELI (Van Lenthe, in press). An experiment, in which 304 subjects participated, was carried out to study the effects of using different scoring rules for generating the score curves. Differences were observed only for proper scoring rules on the one hand and an improper linear scoring rule on the other. The three proper scoring rules (logarithmic, quadratic, spherical) appeared to produce similar external validity scores, and the external validity scores for the improper linear scoring rule were significantly worse. The score curves contained both positive and negative scores. To explore possible artificial effects of a mixed-score range, four additional conditions were constructed by using an additive constant to transform the score curves to the positive domain. No differences were observed between ELI versions with mixed-score curves and ELI versions with positive-score curves. Using proper scoring rules to provide feedforward information appeared to be a promising method for improving the quality of probability assessments. This scoring rule approach appeared to be robust with respect to different types of proper scoring rules and with respect to different score ranges.

The Elicitation Technique ELI

To study their merits, central ideas of the blueprint were implemented in an experimental version of the elicitation technique ELI. This preliminary ELI version was restricted to the estimation of proportions (or percentages). So, the family of beta distributions

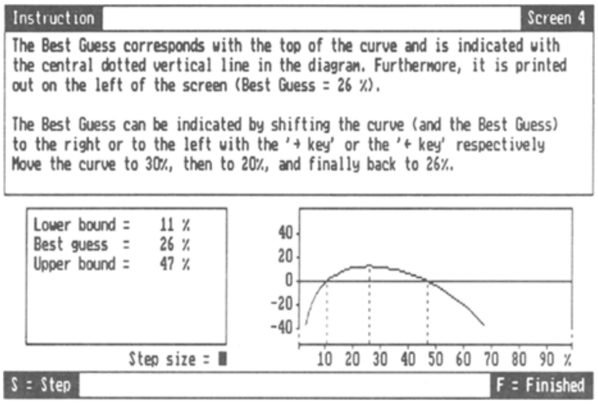
$$f_{p,q}(x) = x^{p-1}(1-x)^{q-1}/B(p,q), \qquad (8)$$

with $0 \leq x \leq 1$, $p > 0$, $q > 0$, and beta function $B(p,q)$, was used for representing uncertain knowledge. A logarithmic scoring rule was implemented to generate score curves from the underlying beta density functions. To prevent practical and motivational difficulties associated with bankruptcies of assessors, the following truncated version of the logarithmic scoring rule was used:

$$S(f_{p,q},t) = 10^{*}\max\{\ln(f_{p,q}(t), -8\}. \qquad (9)$$

Strictly speaking, this truncated version is not quite proper. However, its deviation from properness seems to be very small and next to negligible. A score of -8 corresponds to a density of e^{-8} (=.0003). The contribution of the probability areas with a density of e^{-8} or smaller to the subjectively expected score $ES(\mathbf{r},\mathbf{p})$ appears to be negligibly small—irrespective of whether or not the logarithmic score is truncated. The multiplicative constant of 10 was used to avoid decimals in the scoring.

The direct implementation of the logarithmic scoring rule in a graphically oriented interactive computer program results in a rather easy assessment task. Assessors are not required to key in numbers, and they do not have to bother about probabilities or transformations with scoring rules. The computer program takes care of necessary calculations. The only thing assessors have to do is to select, from a large number of alternatives, the score curve corresponding most closely to their uncertain knowledge.



Instruction
Screen 4

The Best Guess corresponds with the top of the curve and is indicated with the central dotted vertical line in the diagram. Furthermore, it is printed out on the left of the screen (Best Guess = 26 %).

The Best Guess can be indicated by shifting the curve (and the Best Guess) to the right or to the left with the '→ key' or the '← key' respectively Move the curve to 30%, then to 20%, and finally back to 26%.

Lower bound = 11 %
Best guess = 26 %
Upper bound = 47 %

Figure 4. Example of an ELI instruction screen. With the displayed curve, assessors can indicate their best guess as well as their uncertainty.

Each score curve represents a best guess, as well as the uncertainty about the best guess (see Figure 4). The best guess corresponds to the top of the score curve, and the uncertainty is related to the dispersion of the curve. With the preliminary ELI version, assessors can choose from a grid of 99×24 score curves that correspond to a predetermined set of 99×24 beta density functions. The best guess can take on 99 values (.01, .02, . . . , .99); for each best guess, there are 24 degrees of uncertainty. The interactive computer program provides a simple mechanism for selecting a particular curve from the set of 99×24 score curves. The left-arrow and right-arrow keys control the horizontal position of the curve and can be used for choosing a best guess. For example, pressing the left-arrow key erases the current curve and results immediately in a new curve with a best guess that is decreased by .01. Holding the left-arrow key has an animation effect: the curve *walks* to the left.

