12-2008

A Comparison of Two Probability Encoding Methods: Fixed Probablity vs. Fixed Variable Values

Ali E. Abbas

David V. Budescu

Hsiu-Ting Yu

Ryan Haggerty

DECISION ANALYSIS
Vol. 5, No. 4, December 2008, pp. 190–202
ISSN 1545-8490 | EISSN 1545-8504 | 08 | 0504 | 0190

informs

DOI 10.1287/deca.1080.0126

A Comparison of Two Probability Encoding Methods: Fixed Probability vs. Fixed Variable Values

Ali E. Abbas
Department of Industrial and Enterprise Systems Engineering, College of Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, aliabbas@illinois.edu

David V. Budescu
Department of Psychology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801; and Department of Psychology, Fordham University, Bronx, New York 10458, budescu@fordham.edu

Hsiu-Ting Yu
Department of Psychology, Methodology and Statistics Unit, Leiden University, 2300RB, Leiden, The Netherlands, hsiutingyu@gmail.com

Ryan Haggerty
Department of Materials Science and Engineering, College of Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, haggerty@illinois.edu

We present the results of an experiment comparing two popular methods for encoding probability distributions of continuous variables in decision analysis: eliciting values of a variable, X, through comparisons with a fixed probability wheel and eliciting the percentiles of the cumulative distribution, $F(X)$, through comparisons with fixed values of the variable. We show slight but consistent superiority for the fixed variable method along several dimensions such as monotonicity, accuracy, and precision of the estimated fractiles. The fixed variable elicitation method was also slightly faster and preferred by most participants. We discuss the reasons for its superiority and conclude with several recommendations for the practice of probability assessment.

Key words: probability elicitation; decision analysis; continuous distributions; fractile estimation
History: Received on September 23, 2007. Accepted on August 30, 2008, after 2 revisions. Published online in *Articles in Advance* November 5, 2008.

1. Introduction

The elicitation of a representative probability distribution for a continuous variable is a fundamental step in decision making under uncertainty and has engendered a substantial literature. Several sources focus on the steps needed to elicit a probability distribution and methods to evaluate the quality of the estimates (see, e.g., Spetzler and von Holstein 1975, Wallsten and Budescu 1983, Edwards and von Winterfeldt 1987, von Winterfeldt and Edwards 1987, Merkhofer 1987, O'Hagan et al. 2006). Other sources focus on methods to construct the distribution using the moments of the variable of interest (see, e.g., Moder and Rodgers 1968, Perry and Greig 1975, Smith 1993) or use quantiles and/or moments to construct the distribution with a maximum entropy approach (e.g., Abbas 2002, 2006).

One widely used method for constructing probability distributions assumes a functional form for the probability function and uses the judges' assessments to estimate the parameters of this functional form. For example, Lindley (1987) uses three quantile assessments of a given variable to derive the three parameters of a skew logistic distribution. Other curve-fitting approaches use the assessed data to estimate the two parameters of a Beta distribution, which has found widespread popularity among Bayesian analysts, for its ease of updating with Bernoulli likelihood functions and for the wide variety of shapes it can reproduce (Raiffa and Schlaifer 1961).

Hughes and Madden (2002) review several methods for estimating the parameters of a Beta distribution. One approach derives the parameters using information about a location statistic and other

quantile assessments. The location statistic can be the mode (Fox 1966, Gilless and Fried 2000) or the mean (Duran and Booker 1988, Gross 1971, Weiler 1965), and the assessed quantiles may be some interval that contains the location statistics or a particular set of quantiles of the distribution. In related work, Van Dorp and Mazzuchi (2000) fit Beta distributions using two percentile constraints, León et al. (2003) propose two alternative methods for eliciting parameters of a Beta distribution, and AbouRizk et al. (1992) provide several visual methods for fitting Beta distributions to quantile assessments.

The focus of this paper is the comparison of different methods for constructing probability distributions using direct quantile assessments. Spetzler and von Holstein (1975) identify three types of probability encoding methods: fixed probability (FP), fixed variable (FV) value, and a mixture of the two. In its most general form, the FP method asks for the value of the variable that corresponds to a given cumulative probability. In a typical application of the FP approach, one selects a set of cumulative probabilities (p_i, $i = 1, \ldots, n$) and judges are asked to report values (v_i, $i = 1, \ldots, n$) such that $\Pr(V \le v_i) = p_i$. This method is often implemented in practice using a fixed setting on the probability wheel (hence, FP) to represent the chosen cumulative probability, p. The actual implementation takes the form of a choice between two gambles: The first (A) pays a certain amount of money, $\$X$, if the wheel is spun and an arrow lands on the target segment of the wheel. The second (B) pays the same amount $\$X$ if the value of the variable lies below a given value of the variable. If the judge prefers A, then he or she is offered a choice between A and a modified version of B (with a higher variable value) until indifference is achieved.

The most commonly chosen quantiles of the cumulative distribution for the FP method are the median ($p = 0.5$) and the quartiles ($p = 0.25$ and 0.75) (see, e.g., Hora et al. 1992, Lichtenstein et al. 1982). Variants of this approach are widely used in practice when experts provide their High, Base, and Low values (0.1, 0.5, 0.9) for a variable of interest to construct decision trees or to conduct sensitivity using tornado diagrams (Howard 1983, 1988; Felli and Hazen 2004; Keefer and Bodily 1983; McNamee and Celona 2001; Watson and Buede 1987). In some cases, it may be necessary to assess as many as five quantiles (e.g., 0.10, 0.25, 0.50, 0.75, and 0.90), or even seven (e.g., 0.01, 0.10, 0.25, 0.50, 0.75, 0.90, and 0.99) (Lau et al. 1996). Alpert and Raiffa (1982) report, however, that judges perform poorly when judging the extreme quantiles.

Variants of the FP paradigm are also used in practice, where judges are asked to provide their subjective $X\%$ probability intervals of a variable of interest (e.g., Alpert and Raiffa 1982, Budescu and Du 2007). For example, when asking judges for a 50% probability interval for a quantity (e.g., the price of a stock a year from now), the judge is asked to produce the two quartiles of the variable such that it is equally likely that the variable will fall between them or outside the interval. This approach has enjoyed popularity in the psychological literature, because it can be used to illustrate the judges' alleged overconfidence (see Russo and Schoemaker 1992). Soll and Klayman (2004) suggest that the precision of the encoding is improved if, instead of judging $X\%$ probability intervals, judges are asked to estimate separately its end points (i.e., the $(100 - X)/2$ and the $(100 + X)/2$ quantiles).

