A Method for Evaluating Elicitation Schemes for Probabilities

Haiqin Wang[‡], Denver Dash[‡] & Marek J. Druzdzel[†‡]
Decision Systems Laboratory
[†]School of Information Sciences
and [‡]Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260
{whq,ddash,marek}@sis.pitt.edu

Abstract

We present an objective approach for evaluating probability elicitation methods in probabilistic models. Our method draws on ideas from research on learning Bayesian networks: if we assume that the expert's knowledge is manifested essentially as a database of records that have been collected in the course of the expert's experience, and if this database of records is available to us, then the structure and parameters of the expert's beliefs could be reliably constructed using techniques for Bayesian learning from data. This learned model could, in turn, be compared to elicited models to judge the effectiveness of the elicitation process. We describe a general procedure by which it is possible to capture the data corresponding to the expert's beliefs, and we present a simple experiment in which we utilize this technique to compare three methods for eliciting discrete probabilities: (1) direct numerical assessment, (2) the probability wheel, and (3) the scaled probability bar. We show that for our domain, the scaled probability bar is the most effective tool for probability elicitation

Introduction

As more and more decision-analysis models are being developed to solve real problems in complex domains, extracting knowledge from experts is arising as a major obstacle in model building (Druzdzel & van der Gaag 2000). Quite a few methods have been proposed to elicit subjective probabilities from domain experts. These techniques are concerned with balancing quality of elicitation with the time required to elicit the enormous number of parameters associated with many practical models. Systematic evaluation and comparison of different model elicitation methods are thus becoming of growing concern.

In Bayesian probabilistic models, encoded probabilities reflect the degree of personal beliefs of the experts. The sole purpose of probability elicitation is to extract an accurate description of the expert's personal beliefs. In order to judge whether the elicitation procedure has produced an accurate model, therefore, the elicitor must know intimate details about the expert's knowledge. Unfortunately, these details that the elicitor is seeking from the start are hidden from explicit expressions; so it has not been possible to evaluate elicitation schemes directly. Less direct methods are the only possibility.

In this paper we present an objective approach for evaluation of elicitation methods that avoids the assumptions and pitfalls of existing approaches. Our technique is much closer to the ideal "direct" comparison between the elicited network and the expert's beliefs. The main idea is to simulate the training/learning process of an expert by allowing the trainee to interact with a virtual domain. Underlying the domain is a Bayesian network that is used to stochastically update the state of the world in response to the subject's interaction. Then by recording every state of the world that is experienced by the trainee, we can effectively gain direct access to the trainee's knowledge. It is quite an established fact that people are able to learn observed frequencies with an amazing precision if exposed to them for a sufficient length of time (Estes 1976). Therefore, after training, the trainee obtains some level of knowledge of the virtual world and, consequently, becomes an expert at a certain proficiency level. This knowledge, in the form of a database of records, D_{exp}, can be converted to an "expected" model of the expert, $\widehat{M}_{exp}$, by applying Bayesian learning algorithms to D_{exp}. Finally, this expected expert model can be directly compared to the model elicited from the expert to judge the accuracy of elicitation.

Our approach captures a subject's state of knowledge of the probabilistic events in the toy world. The subject's experience with the toy world, rather than the actual model underlying the world, forms the basis of his or her knowledge. For this reason the learned model should be the standard used to evaluate the elicitation schemes, rather than the original toy model. This technique allows us to avoid the expensive process of training subjects to fully-proficient expertise. For example, our expert's experience may have led him to explore some states of the world very infrequently. In this case, even if our elicitation procedure is perfect, the elicited probabilities of these states may be significantly different from the underlying model. Using the expert's experience rather than the original model gets around this problem completely because we know precisely how many times our expert has visited any given state of the world.