The uncertainty about the best guess corresponds with the steepness of the curve and is explained to assessors in terms of a lower and an upper bound—that is, what they think is the lowest and the highest possibly correct answer. The up-arrow and down-arrow key control the steepness of the curve. Uncertainty can be increased by using the down-arrow key. The curve then *grows* flatter, and the interval between lower and upper bound becomes larger. Using the up-arrow key reduces the uncertainty of the response. The curve grows more steep, and the distance between the bounds becomes smaller.

In an interactive instruction, assessors are informed about the score interpretation of the curve. It is explained that the curve returns scores for all possible actual values of the percentage. So, assessors know that actual values equal to the lower or upper bound will yield zero scores. They also can infer from the curve that positive scores will be obtained with actual values between lower and up-

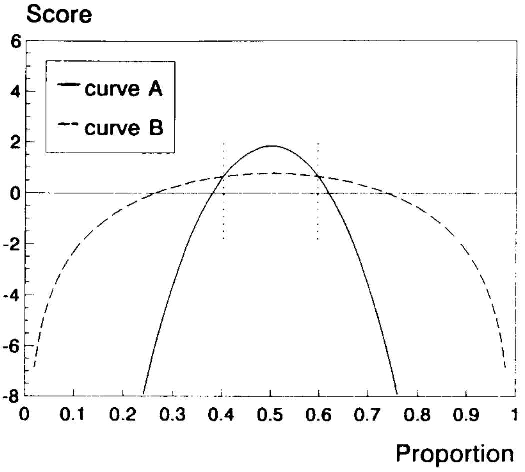
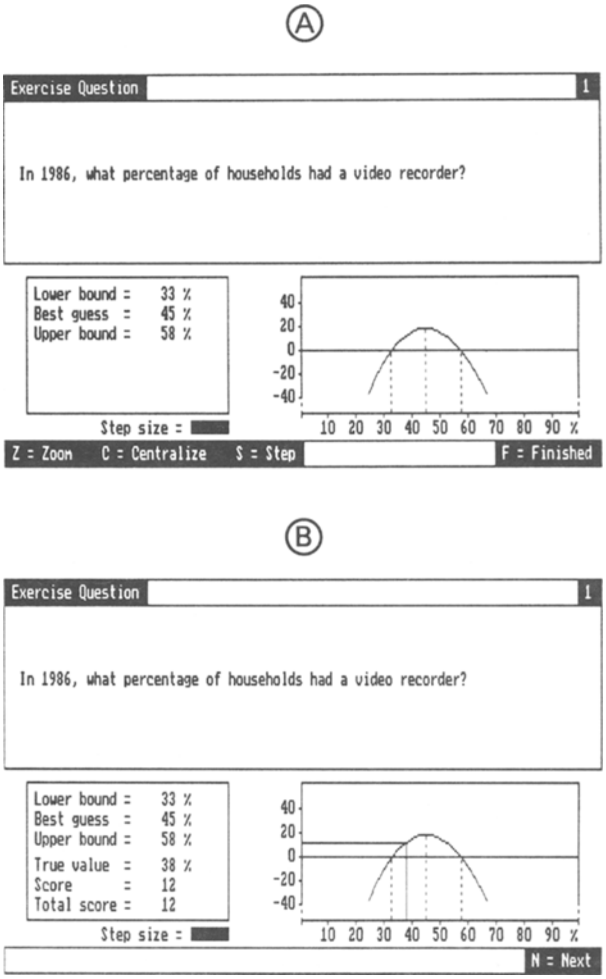


Figure 5. The steep curve A yields higher scores only with actual values within the two points of intersection of the two curves.

per bound and that actual values outside this interval will yield negative scores. Figure 5 presents two score curves: The steep curve A with a small interval between lower and upper bound represents knowledge that is quite certain, and the flat curve B reflects knowledge that is much more uncertain. The maximal attainable score is higher for the steep curve than for the flat curve. But, the steep curve only yields higher scores with actual values within the two points of intersection of the two score curves; the flat curve yields higher scores with actual values outside this interval. To obtain optimal scores, assessors should not only avoid the use of steep curves when they are not sure, but they also should avoid unduly flat curves. Trying to express their actual uncertainty appears to be the optimal strategy for assessors.