The FV method, in contrast, uses a fixed value of the variable (hence, FV) and asks for its corresponding cumulative probability. In this setting, the value of the variable is fixed and the probability wheel setting is adjusted until it corresponds to the judge's cumulative probability. The FV approach selects a set of variable values v_i, $i = 1, \ldots, n$ and asks judges to provide their cumulative probabilities (p_i, $i = 1, \ldots, n$) such that $\Pr(V \le v_i) = p_i$. For example, judges could be asked, "What is the probability that the price of this stock will be less than or equal to $v = \$20$ three months from now?" The assessed probabilities corresponding to the values (v) can be plotted, and the distribution can be obtained by fitting a smooth curve through the points. The FV approach is used in practice when assessing a full cumulative probability distribution for the variable of interest using a probability wheel. The wheel is adjusted until its setting corresponds to the value of the cumulative probability. Applications include competitive bidding situations where a full distribution of the maximum competitive bid is required to calculate the optimal bid (Carpen et al. 1971, Gates 1967). Other applications of the FV method include

assessing probabilities for lower and upper bounds of dose-response function curves (Wallsten et al. 1983).

Both FP and FV approaches are used in practice. Surprisingly, we have not found any prior comprehensive direct comparisons of these two encoding methods. Most previous studies have employed between-judge designs that make direct comparisons difficult (see, e.g., Juslin et al. 1999, Klayman et al. 1999, Soll and Klayman 2004). In general, there is a perception that FP judgments (especially when used to elicit the lower and upper bounds of $X\%$ probability intervals) induce more over confidence than FV judgments, but recently Budescu and Du (2007) have shown that this does not hold for all levels of confidence.

In this paper, we report the results of an online probability encoding experiment designed to compare the two assessment methods on a variety of criteria. This comparison allows us to provide new insights into the probability encoding process that are relevant for both practitioners and researchers. From a practical view, we address several questions, including (i) whether judgments are more consistent and/or precise with one method rather than another, (ii) whether one method is easier than the other, and (iii) whether judges prefer one method over the other. We also examine (iv) how the results from the two methods compare. These analyses provide direct suggestions for optimal structuring of elicitation sessions in practice and the type of encoding method to be applied.

The remainder of this paper is structured as follows. Section 2 describes the experimental methodology and setup. Section 3 presents the data analysis comparing the two encoding methods. Section 4 concludes with a set of recommendations for practitioners and recommendations for future research.

2. Experimental Methodology

2.1. Judges

The judges were 103 students enrolled in the decision analysis classes at Stanford University and at the University of Illinois volunteered to participate in this experiment. The participants included 71 men and 32 women, whose average age was 26.7 years, with a standard deviation of 5.9. Most were management science and engineering majors. All had been exposed to probability encoding in class lectures.

2.2. Procedure

The experiment was conducted online. After logging on to the site and reading the informed consent form, judges answered a few demographic questions. Next they were allowed to choose to assess either the closing value of the Dow Jones Industrial Average on December 12, 2006, or the high temperature in Palo Alto on December 12, 2006 (students at the University of Illinois were instructed to judge only the Dow Jones values). Judges could choose the units—Fahrenheit or Celsius—for the temperature assessment, but for the purpose of the analysis, all temperatures were converted to Celsius (we found no significant differences between the two sets of judges). A subset of the participants was chosen at random and provided with a chart of historical data for their variable of choice (see Figure 1).[1] Table 1 shows the number of judges in each condition.

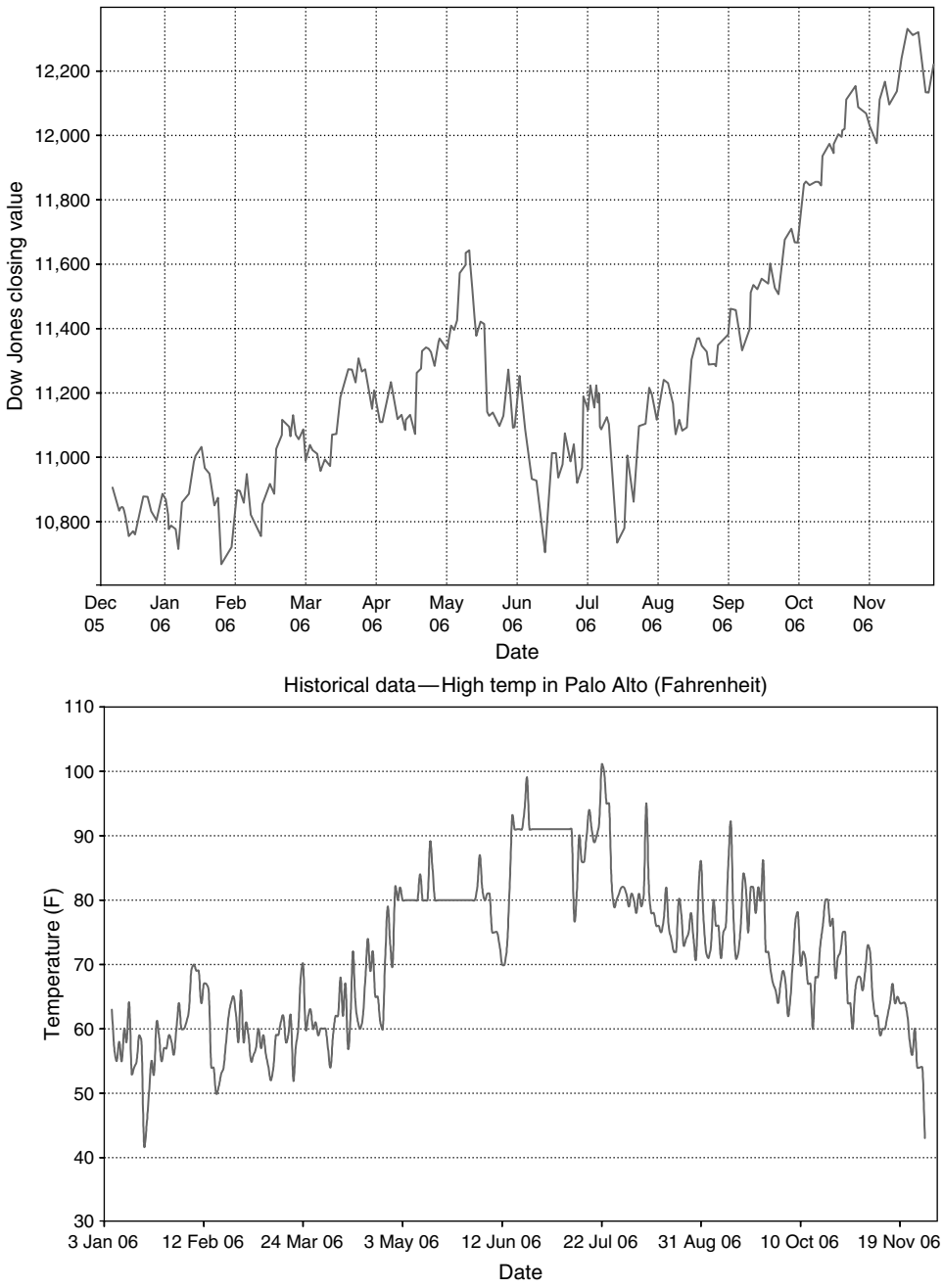
The experiment started a week before the target date. Judges could access the site at any time until December 11, 2006. Judges operated at their own pace, but completed all their judgments in one session that lasted between 30 and 45 minutes. After choosing their assessment variable, judges were asked to provide lower and upper bounds for that variable. Then they performed the two elicitation tasks (FP and FV) for their variable of choice. The order of the assessment, using FP or FV, was determined randomly.

