Pergamon

*Computers in Human Behavior*, Vol. 14, No. 1, pp. 181–194, 1998
© 1997 Elsevier Science Ltd. All rights reserved
Printed in Great Britain
0747-5632/97 $19.00 + 0.00

PII: S0747-5632(97)00039-3

# The Impact of Expert-System-Based Training on Calibration of Decision Confidence in Emergency Management

**Yuan-Liang Su**

*Computer and Communication Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan*

**Dyi-Yih M. Lin**

*Institute of Industrial Engineering, National Chiao Tung University, Hsinchu, Taiwan*

**Abstract** — *In many emergency incidents, human operators need to derive countermeasures based on contingency rules under time pressure. Since people tend to be overconfident regarding their performance levels, it is necessary that the operators be well trained to calibrate proper decision confidence in the safety-related domain. This paper examines the effectiveness of using expert systems to train for the desired calibration. Emergency management of chemical spills was selected to exemplify the rule-based decision task. An expert system in the domain was developed to serve as the training tool. A total of 40 student subjects participated in an experiment in which they were asked to resolve spill scenarios under the manipulation of training and deadline conditions. The experiment*

Request for reprints should be addressed to Dyi-Yih M. Lin, Institute of Industrial Engineering, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 300, Taiwan. E-mail: u8233801@cc.nctu.edu.tw

*results indicate that people tend to overestimate their performance capabilities when reasoning with a rule-based knowledge set, especially with the presence of time constraints. The results also show that the manifestation of overconfidence can be reduced for individuals who undergo the expert-system calibration training. The implications of the findings are examined in this paper.* © *1997 Elsevier Science Ltd*

*Keywords* — expert systems, training, overconfidence

When decision makers complete a decision, one of the most crucial pieces of information they should provide is their confidence in that decision. This point is important because confidence levels usually serve as an indicator of the accuracy of the decisions made and, thus, guide the course of actions (Heath & Tversky, 1991; Wickens, 1992). A critical concern of the accuracy–confidence relationship is related to the concept of calibration. People are said to be calibrated if, over the long term, they are able to assign confidence estimates that equal the real proportion of correct responses.

Research on the appropriateness of confidence has reached a general conclusion that people are poorly calibrated and tend to be overconfident when assessing their performance capabilities (Kleinmuntz, 1990). The phenomenon of overconfidence is indeed a pervasive bias that permeates a wide variety of tasks. Typical examples include general knowledge questions (Arkes, Christensen, Lai, & Blumer, 1987; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein & Fischhoff, 1977, 1980), forecasting (Brown & Murphy, 1987; Fischhoff & MacGregor, 1982), clinical judgment in human behavior (Oskamp, 1965), automotive troubleshooting (Mehle, 1982), and ranking of company stocks (Sen & Boe, 1991).

Despite the widespread manifestation of overconfidence, the calibration issue has been left largely unexplored in a domain where unwarranted judgment of confidence could result in safety-related consequences. This task domain, also the interest of the present study, mainly involves decision making for emergency management of risks. The decision activities in emergency situations are characterized by the need to reach accurate solutions under stringent time pressure (Moray, 1988). These solutions are normally derived from extensive reasoning over a set of knowledge expressed in rule-based form (Johnson & Jordan, 1983). Specifically, the rule-based decision tasks are carried out by recognizing system and/or environment symptoms and associating rules with those symptoms (Rasmussen, 1986).

It has been found that, during their cognitive activities, people are prone to adopt various intuitive strategies to minimize mental workload (Kahneman, Slovic, & Tversky, 1982). These simplifying strategies often make people

assign excessive weight to confirming evidences and disregard contradictory ones, thereby leading to unjustifiably high confidence on the decisions that are, in fact, incorrect (Koriat et al., 1980; Lichtenstein, Fischhoff, & Phillips, 1982). The tendency of relying on the heuristics-based processing of information has been found to become more prevalent when time pressure is present (Payne, Bettman, & Johnson, 1988; Svenson & Mauley, 1993). Therefore, we can extrapolate that, with limited time horizons under emergency circumstances, people will also tend to be overconfident when performing rule-based decision tasks. Hence, our first hypothesis is that people will be overconfident in the rule-based emergency domain.