An elicitation session typically consists of an interactive instruction, a few practice items, and the questions of interest. In the instruction part, assessors become familiar with the manipulation and interpretation of the score curve. For each question, assessors then have to select a curve that corresponds most closely with their subjective knowledge (Figure 6A). Assessors might be provided with outcome and scoring-rule feedback when the question at stake is a practice item and when the actual value is known (see Figure 6B). The feedback involves the communication of the actual consequence of an assessment. From the curve of possible scores, one is pointed out as the score actually obtained. It is possible to construct a training session in which assessors have to complete a set of practice items and in which they are provided with trial-by-trial outcome and scoring-rule feedback.

An experiment was carried out to evaluate ELI on the criteria of reliability and validity of assessments and practical usefulness of the technique (Van Lenthe, 1993). In this study, ELI performance was compared with the performance of a classical elicitation technique and a simple technique that asked subjects to key in numerical values only for the best guess and the lower and upper bound. The empirical results revealed that ELI support contributed to reliable and externally valid SPDs. ELI also appeared to be an efficient and acceptable method. Compared with the other techniques, the reliability and external validity of ELI SPDs appeared to be superior. ELI also turned out to be the most useful technique. Its cognitive support was rated much higher, and its cognitive load was rated much lower. Two external validity outcomes in particular are noteworthy. First, in correspondence with past research, outcomes with the two existing techniques pointed to a strong overconfidence bias. With ELI, overconfidence appeared to be almost eliminated. Second, ELI support produced the highest accuracy scores. Only with ELI were the mean accuracy scores greater than the mean accuracy score of a hypothetical subject applying the ignorance strategy of assessing uniform distributions all the time. As a matter of fact, performance with the other techniques appeared to be much worse than the performance of this hypothetical subject. This observation resembles results of past research



In 1986, what percentage of households had a video recorder?

Lower bound = 33 %
Best guess = 45 %
Upper bound = 58 %

Step size =

In 1986, what percentage of households had a video recorder?

Lower bound = 33 %
Best guess = 45 %
Upper bound = 58 %
True value = 38 %
Score = 12
Total score = 12

Step size =

Figure 6. (A) Example of an ELI training question. The curve provides scores for all possible values of the actual value and can be manipulated with the cursor keys. (B) Example of trial-by-trial outcome and scoring-rule feedback with a training question.

(Fischer, 1982; Lourens, 1984; Schaefer, 1976; Staël von Holstein, 1970a; Winkler, 1971; Yates et al., 1989).

Conclusion and Discussion

The development of a blueprint for a completely new way of eliciting uncertain knowledge was motivated by the notion that the recurrent observation of poor SPD quality might be attributed to the poor quality of existing elicitation techniques. It was anticipated that, equipped with a more appropriate elicitation tool, assessors might prove to be more competent probability assessors than has been suggested by research thus far. In the search for a new elicitation methodology, uncertain knowledge was represented through specific natural conjugate probability distributions. In its interaction with assessors, however, the method uses a score representation instead of a probability representation. A strictly proper scoring rule was applied to transform probability density functions into score functions. An interactive computer program was considered an adequate device for showing graphical displays of alternative score functions and for providing assessors with a selection mechanism to choose the most appropriate one.

There are several reasons for expecting that the ELI procedure, with feedforward based on proper scoring rules, will contribute to improved SPDs. First, the score curves provide an alternative score representation of uncertain knowledge that might be more compatible with the capacities of assessors than is the usual probability representation. Second, with proper scoring rules, assessors can maximize their subjectively expected score by making their SPDs correspond to their subjective knowledge. So, as far as assessors are responsive to the properness characteristic of the scoring rule, they are encouraged to report their true uncertainty. Third, the scoring-rule feedforward characteristic of the score curves may stimulate assessors to reflect on the consequences of their assessment and to reconsider it before making a final choice. Fourth, scoring-rule feedback in a training context fits in adequately with the scoring-rule feedforward interpretation of the curve. From the curve with possible scores, one is pointed out as the score actually obtained. Finally, the logarithmic transformation of the beta density function provides a natural operationalization for the often rather loosely defined concepts of lower and upper bounds.