Judges were presented with a sequence of binary choices regarding a hypothetical \$20 lottery (see Figure 2 for an example): The deal on the left displays a setting on the probability wheel. Judges win the prize if we spin the wheel and the arrow lands on grey. The deal on the right displays a certain value of the variable of interest. Judges win the prize if the value of the variable of interest is less than the displayed value.

If one chooses the deal on the left, one of two things could happen: For the FP method, the next screen would compare the same wheel setting with a higher

[1] This was done by generating a random number from a uniform distribution between 0 and 1. If the number was higher than 0.5, the judge was shown the chart, and if it was less than 0.5, he or she did not see it.

Figure 1 Historical Time Series of the Target Variables



value of the variable of interest, say 30 degrees. For the FV method, the next screen would compare the same value of the variable with a lower grey setting on the wheel, say 25% grey. In contrast, if the judge chooses the deal on the right, one of two things could happen: For the FP method, the next screen would compare the same wheel with a lower value of the variable of interest, say 20 degrees. For the FV method, the next screen would compare the same value with a higher grey setting on the wheel, say 75%. The next value was determined by a halving algorithm (see the appendix). The process stopped

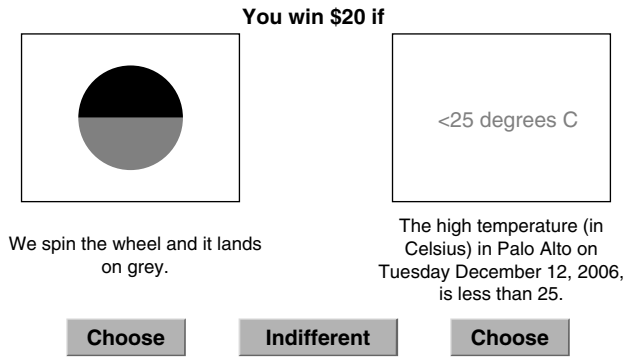
Table 1 Number of Judges in Each Experimental Condition

	No chart	With chart	Total
Dow Jones index	11	20	31
Temperature (Celsius)	13	17	30
Temperature (Fahrenheit)	22	20	42
Total	46	57	103

when the range of values was below a narrow (predetermined) threshold or when the judges expressed indifference between the two deals. At that point the judges were asked to confirm their decisions and a new series of choices with a new FP or FV value was initiated. In all our analyses we treat the midpoints of the ranges elicited as the judges' judgments.

Judges completed 10 series of judgments to determine five fractiles using each method. For the FP method, the fixed probabilities were 5%, 25%, 50%, 75%, and 95%; for the FV method, the values were set at 5%, 25%, 50%, 75%, and 95% of the range of variable values specified by the judges (extended by 20%). Note that although all judges used the same five probabilities in the FP method, the actual value used in the FV case varied across individuals, as a function of the range they specified. In both methods, the first point elicited was the central one (i.e., the median for FP and the midpoint of the range for FV). The other four points were presented in one of several predetermined orders that were counterbalanced across judges. The judges did not have direct access to their previous judgments when making their choices. After completing the assessments, the judges were asked a series of questions to evaluate the process and compare the two methods in terms of their ease and comfort level.

Figure 2 Example of an Elicitation Screen



3. Results

3.1. Monotonicity of Judgments

Recall that under each method the five points were elicited independently, without visible record of the previous points, and in no obvious order. Thus, the first question is whether the judges' judgments are monotonic and whether one of the methods induces a higher level of monotonicity. Monotonicity is satisfied if, for a pair of points, X_i and X_j, and their corresponding cumulative probabilities, we observe

$$X_i \geq X_j \leftrightarrow F(X_i) \geq F(X_j). \tag{1}$$

To determine the degree to which this condition is met, we calculated the Kendall τ_b rank correlation coefficient for judgments based on (i) midpoints of fractiles estimated with the FV method, (ii) midpoints of fractiles estimated with the FP method, and (iii) a combination of both estimates of the FV and FP methods. For any sample of size n there are $\binom{n}{2} = n(n-1)/2$ distinct pairs. In our case there are $n = 5$ fractiles (X_i) and their cumulative probabilities $F(X_i)$, defining 10 pairs for each method. Let C be the number of pairs that are concordant (i.e., that satisfy the condition in Equation (1)) and D be the number of pairs that are discordant (i.e., that violate the condition in Equation (1)). Kendall's τ_b is the difference between the proportion of concordant pairs and the proportion of discordant pairs. Formally:

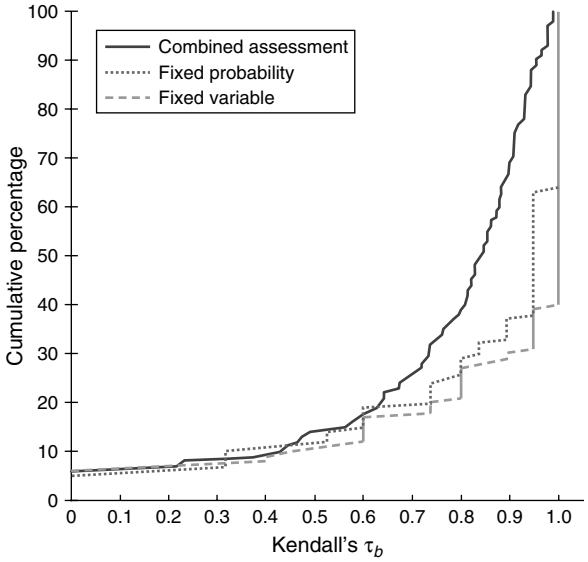
$$\tau_b = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \operatorname{sgn}(X_i - X_j)\operatorname{sgn}(F(X_i) - F(X_j))}{\binom{n}{2}} = \frac{C-D}{\binom{n}{2}} = \frac{C-D}{C+D},$$

where sgn is the sign function.[2]

Kendall's τ_b is a nonparametric measure that does not depend on the domain of assessments, the scales used, or their range. It is therefore convenient for comparing the two encoding methods. It ranges from -1 (all pairs are discordant) to 1 (all pairs are concordant), and it is 0 when there are equal numbers

[2] In the presence of ties, the numerator of the formula is $\sqrt{(C+D+T_x)(C+D+T_y)}$, where T_x is the number of pairs with ties on X (but not on Y), and T_y is the number of pairs with ties on Y (but not on X).