We use these techniques along with a toy cat-mouse game to evaluate the accuracy of three methods for eliciting discrete probabilities from a fixed structure: (1) direct numerical elicitation, (2) the probability wheel (Spetzler & Staël von Hostein 1975), and (3) the scaled probability bar (Wang

& Druzdzel 2000). We use mean squared errors between the learned and the elicited probabilities to evaluate the accuracy of the different methods. We show that for our domain, using the scaled probability bar is the most effective and least time-consuming procedure for probability elicitation.

In the following sections, we first give a brief review of the existing evaluation schemes of probability elicitation. Then we present the assumptions and relevant equations that allow us to capture a subject's beliefs in the form of learned network parameters. Next, we describe the cat-mouse game that we used to train our subjects and collect data for learning. Finally, we present our experimental design and results followed by a discussion of our findings.

Evaluation Schemes of Probability Elicitation Methods

Eliciting probability of a proposition from an expert amounts to obtaining the expert's subjective degree of belief in that proposition. There is a vast amount of behavioral decision theory literature covering this topic. Due to space limitations, we will not discuss these methods in this section. For a detailed review, see for example (Merkhofer 1987; Morgan & Henrion 1990).

The difficulty in evaluating elicitation methods is that the true model is needed in order to be compared to the elicited model. Obviously, since the former is encapsulated in the expert's mind, it is not readily available for comparison. Previous comparisons of elicitation schemes followed essentially three lines of reasoning: (1) expert's preference, (2) benchmark model, and (3) performance measure.

The first approach, *expert's preference*, is based on the assumption that when an elicitation method is preferred by the expert, it will yield better quality estimates. While this assumption is plausible, to our knowledge it has not been tested in practice. There are a variety of factors that can influence the preference for a method, such as its simplicity, intuitiveness, or familiarity and these factors are not necessarily correlated with accuracy.

The second approach, *benchmark model*, compares the results of elicitation using various methods against an existing benchmark (gold standard) model $\widehat{M}$ of a domain (or a correct answer that is assumed to be widely known). Accuracy is measured in terms of deviation of the elicited model from $\widehat{M}$. For example, in Lichtenstein *et al.*'s (1978) study of people's perception of frequencies of lethal events, there was a readily available collection of actuarial data on those events. Similarly, In Price's (1998) study on effects of a relative-frequency elicitation question on likelihood judgment accuracy, general knowledge was used. An important assumption underlying the benchmark model method is that the model $\widehat{M}$ is shared by all experts. While in some domains this assumption sounds plausible, human experts notoriously disagree with each other (Morgan & Henrion 1990; Cooke 1991), and an experimenter is never sure whether the model elicited is derived from a gold standard model or some other model in the expert's mind. A debiasing training of experts with an established knowledge base may help to establish a benchmark model among them. For example, Hora *et al.*(1992) trained their subjects in a formal probability elicitation process directed toward assessing the risks from nuclear power generating stations and compared two elicitation methods for continuous probability distributions. Their subjects were scientists and engineers who quite likely possessed extensive background knowledge about the risks. Effectively, it is hard in this approach to make an argument that the elicited model is close to the experts' actual knowledge, as the latter is simply unknown.

The third approach, *performance measure*, takes a pragmatic stand and compares the predictive performance of models derived using various methods. This reflects, in practice, how well calibrated the expert's knowledge is (Lichtenstein, Fischhoff, & Philips 1982). An example of this approach is the study performed by van der Gaag *et al.*(1999), who used prediction accuracy to evaluate their probability elicitation method in the construction of a complex influence diagram for cancer treatment. While it is plausible that the quality of the resulting model is correlated with the accuracy of the elicitation method, this approach does not disambiguate the quality of the expert's knowledge from the quality of the elicitation scheme. A model that performs well can do so because it was based on superior expert knowledge, even if the elicitation scheme was poor. Conversely, a model that performs poorly can do so because the expert's knowledge is inferior, even if the elicitation scheme is perfect.

The next section introduces an evaluation method that we believe does not suffer from the problems identified with the existing evaluation schemes.