It is crucial for the new elicitation methodology that a *proper* scoring rule is used for generating the score curves. It is less important which particular one is used, as long as the rule is proper (Van Lenthe, in press). So, it is not likely that the superior ELI performance found in the ELI evaluation study (Van Lenthe, 1993) originates from an artifact of the implementation of especially the logarithmic scoring curve. The logarithmic scoring rule apparently achieves more than a reduction in the tendency to use extreme responses. It is possible that the visualized scoring-rule feedforward about the consequences of the assessments stimulates assessors to be careful about the specification of uncertainty and to reconsider their assessment. The resulting increased amount of cognitive processing might be responsible for the positive external validity results with ELI and, it also might enhance the internal validity of assessments (Sniezek, Pease, & Switzer, 1990). So, even when the *strictly proper* characteristic of the logarithmic scoring rule fails to work, the increased amount of cognitive processing might do the job of enhancing internal validity.

The positive results of the ELI evaluation study encourage a further development of the technique.[1] The experimental version was established mainly to study the merits of the central ideas for the new technique. Future advancement will be aimed at making ELI more flexible and more suitable for different applied and experimental settings. For example, the grid of possible curves will be extended, allowing the estimation of very small proportions and the specification of more extreme certainties. Moreover, support will be not restricted to percentages; other continuous quantities will also be included.

In the experimental ELI version, the course of an elicitation session is fixed. With the next version, it will be possible for experimenters or decision analysts to implement their own instructions, their own training items, and their own questions of interest (see Note 1). In other words, the next ELI version will support the construction of complete elicitation sessions. The next version is intended to support different groups of users: (1) individual estimators who use ELI mainly as an elicitation aid (i.e., for the specification of their own uncertain knowledge), (2) experimenters or decision analysts who use ELI primarily as a tool to design an elicitation session, and (3) subjects (e.g., substantive experts) who are asked to complete a particular ELI session. The next ELI version will consist of three central parts that correspond with the three groups of users: (1) an estimate part for the individual estimator, (2) a design part for the experimenter, and (3) an option to run a particular elicitation session. Furthermore, the technique will have a statistics option to examine the characteristics of a particular score function and the corresponding probability distribution. And, of course, an output option will be included for saving relevant data and characteristics in an optional format. An example of a situation in which ELI might be useful is in a medical context to contrast the preconceptions of physicians, surgeons, and nurses about the surgical mortality of some of the serious congenital heart lesions (C. Bull, personal communication, June 1991). A database with information about the surgical mortality can be used to evaluate the assessments. It is also conceivable to use ELI to estimate in advance the surgical mortality of interventions about which no empirical data are available.

In view of the potential advantages of using graphical displays of score functions and in view of the positive experimental results thus far, it is expected that in the future ELI will contribute to the assessment of reliable and valid SPDs. Furthermore, it is anticipated that the technique will be efficient, acceptable, and suitable for different applied and experimental settings. Such an elicitation technique is of both scientific and practical interest. Theoretically or experimentally oriented studies as well as the analyses of decision or risk analyst can profit from the availability of an appropriate elicitation technique.

REFERENCES

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294-305). Cambridge: University Press.

Fischer, G. W. (1982). Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting. *Organizational Behavior & Human Performance*, **29**, 352-369.

Hofstee, W. K. B. (1987, December). *Overconfidence*. (Available from Willem K.B. Hofstee, Department of Personality Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands.)

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association*, **70**, 271-289.

Hogarth, R. M. (1980). *Judgement and choice: The psychology of decision.* Chichester: Wiley.

Huber, G. P. (1974). Methods for quantifying subjective probabilities and multi-attribute utilities. *Decision Sciences*, **5**, 430-458.

Jensen, F. A., & Peterson, C. R. (1973). Psychological effects of proper scoring rules. *Organizational Behavior & Human Performance*, **9**, 307-317.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: University Press.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 107-118.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P . Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge: University Press.

Lourens, P. F. (1984). *The formalization of knowledge by specification of subjective probability distributions*. Unpublished doctoral dissertation, University of Groningen, Groningen.

Ludke, R. L., Stauss, F. F., & Gustafson, D. H. (1977). Comparison of five methods for estimating subjective probability distributions. *Organizational Behavior & Human Performance*, **19**, 162-179.

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, **22**, 1087-1096.