Figure 3 **Cumulative Percentages of Kendall's τ_b for Fixed Probability Assessments, Fixed Variable Assessments, and the Union of the Two Assessments**



of concordant and discordant pairs. Figure 3 shows the cumulative distributions of the Kendall's τ_b values for the two encoding methods. It also includes the values calculated using the combined assessments from both methods ($n = 10$, defining 45 pairs). The values based on the judgments with FV are slightly higher than those for the FP and, not surprisingly, both are superior to the joint set because it includes a larger number of points (the FP and the FV assessments). However, it is reassuring that the monotonicity of the combined assessments is not much lower than the results obtained for each method separately. This result also provides some insight into the stability of the assessments obtained using two different methods.

Table 2 summarizes the medians of Kendall's τ_b values. The first two rows describe the monotonicity of the 5 fractiles elicited within each method (FP and FV) separately, and the third row measures the monotonicity of all 10 fractiles elicited by the two methods combined. The last row in each panel summarizes the degree of order consistency between the two methods (that is, the degree to which judgments elicited with one method are ordinaly consistent with those obtained with the other one). Although it is, clearly, lower than the monotonicity achieved within each method separately, it is quite high, indicating

Table 2 **Monotonicity of the Judgments (Fixed Probability, Fixed Value, and All Assessments)**

Median rank correlation, τ_b	Temperature	Dow Jones	All assessments
FP	0.95	0.95	0.95
FV	1.00	1.00	1.00
Combined points (FP & FV)	0.86	0.85	0.85
Cross methods (FP & FV)	0.80	0.74	0.77

an almost 90% level of rank agreement. These values confirm the impressions from the figure and highlight the impressive level of monotonicity achieved by the judges with either method, as well as for the combined assessments.

One third of the judges had identical Kendall's τ_b for both methods. However, the majority of judges (45%) had higher rank correlation coefficients in the FV assessment, and only a minority (22%) was more monotonic in the FP assessment. This difference is significant by a sign test ($Z = 1.81$; $p < 0.05$ one sided). Interestingly, judges who were monotonic in one method were more likely to be monotonic in the other method as well. The correlation between the two (within judge) measures of monotonicity is 0.51 ($p < 0.05$), and it is consistent across the two variables.

3.2. Fitting the Judgments

We fitted Beta distributions to the midpoints of the fractiles estimated with the FP and FV methods, separately. The Beta is a continuous two-parameter distribution defined over a given bounded range. Its density is given by

$$Beta(\alpha, \beta, a, b, x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{\int_a^b (x-a)^{\alpha-1}(b-x)^{\beta-1}\,dx}, \qquad (2)$$

where a and b are the lower and upper bounds of the domain (respectively), and α and β are the two parameters of the Beta distribution. Of course, Beta is not the only distribution that can be used to model bounded variables (see, e.g., Kotz and van Dorp 2004, van Dorp et al. 2007), but it is frequently used as a prior distribution in Bayesian analysis. We make no claim of superiority or exclusivity for the Beta but use it to illustrate the results and to facilitate the comparison of the two methods in a meaningful fashion.

We used Matlab's "fminsearch" function to minimize the squared residuals and estimate the two

shape parameters $(\hat{\alpha}, \hat{\beta})$ within the range (lower and upper bounds, a and b) defined by the judges' judgments. Note that the optimization takes different forms, depending on what is being minimized. When using fixed probabilities, p_i, $i = 1, \ldots, 5$, we minimized sums of squared deviations in the metric of the variables ($X =$ Temperature or Dow Jones):

$$\min_{\hat{\alpha}, \hat{\beta}} \sum_{i=1}^{5} (X_i - \hat{X}_i)^2, \tag{3}$$

where $\hat{X}_i = BetaInverse(p_i, \hat{\alpha}, \hat{\beta}, a, b)$ and X_i is the elicited value of the variable corresponding to a cumulative probability, p_i. In the case of fixed values, V_i, $i = 1, \ldots, 5$, we minimized total squared deviations in the metric of cumulative probabilities:

$$\min_{\hat{\alpha}, \hat{\beta}} \sum_{i=1}^{5} (p_i - \hat{p}_i)^2, \tag{4}$$

where $\hat{p}_i = Beta(X_i, \hat{\alpha}, \hat{\beta}, a, b)$ and p_i is the elicited value of the cumulative probability corresponding to the variable value, X_i.

The moments of the Beta distribution are simple functions of the two parameters α and β as well as the upper and lower bounds, a and b. More specifically, the mean, μ, and variance, σ^2, are given by

$$\mu = \frac{\alpha b + \beta a}{\alpha + \beta}, \qquad \sigma^2 = \frac{\alpha\beta(b-a)^2}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{5}$$

Figure 4 presents examples of some of the better fits obtained from the experiment. The legends of the figures list the shape parameters (α and β) and the upper and lower bounds (a and b). Table 3 summarizes the means and standard deviations of the distributions of the various variables for each elicitation method. The top panel uses all 103 judges, and the bottom panel presents results based only on those judges who