Datamining Expert Beliefs

To evaluate the accuracy of an elicitation method is to make a judgment about how good the elicited model reflects the expert's real degree of personal belief. The closer the elicited model reflects the expert's real beliefs, the more accurate we say the method of elicitation is. But how can we measure an expert's real degree of personal belief? What can be used as a standard to evaluate the accuracy of a subjective probability? What we need is a method to capture the knowledge/beliefs that are held by our expert, then we need a method to construct a model entailed by that knowledge.

On the other hand, if we have a set of records in the form of a database, there are many machine-learning algorithms that are available to learn various types of models from that database. In this section we will present the theory needed to learn probabilistic network models from data. However, the method that we describe in this paper is general enough to be used on any types of models for which there exist statistical methods for learning.

Capturing the Expert's Knowledge

Complicating this effort is the fact that a person becomes an expert from a novice in a process of acquiring knowledge from a wide array of sources. Sources of knowledge range from reading books, talking to other experts, and most importantly for us, to observing a series of instances in the real world. In the method that we are proposing, we *create* an

expert in a particular toy domain. In the process, we confine the source of knowledge available to that expert to be strictly of the latter type; namely, a series of observations of the real world. Being assured that our expert accumulates only this knowledge allows a particularly simple analysis of what our expert's beliefs about the domain should be. Throughout the paper we will refer to this type of knowledge as *observational knowledge*.

If we assume that we have an expert whose entire knowledge of a domain is observational, then the expert's knowledge can be viewed as originating entirely from a database, D_{exp}, of records filled with instances of the domain our expert has committed to memory. If we further assume that we have recorded all relevant instances of the domain that our expert has actually observed into a database D, then our database D will be identical to D_{exp} under the assumption that the subject has paid attention to the occurrence of each event during his or her observation process. Thus, in any experiment designed to measure D_{exp}, it will be important to incentivate the subject in some way to pay attention to all events in the world.

Learning Parameters From Data

Assuming that we can assess D_{exp} correctly, we must now construct a probabilistic model that is most consistent with that data. Much work has been done on this problem in recent years. We will present just the key results of some of this work here. A good review of the literature can be found in (Heckerman 1998).

Bayesian methods (Cooper & Herskovits 1992) for learning a probabilistic model over a set of variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, assume that the learner begins with a set of prior beliefs governing the domain. In the case of an unrestricted multinomial distribution, each variable X_i is discrete, having r_i possible values $x_i^1, \ldots, x_i^{r_i}$, where $i = 1, \ldots, n$. In this case, it is assumed for convenience that the priors take the form of a Dirichlet distribution, having parameters α_{ijk}. One common sense interpretation of α_{ijk} in a Bayesian network capturing this domain is that it is the number of times an expert has observed variable $X_i = x_i^k$ when the parents of X_i achieved the jth configuration: $Pa_i = pa_i^j$. As a bit of notation, we define θ_{ijk} to be the true probability that $X_i = x_i^k$ given that $Pa_i = pa_i^j$. In other words, it is the conditional probability parameter corresponding to the α_{ijk}. We use $\boldsymbol{\theta}_{ij} = \{\theta_{ijk} | 1 <= k <= r_i\}$ to denote the conditional probability distribution of X_i under the jth parent configuration. We assume *parameter independence*, which states that $\boldsymbol{\theta}_{ij}$ is independent of $\boldsymbol{\theta}_{ij'}$ for all $j \neq j'$.