In order to reduce the common bias of overconfidence, several interventions, emphasizing training with feedback, have been attempted and some of them were at least partially successful. Generally, these debiasing efforts resort to examining feedback in error identification and correction (Kulhavy, 1977) with the aim of reducing overconfidence. For example, Lichtenstein and Fischhoff (1980) succeeded in reducing overconfidence by providing feedback on the proportion of correct answers. Arkes et al. (1987) also found a significant improvement of calibration by providing subjects with feedback regarding the answers to individual questions.

Considering the successful efforts, appropriate calibration of decision confidence in emergency management seems possible if humans are trained with feedback concerning the symptomatic search of rule-based knowledge. Being excellent rule-based reasoners, expert systems (ES) are considered by the present study to serve as the calibration prescription. The viability of using ES-based training for confidence calibration is built on the following rationale: ES, by employing normative artificial intelligence (AI) techniques, are capable of deriving rule-based solutions that are always logically true (Luger & Stubblefield, 1989); furthermore, ES are distinct in being able to make the process of the rule-based reasoning transparent to human users through so-called explanatory justifiers (Hayes-Roth, Waterman, & Lenat, 1983; Luger & Stubblefield, 1989). In other words, the value of the ES justifier is to provide decision makers with explanations concerning the normative search mechanism — essential for accurate and consistent rule-based decision making.

These prominent features enable exploitation of the ES-based decision process and results as feedback, in order to identify and correct possible intuitive decision biases that would cause overconfidence. Our second hypothesis is that ES-based training with normative feedback will provide a resource to which humans can revert when the tendency of heuristic processing of rule-based information is occurring. Such training experience is expected to enhance a person's ability to mediate an appropriate

confidence–accuracy relationship when emergency incidents have to be resolved under extreme time pressure. In short, the present study is aimed at investigating the effectiveness of utilizing ES as a training system for calibration of decision confidence in emergency management.

## METHODOLOGY

### *Independent Variables*

Given the purpose of the present study, training and time pressure were manipulated as independent variables. Training was designed as a between-subjects factor and defined by two treatment levels. The ES group received lines of reasoning generated by an ES. The control group, however, received no such information. Time pressure was examined using two treatment levels and was designed as a within-subjects factor due to people's adaptivity in reacting to deadline conditions (Payne et al., 1988). The subjects in the "no-time-pressure" condition were allowed to complete a task at whatever pace they wished, whereas the subjects in the "time-pressure" condition were required to finish a task within 90 s (this was found to constitute time pressure in a pilot study).

### *Subjects*

Subjects were 40 undergraduate students in industrial engineering at a major university. Participation in the experiment earned credit toward fulfillment of a course requirement. None of the 40 subjects had taken ES/AI-related courses prior to participating in the experiment. These subjects were randomly assigned to the two training conditions, with each group having 20 participants. All subjects completed the experiment successfully.

### *Expert-System Development*

An ES in emergency management of chemical spills (Johnson & Jordan, 1983) was developed to serve as the training tool. The knowledge base of the ES consisted of 52 domain-specific rules, in the form of "IF symptoms THEN action", that dealt with various aspects of spill management (Appendix A shows some sample rules). A "HOW" explanatory facility (Luger & Stubblefield, 1989) was programmed to demonstrate how the ES employs a normative search strategy to chain relevant rules to prove a query. This facility is particularly significant because it is the reasoning lines

generated by the justifier that will function as feedback for debiasing overconfidence. The ES also included an interface in which the interaction with the ES-based feedback could be conducted in a friendly, natural language environment.

### *Stimulus Material/Query Systems*

The query systems represented a set of scenarios that consisted of queries and associated facts simulating spill incidents (Appendix B shows a sample scenario). There were two sets of query systems. One was for training and included 4 scenarios. The other was for experimental tests and included 10 scenarios. Both sets of the query systems were manipulated to chain the same number of rules so that the 14 queries were identical in terms of processing difficulty. The answer to each of the 14 queries comprised five alternatives, only one of which was correct.