Figure 4 Examples of Beta Fits

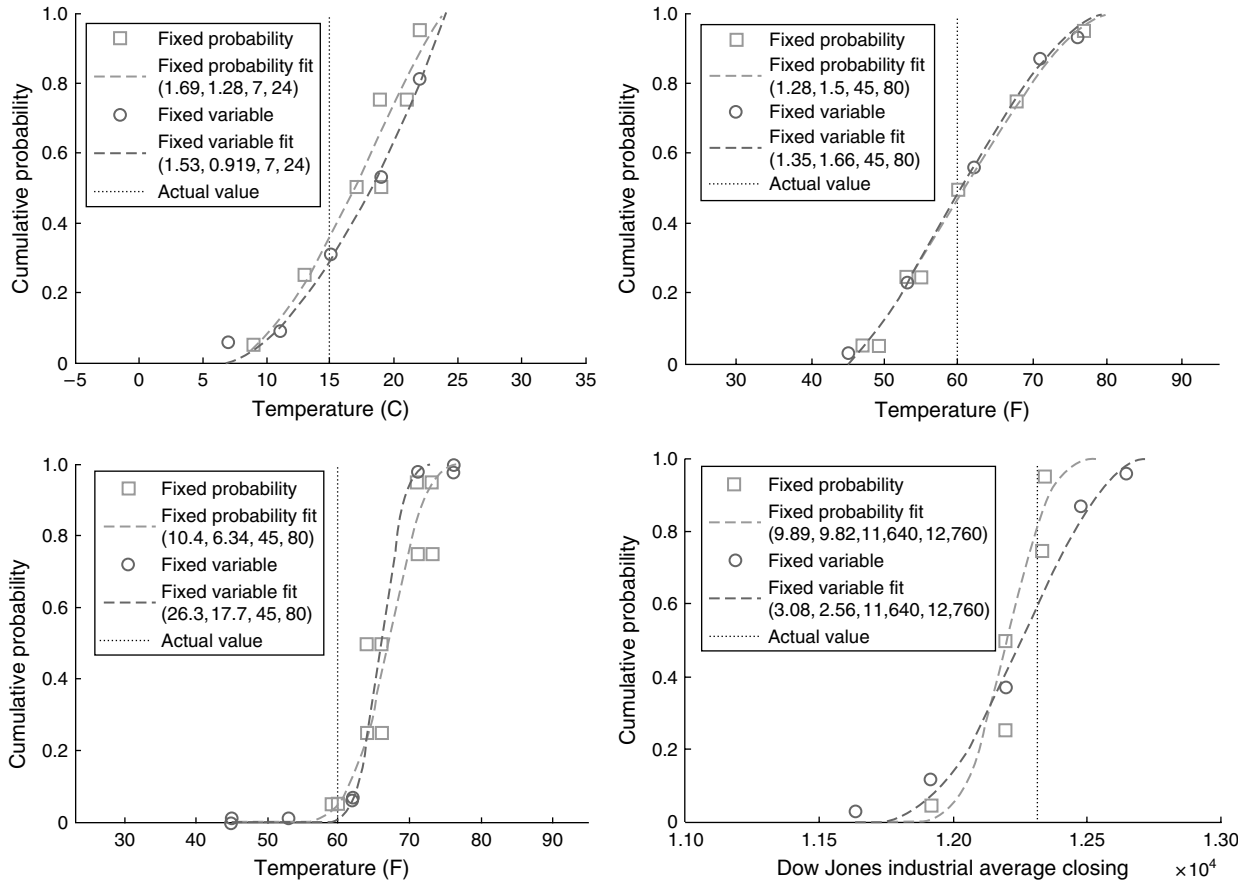


Table 3 Average Means and Standard Deviations of the Fitted Beta Distributions

	Temperature		Dow Jones	
Method	Mean	SD	Mean	SD
FP	14.99	3.58 (2.19)	12,523	747 (1,470)
(X residuals)	(4.16)		(623)	
FV	15.36	5.19 (2.74)	12,427	864 (1,648)
($F(x)$ residuals)	(3.90)		(393)	
Only cases with $\tau_b \geq 0.8$:				
FP	15.03	4.12 (1.98)	12,385	494 (1,110)
(X residuals)	(3.79)		(488)	
FV	15.15	5.08 (2.44)	122,967	615 (1,238)
($F(x)$ residuals)	(3.94)		(351)	

displayed high levels of monotonicity ($\tau_b \geq 0.8$) with both assessment methods (FP and FV) ($n = 74$). The values are reasonable (the temperature on the target date was 15°C = 59°F, and the Dow Jones closed at 12,316) and are quite similar in the two methods.

Table 4 compares the moments of the two distributions fitted for each judge and counts the number of cases (i.e., judges) where one method had a higher mean or variance (note that these are all within-judge comparisons). There are only small differences in the fitted means (and almost equal splits of judges with higher/lower means under each method, with 28/44 for temperature and 15/16 for Dow Jones). In contrast, for 85% of the judges (59 judges for temperature and 29 judges for Dow Jones), the variance of the distribution extracted from the FV is higher than its counterpart based on FP.[3]

Table 5 summarizes the goodness of fit of the solutions as measured by the root mean squared error (RMSE) for each method and variable (averaged across all judges), as well as the count of cases where one method outperformed the other. As indicated earlier, when fitting the distributions we minimized different types of residuals ($F(X)$ in the FV method and X in the FP method). In the latter case, we normalized the values relative to the range stated by the judges, so all the RMSEs are in the 0–1 range and can be compared meaningfully. The two methods fit equally well,

[3] We repeated the curve-fit analysis for each method using the other optimization procedure (optimizing Equation (3) for FV, and minimizing Equation (4) for FP) and found again that the distribution fitted to the FV judgments had higher variances, indicating that this is not an artifact of the curve-fitting method.

Table 4 Comparison of the Moments of the Distributions Fitted from the Two Elicitation Methods

	Temperature	Dow Jones
Mean(FP) − Mean(FV)	−0.37	59
|(Mean(FP) − Mean(FV))|	1.28	153
No. of positive diffs./no. of negative diffs.	28/44	15/16
SD(FP)/SD(FV)	0.75	0.59
No. of ratios > 1/no. of ratios < 1	13/59	2/29

and there is no clear advantage to one method over the other.

3.3. Accuracy of the Judgments

In this section, we address the question of how well the probability distributions provided by the various judges under the two methods fared with the historical record of the temperatures in Palo Alto.[4] Given that temperatures at a particular location are relatively stable over time and vary only negligibly within a week, we constructed the distribution of the temperatures in Palo Alto on December 12 ± 3 days (i.e., December 9–15) based on the data recorded between 1955 and 2007 (we obtained 345 data points for this location and dates at http://www.wunderground.com/history/airport/KPAO/2007/12/15/DailyHistory.html). As a measure of proximity, we used the Kolmogorov-Smirnov statistic (the maximal absolute difference between the estimated cumulative distributions and the historical distribution of temperatures) for the FV and the FP judgments. On average, the Kolmogorov-Smirnov scores of the FV estimates are smaller than their FP counterparts (mean difference = 0.064), and they are significantly closer ($t(71) = 2.37$, $p < 0.05$) to the historical distribution. This pattern also holds for a small (but statistically significant) majority of judges (56%; $Z = 2.39$; $p < 0.05$). Thus, the distributions extracted from the FV method fit the historical data slightly better.

The same pattern is observed when we compare the sum of squared differences between the FV estimates and the historical data and their counterparts based on the FP estimates. They are lower (mean difference = 0.102, ($t(71) = 1.94$, $p = 0.06$); this also

[4] Evidently, this analysis is not meaningful for the Dow Jones values that vary systematically over time.

Table 5 Comparison of Goodness of Fit for Distributions Fitted with the Two Methods

	Temperature	Dow Jones	Overall
RMSE(FP)	0.072	0.081	0.075
RMSE(FV)	0.078	0.098	0.084
RMSE(FP) − RMSE(FV)	−0.006	−0.017	−0.009
|RMSE(FP) − RMSE(FV)|	0.056	0.072	0.061
No. of positive diffs./ no. of negative diffs.	35/37	17/14	52/51

holds for a small, but significant, majority of the individual judges (58%, $Z = 2.82$; $p < 0.05$)).

Whereas the previous sections analyzed the quality of the judgments extracted by the two methods, in the next section we focus on a comparison of the methods in terms of the judges' performance and perceptions. More specifically, we ask whether the judges find one method easier to use by analyzing both objective and subjective measures.

3.4. Reaching Indifference

Our elicitation procedure yields upper and lower bounds for each of the fractiles obtained with either encoding method (recall that the elicitation procedure terminated when the difference between the upper and lower bounds was below a specified low threshold or when judges clicked "Indifferent"). Upper and lower bounds may also appear in practice if there is not enough time to reach indifference or if the information available is too vague and prevents one from identifying a precise indifference point (e.g., Wallsten et al. 1983). Table 6 summarizes the proportion of cases where judges actually converged to a single point. There is an impressive level of convergence (for example, 78% of judges expressed indifference for the assessments of Dow Jones values with FP). However, the FV method induces higher percentages of indifference for temperature. This difference is statistically significant for the temperature ($p < 0.05$ by a sign test). The FP method yielded a higher percentage of point estimates for the Dow Jones, but this difference is not significant.

Table 6 Proportion of Cases Where Judges Converged to a Point

Method	Dow Jones (%)	Temperature (%)
FP	78.07	60.00
FV	74.84	83.05

Table 7 Mean Judgment Time (in Seconds) as a Function of the Target Variable, Presence of the Chart, and the Elicitation Method

			FP		FV		Difference	
Variable	Chart	*N*	Mean	SD	Mean	SD	Mean	SD
Dow Jones	N	11	8.97	(3.00)	3.92	(0.67)	5.05	(3.06)
Dow Jones	Y	20	7.41	(1.20)	6.82	(0.91)	0.59	(0.90)
Temperature	N	35	6.79	(0.77)	5.23	(0.69)	1.55	(0.59)
Temperature	Y	37	7.00	(0.60)	6.21	(0.46)	0.78	(0.73)
Overall		103	7.22	(0.51)	5.75	(0.35)	1.46	(0.50)

3.5. Response Time

Are judgments faster (and, presumably, easier) in one of the methods? Judges used different numbers of questions to reach the upper and lower bound for various fractiles. Thus, we compared the average response times per question in the various conditions. The mean response times of the judges (across all five series) were analyzed in a three-way ANOVA with two between-judges factors (units and presence of chart) and one within-judge factor (elicitation method). These means are presented in Table 7.

The effect of the elicitation method is significant ($F(1, 99) = 13.04$, $p < 0.05$), with the FV method inducing faster responses. The presence of the chart slowed down the response time by about half a second (6.86 vs. 6.22). This difference was also significant ($F(1, 99) = 5.62$, $p < 0.05$).

3.6. Perceptions and Preferences of the Judges

At the conclusion of the experiment, judges were asked to compare the two methods along three dimensions using seven-point scales. For the purpose of this analysis we collapsed these ratings into three coarser categories: The midpoint of the scale (4) is interpreted as indicative of indifference/neutrality between the methods, and all responses on one side of the scale (1–3 and 5–7) were classified as favoring one of the methods. A clear majority (64%) of the respondents thought that the FV elicitation method is simpler and more natural. Table 8 also shows a clear preference for this elicitation method. The results are identical for both variables and with/without the benefits of historical charts.

Are the judges' preferences for a method reflected in the quality of their judgments? Table 9 cross-tabulates the judges' preferences for a method and the method in which they had higher Kendall's τ_b.

Table 8 Preference for Elicitation Method in the Various Groups

			Which method do you prefer?		
Variable	Chart	*N*	FP (%)	Equal (%)	FV (%)
Dow Jones	No	11	27	9	64
	Yes	20	30	10	60
Temperature	No	35	14	9	77
	Yes	37	27	11	62
Combined	No	46	17	9	74
	Yes	57	28	11	61
		103	23	10	67

Table 9 Relationship Between Preferred Method and Monotonicity of Judgments

	Higher monotonicity			
Method preferred	FV	Equal	FP	Total
FP	7	10	7	24
Indifferent	9	1	0	10
FV	30	23	16	69
Total	46	34	23	103

Among those who prefer the FP method (top row), there is no difference between the monotonicity under the two methods but, remarkably, for those who prefer the FV elicitation (or were indifferent between the two methods), the preferred method induces, indeed, higher levels of monotonicity in a clear majority of the cases.

4. Summary, Conclusions, and Recommendations

We used a Web-based system based on simple binary choices to elicit fractiles of probability distributions. Our main goal was to compare two competing methods: FP and FV values. All the assessments were made in real time, one fractile at a time, so the judges could not see their previous judgments. The FV and FP methods were successful: The judges reported no major problems, and provided high-quality—monotonic, reasonable, and meaningful—judgments that were consistent across the two methods.

The results of our experiment show that the two methods were practically indistinguishable in many ways (e.g., the means and the goodness of fit of the Beta distributions based on the FP and FV judgments). We did find, however, several systematic differences between the two methods, and these differences point to a slight superiority of the FV method. The judges were able to make these judgments faster and were more likely to reach full indifferences (rather than establishing narrow intervals) with the FV assessments. It is not surprising that the majority of the judges express a clear preference for this method in the postexperimental evaluations. Convenience and ease of use do not guarantee quality, so it is reassuring that the FV method also resulted in judgments with higher levels of monotonicity and matched slightly better the historical distribution of the target variable. The distributions based on the FV method had higher variances than their counterparts based on FP. Given the recurring concern that subjective probabilities are too narrow (reflecting overconfidence), we view this as a positive feature of the method.

We believe that two factors can explain the superiority of the FV method in our study. The first is, simply, the nature of the response scale (probabilities), which is bounded by 0 (impossibility) and 1 (certainty) and universal in the sense that it applies to all events and is independent of the measurement units of the particular variables. The second factor is the fact that people more often face and are more familiar with problems that resemble the FV judgments. These judgments match the single-event format of most decisions (under risk) that we encounter daily. We are likely to answer (to ourselves, or to others) questions regarding the likelihood that certain future events will exceed predetermined thresholds. For example, we may need to judge how likely it is that (i) the temperature will be above 30°, (ii) the rain will last more than 1 hour, (iii) one's blood pressure will be below the threshold requiring medication for hypertension, (iv) one's child's SAT score will be above the admission cutoff of her favorite college, etc. To answer such questions, one relies on his or her life-long experience with these variables and specific cues about the particular target (what I know about today's weather or my child's abilities). However, we rarely need to estimate the required level of a certain event to reach a certain probability. Questions such as how hot it should be next Sunday, to exceed 90% of all summer days, or how high our child's SAT score should be to place her in the top 15% of the applicants to her favorite college, etc. are more complex because they require more knowledge about other possible outcomes.

4.1. Some Practical Recommendations

Recall that in our study the various fractiles were elicited in isolation using binary comparisons with no direct access to previous assessments. This restriction makes perfect sense in a research setting but could be relaxed in a decision analysis. It is safe to assume that when judges have access to their previous assessments and events are presented in a systematic fashion (e.g., in ascending or in descending order), the performance would be improved in all respects (e.g., monotonicity, speed, level of convergence). Thus, our results provide some sense of the lower bounds for the monotonicity and accuracy of the FP and FV methods in decision analysis.

There are two somewhat surprising findings in our study. About half the judges in our sample had access to historical charts of the relevant variables, but for the most part this extra information did not make a difference—the quality of their judgments and the distributions extracted from them was, essentially, identical to that of the other half, who did not have access to this aid. The simplest explanation is that we observe a "ceiling effect." Recall that (i) we allowed judges to select which distribution they preferred to assess (and in the case of temperature, to choose their units) and (ii) both variables were familiar (and "experienced" on a daily basis) to begin with. In other words, it is likely that judges selected to judge the variables about which they were most knowledgeable, so there was not much information in the charts that was not available to them anyway! We hypothesize that more information in the form of past results would have been beneficial if they were to judge variables with which they were less familiar (say temperatures and stock market outcomes in other countries). This has to be verified in future work, but *we recommend* (*tentatively*) *having such information in the system and allowing the judge to determine whether he or she wants to access it while making the judgments*.

We observed that judges who took longer times to make the judgments were not necessarily more consistent than those who answered faster (there was no correlation between time to answer and global monotonicity). Judges were not instructed to answer quickly and were not offered any incentives for slower/faster response rates. It makes sense to assume that they answered at the rate that was most convenient and natural for them and *we recommend* (*tentatively*) *not imposing time constraints but allowing judges to respond at their preferred rate*.

The theoretical literature indicates that it is possible to fit Beta distributions based on as few as two points, but in many applications of decision analysis the norm is three fractiles (McNamee and Celona 2001). We achieved satisfactory fits with five points in both methods. In sensitivity analyses (not reported here), we found that the fitting procedure was highly robust to significant changes in the domain (up to $\pm 20\%$) of the fitted Beta distribution. Our results indicate that the removal of the end point fractiles from the elicitation led to the highest change in variance, and the removal of the mid fractile led to the lowest change; Budescu and Du (2007) documented the differential pattern of miscalibration of the subjective 90% probability intervals. In light of these results, *we recommend asking judges to* (i) *estimate the range of the target variable,* (ii) *encourage them to be generous in this task and consider all feasible values of the variable, and* (iii) *elicit at least five fractiles*.

Our method relied on a self-terminating series of binary questions that identify narrower ranges at every step. Ideally, this series of questions ends when the judge declares his or her indifference between the wheel and the deal that depends on the target quantity. In our algorithm for the FV method (that used up to seven consistent answers), this ideal was achieved in more than 80% of the cases overall, and the ranges identified in cases where convergence to indifference was not achieved were quite narrow (lower than 0.02 overall). Algorithms that terminate before seven questions would lead to fewer indifferences and wider ranges.

4.2. Future Research

Although the results of this study favor the FV method, we recognize that their generalizability should be reexamined in future studies using different variables and judges, including populations of acknowledged experts. An additional factor that should be studied is the robustness of our findings under various changes to the algorithm we employed here. For example, future work should test whether the results hold if the original range of values is predetermined by the experimenter (rather than being

selected by the judge) and the iterative sequence of preferences is replaced by a more direct equivalence judgment.

Acknowledgments

This material is based on work supported by the National Science Foundation under Award SES 06-20008. The authors thank Ronald A. Howard for granting them access to the students in the decision analysis class at Stanford University and Ryan Mulligan for programming the study.

Appendix. The Elicitation Algorithm

A range is a certain interval (with lower and upper bounds) that brackets the value we are interested in. The range is reduced through a sequence of questions (choice options) and answers (choices). Two types of range reductions were done. One was the range of the fraction (percentage) of the grey section of the probability wheel, and the other was the range of the target values. The first case reduces to a special case of the second case when the variable range is 100. As such, there was only one reduction algorithm in the experiment.

The reduction algorithm consists of halving the range with each question. This approach provides the maximum reduction in the entropy of the range if we believe that a value is uniformly distributed across the range. There were two stopping conditions:

1. The user was indifferent between the two deals in the question. This indicates the range should be reduced to this point.
2. The range interval is reduced to less than three units.

The minimum number of questions to reach a stopping condition is one, while the maximum number of questions is $\log_2(Range) - 1$.

Confirmation questions were also asked as a consistency check. The number of questions asked varied, depending on the stopping condition of the above algorithm. If the stopping condition was indifference, only one confirmation question was asked. If the stopping occurred because the range was narrow enough, two confirmation questions were asked (at the upper and lower values of the range). Taking the confirmation questions into account, the maximum number of questions asked for a given point on the marginal distribution was $\log_2(Range) + 1$. For the special case of the fixed quantity, this means the maximum number of questions is seven when indifferent and eight otherwise.

References

Abbas, A. E. 2002. Entropy methods for univariate distributions in decision analysis. C. Williams, ed. *Proc. 22nd Internat. Workshop on Bayesian Inference and Maximum Entropy Methods Sci. Engrg.*, American Institute of Physics, Melville, NY, 339–349.

Abbas, A. E. 2006. Entropy methods for joint distributions in decision analysis. *IEEE Trans. Engrg. Management* **53** 146–159.