Given a network structure S, a complete data set D without any missing data, a set of Dirichlet prior parameters α_{ijk}, and the assumption of parameter independence, it can be shown that the expected value of the parameters of the network with respect to the posterior distribution $p(\boldsymbol{\theta}_{ij} | D, S, \boldsymbol{\alpha}_{ij})$ can be expressed as:

$$\hat{\theta}_{ijk} = \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}, \qquad (1)$$

where N_{ijk} are the number of times in D that the variable X_i took on value x_i^k when the parents of X_i took on configuration pa_i^j, $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$, and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

For a domain where the expert-to-be has little or no previous experience, we assume that all α_{ijk} are equal and small. Under this assumption, when no data are present for a particular (i, j) configuration of the world (i.e., $N_{ij} = 0$), then the N_{ijk} terms drop out of Equation 1 and the small equal priors produce a uniform distribution. However, even if a small amount of data is involved, then the priors have little influence on the parameters learned.

Evaluating Elicitation Schemes with a Toy Virtual World

We designed a game in which a subject can move a cat to capture a mouse. We recorded the state changes of the cat-mouse game during the game playing process. What each subject experiences is unique and depends on the subject's actions. The recorded data allows for the learning of the probabilistic model of the toy world as seen by the subject. This learned model in turn gives us a standard by which to measure the accuracy of the model elicited from the subject.

The Cat and Mouse Game: A Toy Virtual World

Our toy world includes three characters: a cat and two mice. The objective of the game is for the cat to capture a mouse. There are twelve possible positions indicated by the grid cells in a horizontal line (see Figure 1). The cat can move one cell at a time between the current cell and either adjacent cell. One and only one mouse is present at any given time, and it can only bounce back-and-forth between two positions on each side of the screen. The two special positions for the mice are called *left-pos* and *right-pos* respectively. When the cat enters the cell/position where the mouse is located, it catches the mouse and the game is over.

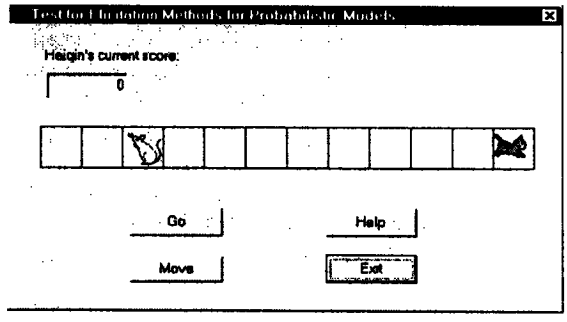


Figure 1: A screen snapshot of the cat-mouse game

The two mice are characterized by a color: *yellow* or *grey*. The cat can be in one of the four states: *normal, angry, frustrated*, and *alert*. Four figures are used to represent the states of the cat. Table 1 and 2 illustrate the figures we used in the game.[1]

[1]Our experimental subjects only saw the figures as the representation of the cat's states and mouse color. The verbal expressions

Table 1: Yellow mouse and grey mouse

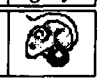
yellow	*grey*
	

Table 2: Four states of the cat

normal	*angry*	*frustrated*	*alert*
			

Two buttons, labeled *move* and *go* respectively, are provided for the subject to manipulate the position of the cat. After the subject clicks a button, the cat moves to either the left or the right. Its moving direction is uncertain and depends on the current state of the world (i.e., which mouse is present, the position of the mouse, the state of the cat, and which button the subject has clicked). There is a short delay (half a second in our experiment) between button clicks during which the buttons are disabled. This prevents the subject from clicking the buttons too frequently and paying little attention to probabilistic relationships among the variables. It allows the subject to have enough time to observe how the moving direction of the cat is influenced by the state of the world and the subject's own actions.[2]

After this delay, the toy world is updated to a new state. One mouse may disappear and another may show up instead. The mouse may appear in a different position. The cat may change its state. The two buttons for the subject's action become enabled.

In the beginning, the yellow mouse is put in the *left-pos* position. The cat is put in the farthest position away from the mouse. After the cat has caught a mouse, the game ends and a new round of the game begins. A new game always begins with the same initial positions for both the mouse and the cat. But the states of the rest of the world are uncertain.