### *Performance Measures*

In order to test the aforementioned hypotheses, three aspects of rule-based performance were measured. The first measure related to performance accuracy and was defined as the proportion of the queries that were solved correctly. The second measure concerned confidence judgment and was evaluated as the probability that the answer to a test query was correct. This confidence rating ranged from 1.0 (absolutely confident) to 0.2 (a random guess). The third measure was to test overconfidence. According to Lichtenstein et al. (1982), a judgment is calibrated if, for all propositions assigned a given probability, the proportion that is true equals the probability assigned. Therefore, the measure was derived by computing, for all scenarios assigned a given probability, the average deviation between the confidence rating and the proportion of scenarios solved correctly, weighted by the number of decisions made within the confidence category. This score ranged from 1.0 (completely overconfident) to $-1.0$ (totally underconfident), with 0.0 indicating perfect calibration. The mathematical expression of the confidence score was as follows: overconfidence/underconfidence $= (1/N)\ \Sigma_{t=1\sim T}\ n_t(r_t - c_t)$, where

$N$ = total number of decisions made,
$T$ = total number of confidence categories,
$c_t$ = the proportion correct in confidence category t,
$n_t$ = the number of decisions made in confidence category t,
$r_t$ = the probability assigned to confidence category t.

### Procedures

The experiment consisted of the following stages:

1. *Memorization session.* All subjects were required to memorize the 52 domain rules, with an emphasis on being able to recognize the association between the symptoms in the IF part and the action in the THEN part. The memorization task was performed as a take-home assignment.
2. *Pretraining session.* In this session, all subjects were required to take a test to demonstrate their knowledge of the rules. The test included 22 blank-filling questions in which the subject was asked to provide associated IF symptoms, given a THEN part, or associated THEN actions, given an IF part. Only those who answered all the questions correctly qualified to enter the next training session. The adoption of such a strict measure was to exclude an extraneous situation where the failure in confidence judgment resulted from forgetting the rules. The subjects who scored unsatisfactorily were instructed to review the rules and to retake the test.
3. *Training session.* In both training groups, each subject was required to solve the four training queries that were presented through a computerized scenario window. After completing each training query, the ES subjects were instructed to interact with the ES and the associated HOW explanatory justifier to observe the ES-based feedback. The control subjects, however, were not allowed to access the ES and received no feedback at all concerning the correctness of their decision outcomes and processes.
4. *Test session.* In this session, all subjects were required to solve the test query system on a computerized data collector. The 10 test queries were presented through a scenario window in the data collector and were separated, by the time-pressure variable, into two categories. Half of the queries were designated for the 90-s deadline condition, and the other half for the no-time-pressure condition. The order in which these 10 replicates (i.e., test queries) were presented was randomized independently for each of the 40 subjects. A clock displaying the 90-s countdown appeared in the data collector as soon as a time-constrained test query was presented. Immediately on finishing each of the test queries, all subjects were required to enter the answer and associated confidence rating into the data collector.

## RESULTS AND ANALYSIS

The descriptive statistics for the three response measures are summarized in Table 1. Separate ANOVAs with one between-subjects factor (training) and

**Table 1. Means (and Standard Deviations) of Performance Measures for Each Training Condition as a Function of Time Pressure**

| | Control | | ES | |
|---|---|---|---|---|
| Measure | No time pressure | Time pressure | No time pressure | Time pressure |
| Accuracy | 0.683 | 0.381 | 0.704 | 0.522 |
| | (0.176) | (0.224) | (0.165) | (0.198) |
| Confidence rating | 0.747 | 0.622 | 0.716 | 0.505 |
| | (0.195) | (0.147) | (0.149) | (0.148) |
| Overconfidence/ | 0.071 | 0.248 | 0.011 | −0.013 |
| underconfidence | (0.134) | (0.144) | (0.063) | (0.109) |

Abbreviation: ES = expert systems.

one within-subjects factor (time pressure) were performed on the three measures. