AHFD

Aviation Human Factors Division
Institute of Aviation

University of Illinois
at Urbana-Champaign
1 Airport Road
Savoy, Illinois 61874

Developing a Methodology for Eliciting Subjective Probability Estimates During Expert Evaluations of Safety Interventions: Application for Bayesian Belief Networks

Douglas A. Wiegmann

Final Technical Report
AHFD-05-13/NASA-05-4

October 2005

Prepared for

NASA Langley Research Center
Hampton, VA

Contract NASA NNL04AA50G

Abstract

The NASA Aviation Safety Program (AvSP) has defined several products that will potentially modify airline and/or ATC operations, enhance aircraft systems, and improve the identification of potential hazardous situations within the National Airspace System (NAS). Consequently, there is a need to develop methods for evaluating the potential safety benefit of each of these intervention products so that resources can be effectively invested to produce the biggest benefit to flight safety. Of interest to the present project is the *process* of using expert judgments to develop Bayesian Belief Networks (BBN's) that model the potential impact that specific interventions may have. Specifically, the present report summarizes methodologies for improving the elicitation of probability estimates during expert evaluations of AvSP products for use in BBN's. The work involved joint efforts between Professor James Luxhoj from Rutgers University and researchers at the University of Illinois. The Rutgers' project to develop BBN's received funding by NASA under contract NAS1-03057 entitled "Probabilistic Decision Support for Evaluating Technology Insertion and Assessing Aviation Safety System Risk." The proposed project was funded separately but supported the existing Rutgers' program.

PROJECT DESCRIPTION

The NASA Aviation Safety Program (AvSP) has defined several products that will potentially modify airline and/or ATC operations, enhance aircraft systems, and improve the identification of potential hazardous situations within the National Airspace System (NAS). Consequently, there is a need to develop methods for evaluating the potential safety benefit of each of these intervention products so that resources can be effectively invested to produce the biggest benefit to flight safety. The present project assisted in the evaluation AvSP interventions through joint efforts between researchers in Professor James Luxhoj's laboratory at Rutgers University and researchers at the University of Illinois. The Rutgers' project focused heavily on the engineering components of Bayesian Belief Networks (BBN's) that model the potential impact that specific interventions may have on safety. The Rutgers' project is currently funded by NASA under contract NAS1-03057.

The development of BBN's, however, also involves a major psychological component, including experts' subjective beliefs that can be biased by a variety of factors. Of interest to the present project, therefore, is the *process* of eliciting expert judgments for use in the development and subsequent validation of BBN's. In this report we briefly summarize the potential problems associated with expert probability estimates and then provide a review of methods designed to overcome these problems. Finally, we present a tool developed as part of this project that integrates many of these improved elicitation methods.

Brief Overview of Bayesian Belief Networks

Bayesian Belief Networks were developed for representing and reasoning with uncertainty (Pearl, 1988) and have come into widespread use for decision support. These networks represent knowledge both qualitatively and quantitatively. The qualitative component consists of a directed graph, indicating which variables in the domain of interest influence others (Renooij, 2001). The variables are represented by nodes, and their influences on other variables are represented by arcs connecting the nodes. The quantitative part consists of the probabilities that a variable will assume each of its possible values, conditioned on the values of each of the variables that directly influence the variable of interest (Renooij, 2001). The probabilities represent the magnitudes of each variable's influence. The qualitative structure of the network is determined with the help of domain experts. Some of the probabilities required for the quantitative part of the network can be determined by reference to databases and scientific literature (Renooij, 2001). However, many of the conditional probabilities required to quantify a belief network cannot be derived from those sources, so the probabilities must be elicited from domain experts, based on their knowledge and experience (Druzdzel & van der Gaag, 1995).

There are many potential problems associated with eliciting probabilities from experts, which most often center on the issue of bias. The biases encountered during probability elicitation can be classified as either motivational, leading to the overconfidence bias, in which experts think they should be more certain about effects than they really are, or cognitive, due to the thought processes the experts use (Skinner, 1999; Renooij, 2001). Cognitive biases are often caused by experts' use of the heuristics of availability, in which probability is determined based on how easy it is to recall events from memory; anchoring, in which probability is determined by starting with a pre-set value and adjusting up or down from it; representativeness, in which the

probability of one event leading to another is assessed based on the similarity between the two events; and control, in which the probability of an event occurring is based on the falsely perceived control over its likelihood (Kahneman et al., 1982). The representativeness heuristic can result in biases due to the conjunction fallacy, the gambler's fallacy, and base-rate neglect. Other biases that may affect elicited probabilities are overestimation of single-event probabilities, conservatism, optimism, and fallacies of causal and diagnostic reasoning, in which causal data and inferences are given more weight than diagnostic data and inferences (Fenton, 1998).

Methods for Improving Probability Elicitation

There is a wide variety of elicitation methods designed to suppress or eliminate the biases associated with directly stating numerical probabilities. In this section, we will summarize the strengths and weaknesses of each approach.

Frequency Estimation Method. Stating probability elicitation questions in frequency format, in which experts are asked to state the number of times out of some multiple of 10 that they would expect an event to occur, given conditions set by the variables that influence the variable in question, suppresses overconfidence, base-rate neglect, the conjunction fallacy, control bias, and overestimation of single-event probabilities (Anderson, 1998; Gigerenzer & Hoffrage, 1995). Probability scales allow experts to mark probabilities on a graphic scale, which are fast and easy to understand; however, they tend to be inaccurate and prone to scaling biases (Renooij, 2001; von Winterfeldt & Edwards, 1986). In an attempt to remedy this situation, Renooij and Witteman (1999) developed a scale with numerical anchors on one side and verbal probability anchors on the other side to allow experts to use the scale with which they are most comfortable for each question.

Gamble Methods. Probabilities can also be determined using two gamble-like methods. In the certain-equivalent method, the expert chooses either a certain payoff or a lottery where the payoff depends on the probability in question, and the elicitor adjusts the amount of the certain payoff until the expert is indifferent between the two choices. In the lottery-equivalent method, the expert chooses either a lottery where the outcome depends on a probability set by the elicitor or a lottery where the outcome depends on the probability in question. Gamble-like methods suffer from a high time cost, can be hard to conceive due to rare or unethical hypothetical situations that must be considered in some cases, and the certain-equivalent method is subject to risk attitude effects, which are reduced in the lottery-equivalent method (Renooij, 2001). Another gamble-like method, the probability wheel, is a pie chart with a spinnable pointer and red and green sections that the elicitor adjusts until the expert thinks that the probability of the pointer landing in the red section is equivalent to the probability in question. There are no risk attitude effects, but since the method is very close to direct estimation, it may be subject to the same biases, and the time cost may be too high (Renooij, 2001).

Hierarchical Methods. Druzdzel and van der Gaag (1995) developed a method to allow experts to provide either qualitative or quantitative information, whichever they were most comfortable providing. They then used the information to define a system of (in)equalities limiting the set of possible joint probability distributions and derived second-order probability distributions to determine the most likely true joint distribution. This method avoids biases

because experts are not forced to provide numerical probabilities, and it allows detection of inconsistent information that can be refined with further elicitation. Monti and Carenini (2000) adapted the Analytical Hierarchy Process for use in belief network probability elicitation. Probabilities are derived from comparisons of the likelihood of each possible pair of events, bypassing direct elicitation biases. Because of the redundancy of the method, it is easy to compute the consistency of the expert's responses and provide immediate feedback to the expert for refinement. However, the number of comparisons required far exceeds the number of probabilities to be assessed, many events are so different that they are hard to compare, and it is hard to determine the acceptability of the expert's consistency based on the statistical methods used (Monti & Carenini, 2000; Renooij, 2001).

A few studies have compared some of these varied methods for eliciting expert judgments. For example, Wang, Dash, and Druzdzel (2002) compared direct numerical elicitation with the probability wheel and the scaled probability bar and found that accuracy and speed were highest with scaled probability bars, followed by the probability wheel. Whitcomb, Önkal, Benson, and Curley (1993) found that for direct numerical elicitation, odds, and the probability wheel, test-retest reliability was high within and between all of the methods. The results of existing studies have often been inconsistent, however. Indeed "What is lacking are large multi-method studies where experts are asked to assess a large number of probabilities with every single method" (Renooij, 2001, p. 268). There is a need to determine which methods are best, or whether certain methods work better in different contexts (e.g., individual or group settings).

Using Multiple Experts

Using more than one expert for probability elicitation is believed to increase the accuracy of the final probabilities by balancing multiple viewpoints and drawing from a larger pool of knowledge. The two major ways of combining the probabilities from multiple experts are aggregating individual assessments and group consensus. In most cases, aggregating by simple averaging works well, but more complex modeling rules can be applied if information about the quality of and dependence among the experts' assessments is available (Clemen & Winkler, 1999). Information about the quality of assessments can be collected using seed variables, for which the true probabilities are known (Roelen, Wever, Hale, Goossens, Cooke, Lopuhaä, Simons, & Valk, 2002). However, caution should be used, since the sensitivity of complex modeling rules can eliminate their advantages over simple averaging and often results in worse performance (Clemen & Winkler, 1999). Expert consensus has the advantage over aggregating probabilities from individual assessments that the experts share their knowledge, and Clemen and Winkler (1999) conclude that this method works almost as well as mathematical aggregation. However, expert consensus introduces potential problems and biases associated with group interaction (Renooij, 2001). Clemen and Winkler (1999) suggest that eliciting and aggregating individual assessments after group interaction, rather than forcing the group to come to consensus, allows knowledge sharing without group interaction problems.

Further work is still needed to determine the best method for assessing probabilities from multiple experts for belief networks. None of the studies reviewed by Clemen and Winkler (1999) involved field application of the techniques studied, and the authors were only able to outline general rules concerning which techniques to use in the field. Roelen et al. (2002) used

the Classical Model to combine expert judgments because it conformed to the rules they set forth for expert judgment techniques: scrutability, performance control, neutrality, and fairness, though it would seem that many other combination techniques would meet those requirements, as well. Thus, empirical results from field studies comparing all of the combination techniques feasible for use in Bayesian Belief Networks are needed to determine which is best or to define explicit guidelines for determining which to use in any given situation.

Another potential problem in selecting and using experts is their nature of expertise, which is generally not addressed in the literature, given the domains usually studied. In the context of aviation and aviation technologies, particularly when human factors issues are involved, criteria for determining expertise may vary. Is it better to use experts in industry and operations or to use experts in human factors and technologies affecting human performance, or both? Winkler and Poses (1993) compared all possible simple average combinations of probabilities elicited from four individuals working in an intensive care unit on the probability of survival of each patient and found that the best combination was the two most experienced, yet least similar people in their area of expertise. Clemen and Winkler (1999) conclude that it is best to use experts who differ from each other in terms of viewpoint and knowledge to minimize redundant information and maximize the effectiveness of aggregation. They also conclude that the optimum number of experts to use is three to five. Therefore, it seems that using experts from both the industry/operations and human factors areas will provide the best results.

Transitioning to Field Applications

Most of the research on probability elicitation for Bayesian Belief Networks discuss the process in conceptual form or detail the methods for experimental implementation, but few provide a protocol for field implementation. Renooij (2001) gives a five-stage process and explains what is involved with each stage and the various elicitation methods that can be used but does not enumerate specific procedures and methods that should be used. Roelen et al. (2002) give a protocol used for building an aviation safety model and explain in detail how they performed each step. They report that the protocol resulted in a workable belief network but were somewhat disappointed in the poor calibration of the individual experts. This may be due in part to their elicitation method, which forced the experts to give frequencies for the 5th, 50th, and 95th quartiles. Though the questions and the answers were generally in frequency format, the use of quartiles, asking the experts to determine values they were 5%, 50%, and 95% certain were greater than the actual value, made the technique vulnerable to the biases associated with direct numerical elicitation of probabilities (Gigerenzer & Hoffrage, 1995; Wang, Dash, & Druzdzel, 2002). To date, there have been few field studies to determine the best way to apply the methods explained in the literature to field use, including the best methods for probability elicitation and combining the experts' judgments.

The Rutgers' Elicitation Tool

Researchers at Rutgers University in consultation with the human factors team at the University of Illinois developed The *GRID Feedback Workbook* to aid in the elicitation of probabilities by aviation safety experts. Specifically, The tool was created to elicit the judgment of the experts as they envisioned the AvSP products influencing the risk level of the various accident precursors used in a risk management model.

The workbook is arranged with precursors down the left-hand column and products listed across the top row (see Figure 1). All precursor names and technology names are hyperlinked to a definition page in the workbook. The precursors are defined and a simple example given in many instances for clarity. The technology products are defined from the AvSP Product Dictionary and most of the definitions in the GRID Feedback Workbook contain graphic PowerPoint slides for fluency. Each of the technology products has two entry cells for the precursors, one for direct effect and one for indirect effects. (Due to the absence of a system structure, such as the BBN model, the effects must be elicited as direct and indirect to represent parent and grandparent nodes.)

Figure 1. Sample Page from the Rutgers' *GRID Feedback Workbook*.

Within the workbook, products are grouped according to the suites as outlined in the AvSP. Users of the workbook complete the GRID for their particular product expertise. They are instructed to:

"Enter a number between 0-1 in the cell intersection of AvSP Product and Causal Factor to represent the projected impact on safety risk reduction. Both Direct and Indirect Effects can be considered, but they will be scored separately. The rank for a product's effect on each causal factor and accident type is based upon your estimate of a relative risk reduction (i.e., if a product will reduce likelihood of a causal factor occurring by 25%, rating = 0.25)."

This tool was specifically designed to integrate a variety of methods for improving the elicitation probabilities described previously. For example, it provides instructions that restate probability elicitation questions in frequency format, which has shown to suppress overconfidence, base-rate neglect, the conjunction fallacy, control bias, and overestimation of single-event probabilities (Anderson, 1998; Gigerenzer & Hoffrage, 1995). Second, it uses numerical anchors on one side and verbal probability anchors on the other side to allow experts to use the scale with which they are most comfortable for each question. This approach has been shown to reduce scaling biases (Renooij, 2001; von Winterfeldt & Edwards, 1986; Renooij & Witteman 1999). In addition, the tool allows for the use of more than one expert for probability elicitation, which is believed to increase the accuracy of the final probabilities by balancing multiple viewpoints and drawing from a larger pool of knowledge (Clemen & Winkler, 1999). Finally, the tool allows individual experts to generate estimates independently rather than attempting to achieve consensus as a group. As stated earlier, expert consensus can introduce potential problems and biases associated with group interaction (Renooij, 2001). Hence, eliciting and aggregating individual assessments after group interaction, rather than forcing the group to come to consensus, allows knowledge sharing without group interaction problems (Clemen & Winkler, 1999). As such, this *GRID Feedback Workbook* should serve as a valuable tool for generating reliable estimates from experts concerning the impact that each NASA product may have on improving aviation safety in the future.

Conclusion

The NASA Aviation Safety Program (AvSP) has defined several products that will potentially modify airline and/or ATC operations, enhance aircraft systems, and improve the identification of potential hazardous situations within the National Airspace System (NAS). Consequently, there is a need to develop methods for evaluating the potential safety benefit of each of these intervention products so that resources can be effectively invested to produce the biggest benefit to flight safety. Of interest to the present project is the *process* of using expert judgments to develop Bayesian Belief Networks (BBN's) that model the potential impact that specific interventions may have. Specifically, the present report summarizes methodologies for improving the elicitation of probability estimates during expert evaluations of AvSP products for use in Bayesian Belief Networks. A specific tool for generating expert estimations developed as part of collaborative efforts between engineers at Rutgers University and human factors researchers at the University of Illinois.

REFERENCES

Anderson, J. L. (1998). Embracing uncertainty: The interface of Bayesian statistics and cognitive psychology. *Conservation Ecology* (online) *2*(1), (http://www.consecol.org/vol2/iss1/art2).

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis, 19*, 187-204.

Druzdzel, M. J., & van der Gaag, L. C. (1995). Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 141-148.

Fenton, N. (1998). Probability elicitation and bias. (http://www.dcs.qmul.ac.uk/~norman/BBNs/BBNs.htm).

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684-704.

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

Monti, S., & Carenini, G. (2000). Dealing with the expert inconsistency in probability elicitation. *IEEE Transactions on Knowledge and Data Engineering, 12*(4), 499-508.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Francisco: Morgan Kaufmann Publishers.

Renooij, S. (2001). Probability elicitation for belief networks: Issues to consider. *The Knowledge Engineering Review, 16*(3), 255-269.

Renooij, S, & Witteman, C. L. M. (1999). Talking probabilities: Communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning, 22,* 169-194.

Roelen, A. L. C., Wever, R., Hale, A. R., Goossens, L. H. J., Cooke, R. M., Lopuhaä, R., Simons, M., & Valk, P. J. L. (2002). *Causal modeling of air safety: Demonstration model* (Tech Report NLC-CR-2002-662). National Aerospace Laboratory NLR.

Skinner, D. C. (1999). *Introduction to decision analysis.* Gainesville, FL: Probabilistic Publishing.

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

Wang, H., Dash, D., & Druzdzel, M. J. (2002). A method for evaluating elicitation schemes for probabilities. *IEEE Transactions on Systems, Man, and Cybernetics, 32*(1), 38-43.

Whitcomb, K. M., Önkal, D., Benson, P. G., & Curley, S. P. (1993). An evaluation of the reliability of probability judgments across response modes and over time. *Journal of Behavioral Decision Making, 6,* 283-296.

Winkler, R.L., & Poses, R.M. (1993). Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Management Science, 39,* 1526-1543.