Scoring rules are adopted to encourage the subject's involvement in the game. Whenever the cat captures a mouse, the subject's score increases as an incentive. Also, the game emits a celebratory sound as a reward for the subject.

are used to encode the cat's states and mouse color in the Bayesian network for the cat-mouse world due to the restraint of the modeling environment. These labels, "normal", "angry", etc., were not provided to the subjects during game play but were used, together with the pictures, to identify the states of the cat during the elicitation process.

[2]The delay length of the disabled state of the buttons was selected based on our experiments with pilot subjects. We first tried 1 second and 2 seconds as the delay, but our pilot subjects soon complained the delay was too long and made the game boring. So we selected the maximum delay (half a second) with which the subjects still felt comfortable.

The Bayesian Network for the Cat-mouse World

The cat-mouse world is based on a simple Bayesian network (Figure 2) consisting of five variables, *Action*, *Mouse Color*, *Mouse Position*, *Cat State*, and *Cat Moving Direction*.

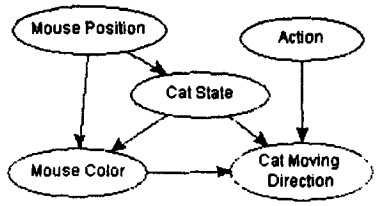


Figure 2: The Bayesian network of the cat-mouse world

Variable *Action* with two outcomes, *move* and *go*, models the observed subject's action. *Mouse Color* which could be *yellow* and *grey*, defines which of the two mice is present. *Mouse Position* indicates the current position of the present mouse: *left-pos* and *right-pos*. *Cat State* represents four possible states of the cat: *normal*, *angry*, *frustrated*, and *alert*. The last variable *Cat Moving Direction* reflects the moving direction of the cat in the current step. Two directions are defined, *left*, and *right*.

The five variables influence each other probabilistically. The states of the variables change at each step according to the probabilities encoded in the network. Their probability distributions, either prior or conditional, were assigned randomly when the network was built to avoid biases to a particular probability distribution. One exception is the probability distribution of the *Action* node. The value of the *Action* node is always instantiated to the state that corresponds to the subject's action, and hence, the prior probability distribution becomes irrelevant. We chose the two nearly identical action words, *move* and *go*, to avoid any semantic difference which could have a potential influence on the subjects' preference.

The State Change of the World by Sampling

After the subject has clicked a button to take an action, the state of the world and the cat's moving direction are updated. The new states are selected by generating a stochastic sample on the cat-mouse network following the partial parent order of the graph. We use probabilistic logic sampling (Henrion 1988) to generate node states on the basis of their prior probabilities of occurrence. By choosing more likely states more often, we simulate the state changes of the toy world. The subjects are exposed to changes in the world that correspond to the underlying joint probability distribution and their actions.

Collecting Data for Expert's Knowledge

Every time the state of the toy world changes, it is recorded automatically. In our data set, a case consists of the outcomes of all of the five variables encoded in the cat-mouse Bayesian network. The database of a subject's experience is assumed to contain all states of the world that the subject has seen. It is the subject's observational knowledge about

the toy virtual world. This knowledge comes completely from the subject's game-playing experience. Therefore, the records constitute a perfect data set for learning the subject's knowledge about the cat-mouse domain.

Experimental Design

We demonstrated our method in an experimental study that investigates the effectiveness of three elicitation methods: asking for numerical parameters directly, translating graphical proportions by using the probability wheel, and using the scaled probability bar. We used the graphical modeling system *GeNIe* (GeNIe 1999) and build a module of cat-mouse game in *GeNIe* as well.

Subjects

The subjects were 28 graduate students enrolled in an introductory decision analysis course at the University of Pittsburgh, who received partial course credit for their participation.

Design and procedure

The subjects were first asked to read the instructions from a help window that introduced the game characters and the game rules. Also, they were asked to pay attention to the probabilistic influences from the state of the toy world and their action choice to the direction of the cat's movement. The subjects were told that knowledge of these probabilistic relationships would help to improve their performance. To motivate the subjects to perform well, extra credit was offered for higher scores in the cat-mouse game and lower errors of estimates of the probabilities in elicitation.