Murphy, A. H., & Winkler, R. L. (1970). Scoring rules in probability assessment and evaluation. *Acta Psychologica*, **34**, 273-286.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

Pickhardt, R. C., & Wallace, J. B. (1974). A study of the performance of subjective probability assessors. *Decision Sciences*, **5**, 347-363.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior & Human Decision Processes*, **40**, 193-218.

Schaefer, R. E. (1976). The evaluation of individual and aggregated subjective probability distributions. *Organizational Behavior & Human Performance*, **17**, 199-210.

Schütt, K.-P. (1981). *Wahrscheinlichkeitsschätzungen im Computer-Dialog [Probability estimates with computer dialogue]*. Stuttgart: Poeschel Verlag.

Seaver, D. A., Von Winterfeldt, D., & Edwards, W. (1978). Eliciting subjective probability distributions on continuous variables. *Organizational Behavior & Human Performance*, **21**, 379-391.

Sniezek, J. A., Pease, P. W., & Switzer, F. S. (1990). The effect of choosing on confidence in choice. *Organizational Behavior & Human Decision Processes*, **46**, 246-282.

Spetzler, C. S., & Staël von Holstein, C.-A. S. (1975). Probability encoding in decision analysis. *Management Science*, **22**, 340-358.

Staël von Holstein, C.-A. S. (1970a). *Assessment and evaluation of subjective probability distributions*. Stockholm: Economic Research Institute.

Staël von Holstein, C.-A. S. (1970b). Measurement of subjective probability. *Acta Psychologica*, **34**, 146-159.

Staël von Holstein, C.-A. S. (1971). Two techniques for assessment of subjective probability distributions—An experimental study. *Acta Psychologica*, **35**, 478-494.

Staël von Holstein, C.-A. S. (1972). Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior & Human Performance*, **8**, 139-158.

Staël von Holstein, C.-A. S. (1977). The continuous ranked probability score in practice. In H. Jungermann & G. de Zeeuw (Eds.), *Decision making and change in human affairs* (pp. 263-273). Dordrecht, Holland: D. Reidel.

Terlouw, P. (1989). *Subjective probability distributions, a psychometric approach*. Unpublished doctoral dissertation, University of Groningen, Groningen.

Van Lenthe, J. (1993). ELI: An interactive elicitation technique for subjective probability distributions. *Organizational Behavior & Human Decision Processes*, **55**, 379-413.

Van Lenthe, J. (in press). Scoring-rule feedforward and the elicitation of subjective probability distributions. *Organizational Behavior & Human Decision Processes*.

Van Naerssen, R. F. (1962). A scale for the measurement of subjective probability. *Acta Psychologica*, **20**, 159-166.

Van Steen, J. F. J., & Oortman Gerlings, P. D. (1988). *Het gebruik van expertmeningen in veiligheidsstudies [The use of expert opinion in safety studies]*. Internal report, TU Delft/TNO.

Von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge: University Press.

Wallsten, T. S., & Budescu, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, **29**, 51-173.

Winkler, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Society*, **66**, 675-685.

Winkler, R. L. (1986). On "good probability appraisers." In P. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques* (pp. 265-278). Amsterdam: Elsevier.

Wright, G., & Wisudha, A. (1982). Distribution of probability assessments for almanac and future event questions. *Scandinavian Journal of Psychology*, **23**, 219-224.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior & Human Decision Processes*, **30**, 132-156.

Yates, J. F. (1988). Analyzing the accuracy of probability judgments for multiple events: An extension of the covariance decomposition. *Organizational Behavior & Human Decision Processes*, **41**, 281-299.

Yates, J. F., Zhu, Y., Ronis, D. L., Wang, D.-F., Shinotsuka, H., & Toda, M. (1989). Probability judgment accuracy: China, Japan, and the United States. *Organizational Behavior & Human Decision Processes*, **43**, 145-171.

NOTE

[1] The development of the next ELI version is a combined effort with the interuniversity expertise center ProGAMMA, a nonprofit center established by a group of Dutch universities to stimulate the development and diffusion of computer applications in the behavioral and social sciences. IecProGAMMA has the copyrights and is the distributor of the next ELI version. This version will be available by the end of 1993 from Iec ProGAMMA, P.O. Box 841, 9700 AV Groningen, The Netherlands (e-mail: gamma.post@gamma.rug.nl).

(Manuscript received January 25, 1991;
revision accepted for publication May 28, 1993.)