AbouRizk, S. M., D. W. Halpin, J. R. Wilson. 1992. Visual interactive fitting of beta distributions. *J. Construction Engrg. Management* **117** 589–605.

Alpert, M., H. Raiffa. 1982. A progress report on the training of probability assessors. D. Kahneman, P. Slovic, A. Tversky, eds. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK, 294–305.

Budescu, D. V., N. Du. 2007. The coherence and consistency of investors' probability judgments. *Management Sci.* **53**(11) 1731–1744.

Carpen, E. C., E. V. Clapp, W. Campbell. 1971. Competitive bidding in high risk situations. *J. Petroleum Tech.* **23** 641–653.

Duran, B. S., J. M. Booker. 1988. A Bayes sensitivity analysis when using the beta distribution as a prior. *IEEE Trans. Reliability* **37** 239–247.

Edwards, W., D. von Winterfeldt. 1987. Public values in risk debates. *Risk Anal.* **7** 141–158.

Felli, J., G. Hazen. 2004. Javelin diagrams: A graphical tool for probabilistic sensitivity analysis. *Decision Anal.* **1**(2) 93–107.

Fox, B. L. 1966. A Bayesian approach to reliability assessment. Memorandum RM-5084-NASA, The RAND Corporation, Santa Monica, CA.

Gates, H. 1967. Bidding strategies and probabilities. *J. Construction Division, ASCE, Paper 5159*, **93** 75–107.

Gilless, J. K., J. S. Fried. 2000. Generating Beta random rate variables from probabilistic estimates of fireline production times. *Ann. Oper. Res.* **95** 205–215.

Gross, A. J. 1971. The application of exponential smoothing to reliability assessment. *Technometrics* **13** 877–883.

Hora, S. C., J. A. Hora, N. G. Dodd. 1992. Assessment of probability distribution for continuous random variables: A comparison of bisection and fixed value methods. *Organ. Behav. Human Decision Process* **51** 135–155.

Howard, R. A. 1983. The evolution of decision analysis. R. A. Howard, J. E. Matheson, eds. *The Principles and Applications of Decision Analysis*. Strategic Decisions Group, Menlo Park, CA, 5–16.

Howard, R. A. 1988. Decision analysis: Practice and promise. *Management Sci.* **34**(6) 679–695.

Hughes, G., L. V. Madden. 2002. Some methods for eliciting expert knowledge of plant disease epidemics and their application in cluster sampling for disease incidence. *Crop Protection* **21** 203–215.

Juslin, P., P. Wennerholm, H. Olsson. 1999. Format-dependence in subjective probability calibration. *J. Experiment. Psych.: Learn., Memory, Cognition* **25** 1038–1052.

Keefer, D. L., S. E. Bodily. 1983. Three-point approximations for continuous random variables. *Management Sci.* **29**(5) 595–609.

Klayman, J., J. Soll, C. Gonzalez-Vallejo, S. Barlas. 1999. Overconfidence: It depends on how, what, and whom you ask. *Organ. Behav. Human Decision Processes* **79** 216–247.

Kotz, S., J. R. van Dorp. 2004. *Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications*. World Scienific Press, Singapore.

Lau, A. H. L., H. S. Lau, Y. Zhang. 1996. A simple and logical alternative for making PERT time estimates. *IIE Trans.* **28** 183–192.

León, C. J., F. J. Vázquez-Polo, R. L. González. 2003. Elicitation of expert opinion in benefit of environmental goods. *Environ. Resource Econom.* **26** 199–210.

Lichtenstein, S., B. Fichhoff, L. D. Phillips. 1982. Calibration of probabilities: The state of the art to 1980. D. Kahneman, P. Slovic,

A. Tversky, eds. *Judgment Under Uncertainty*: *Heuristics and Biases*. Cambridge University Press, Cambridge, UK, 306–334.

Lindley, D. V. 1987. Using expert advice on a skew judgmental distribution. *Oper. Res.* **35**(5) 716–721.

McNamee, P., J. Celona. 2001. *Decision Analysis for the Professional*, 3rd ed. SmartOrg Inc., Menlo Park, CA.

Merkhofer, M. W. 1987. Quantifying judgmental uncertainty: Methodology, experiences, and insights. *IEEE Trans. Systems, Man, Cybernetics* **17** 741–752.

Moder, J. J., E. G. Rodgers. 1968. Judgment estimate of the moments of PERT type distributions. *Management Sci.* **18**(2) 76–83.

O'Hagan, A., C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, T. Rakow. 2006. *Uncertain Judgments*: *Eliciting Experts' Probabilities*. John Wiley & Sons Ltd., West Sussex, UK.

Perry, C., I. D. Greig. 1975. Estimating the mean and variance of subjective distributions in PERT and decision analysis. *Management Sci.* **21**(12) 1477–1480.

Raiffa, H., R. Schlaifer. 1961. *Applied Statistical Decision Theory*. Harvard University Press, Boston.

Russo, J. E., P. J. H. Schoemaker. 1992. Managing overconfidence. *Sloan Management Rev.* **33** 7–17.

Smith, J. E. 1993. Moment methods for decision analysis. *Management Sci.* **39**(3) 340–358.

Soll, J. B., J. Klayman. 2004. Overconfidence in interval estimates. *J. Experiment. Psych.*: *Learn., Memory, Cognition* **30** 299–314.

Spetzler, C. S., C.-A. S. S. von Holstein. 1975. Probability encoding in decision analysis. *Management Sci.* **22**(3) 340–358.

Van Dorp, J. R., T. A. Mazzuchi. 2000. Solving for the parameters of a Beta distribution under two quantile constraints. *J. Statist. Comput. Simulation* **67** 189–201.

van Dorp, J. R., S. Rambaud, J. Prez, R. H. Pleguezuelo. 2007. An elicitation procedure for the generalized trapezoidal distribution with a uniform central stage. *Decision Anal.* **4**(3) 156–166.

von Winterfeldt, D., W. Edwards. 1987. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, UK.

Wallsten, T. S., D. V. Budescu. 1983. Encoding subjective probabilities: A psychological and psychometric review. *Management Sci.* **29**(2) 151–173.

Wallsten, T. S., B. Forsyth, D. V. Budescu. 1983. Stability and coherence of health experts' upper and lower subjective probabilities about dose-response curves. *Organ. Behav. Human Decision Processes* **31** 277–302.

Watson, S. R., D. M. Buede. 1987. *Decision Synthesis*: *The Principles and Practice of Decision Analysis*. Cambridge University Press, Cambridge, UK.

Weiler, H. 1965. The use of incomplete Beta functions for prior distributions in binomial sampling. *Technometrics* **7** 335–347.