Each trial included two stages. The subjects first played the cat-mouse game for 30 minutes. The data about their experienced states of the toy virtual world were automatically recorded. The data sets in our experiment typically contained between 400 and 800 records.

The second stage involved probability elicitation by each of the three elicitation methods. The subjects were shown the Bayesian network structure in Figure 2 and were asked to estimate the conditional probability table (CPT) for the node *Cat Moving Direction* by

1. typing the numerical parameters directly in conditional probability tables, and
2. giving graphical proportions in the probability wheel, and
3. giving graphical proportions in the scaled probability bar.

We applied here a within-subject design in which each subject used the three elicitation methods. To offset the possible carry-over effects, we counterbalanced the order of method usage across our subjects.

The CPT elements θ_{ijk} elicited were compared to $\hat{\theta}_{ijk}$, the CPT elements learned by applying Equation 1 to the subjects' acquired data. The mean-squared error (MSE) of the parameters was calculated as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\theta_{ijk} - \hat{\theta}_{ijk})^2$$

In order to evaluate the speed of the elicitation methods, we also recorded the time taken for each elicitation procedure.

Results

Figure 3 plots the mean squared errors of the three elicitation methods when compared to the learned probabilities. The plot also shows the times spent on elicitation for each of the three methods.

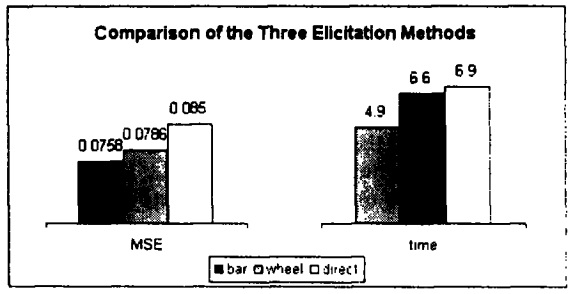


Figure 3: MSE($\alpha = 5$) and elicitation time for each of the three methods tested

For each pair of elicitation methods, we conducted one-tailed, paired sample t test for comparison of accuracy and time. The t tests showed that scaled probability bar performed significantly better than direct numerical elicitation ($p = 0.03$ for MSE and $p = 0.007$ for time). Probability wheel was marginally better than direct numerical elicitation ($p = 0.07$ for MSE) but did not improve the time compared with direct numerical assessment ($p = 0.37$). However, probability wheel was almost as accurate as scaled probability bar. Even though the latter had a slightly lower MSE, the difference was not statistically significant ($p = 0.19$).

Discussion

One objection that could be raised to our technique is that in a thirty-minute training session the trainees used in our experiment probably do not achieve truly proficient expert status. This would be a key objection if we were comparing the elicited models to the *original* model underlying the toy-world; however, the main point in using the trainees' actual acquired knowledge is to deflect this criticism: we are comparing the elicited model precisely to the knowledge that we know our trainee has observed. In principle this technique should work regardless of the expertise of the trainee. Nonetheless, we acknowledge that there may be some transition during the process of achieving true expertise which alters the trainees' elicitation behavior. We assume that these effects will affect the elicitation techniques in a uniform way, so that the relative assessment of elicitation techniques is not affected.

It may be that the effectiveness of different elicitation techniques varies from expert to expert. In that case, our evaluation technique can provide a relatively quick and effective way to judge which elicitation procedure is most effective for a given expert. The expert can quickly be trained

on a toy model, and then our experimental procedure can be used to decide which elicitation technique is most effective for that particular expert.

Conclusion and Future Research

We proposed a method that allows for objective evaluation of methods for the elicitation of probability distributions for probabilistic models. Our method is based on machine learning the expert's beliefs when data of the expert's learning knowledge are available. We illustrated the evaluation approach with a toy virtual world and evaluated three elicitation methods for probabilities: direct numerical elicitation, the probability wheel, and the scaled probability bar. Based on the results of our experiment, we concluded that the probability wheel and the scaled probability bar both performed better than direct numerical elicitation. The scaled probability bar was the most efficient in terms of being most accurate and taking the least time. Our conclusion supports the proposition that graphical tools are useful in eliciting experts' beliefs.

One interesting possibility for future work is to apply our method to experts in real domains and form a baseline for determining the capability of a particular expert to generate a model using a particular elicitation method. In other words, we can use our technique to discover the most effective elicitation schemes for a given expert.

Though we only deal with comparison of probability elicitation methods in this paper, our evaluation scheme can be equally useful for comparing structure elicitation methods for probabilistic models. The elicited structure of the Bayesian network can be compared to the structure learned from the users' data using machine learning techniques for Bayesian networks.

Acknowledgements

The research was supported by the Air Force Office of Scientific Research under grant F49620-00-1-0122, and the National Science Foundation under Faculty Early Career Development (CAREER) grant IRI-9624629. We are grateful to our colleagues in the Decision Systems Laboratory for their suggestions about the experiment design and their willingness to be pilot subjects in our experiments. Special thanks go to Dr. Michael Lewis for his encouragement and Dr. Satish Iyengar for his advice on statistics.

References

Cooke, R. 1991. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press.

Cooper, G. F., and Herskovits, E. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9(4):309–347.

Druzdzel, M. J., and van der Gaag, L. C. 2000. Building probabilistic networks: "Where do the numbers come from?" guest editors' introduction. *IEEE Transactions on Knowledge and Data Engineering* 12(4):481–486.

Estes, W. 1976. The cognitive side of probability learning. *Psychological Review* 83(1):37–64.

GeNIe. 1999. GeNIe: A development environment for graphical decision-theoretic models. Available at http://www2.sis.pitt.edu/~genie.

Heckerman, D. 1998. Bayesian networks for data mining. *Data Mining and Knowledge Discovery* 1(1):79–119.

Henrion, M. 1988. Propagating uncertainty in Bayesian networks by probabilistic logic sampling. In Kanal, L.; Levitt, T.; and Lemmer, J., eds., *Uncertainty in Artificial Intellgience 2*. New York, N. Y.: Elsevier Science Publishing Company, Inc. 149–163.

Hora, S. C.; Hora, J. A.; and Dodd, N. G. 1992. Assessment of probability distributions for continuous random variables: A comparison of the bisection and fixed value methods. *Organizational Behavior and Human Decision Processes* 51:133–155.

Lichtenstein, S.; Slovic, P.; Fischhoff, B.; Layman, M.; and Combs, B. 1978. Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory* 4(6):551–578.

Lichtenstein, S.; Fischhoff, B.; and Philips, L. 1982. Calibration of probabilities: The state of the art to 1980. In *Judgement under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

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

Morgan, M. G., and Henrion, M. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge: Cambridge University Press.

Price, P. C. 1998. Effects of a relative-frequency elicitation question on likelihood judgment accuracy: The case of external correspondence. *Organizational Behavior and Human Decision Processes* 76(3):277–297.

Spetzler, C., and Staël von Holstein, C.-A. 1975. Probability encoding in decision analysis. *Management Science* 22:340–358.

van der Gaag, L.; Renooij, S.; Witteman, C.; Aleman, B.; and Taal, B. 1999. How to elicit many probabilities. In *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–99)*, 647–654. San Francisco, CA: Morgan Kaufmann Publishers.

Wang, H., and Druzdzel, M. J. 2000. User interface tools for navigation in conditional probability tables and graphical elicitation of probabilities in Bayesian networks. In *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000)*, 617–625. San Francisco, CA: Morgan Kaufmann Publishers.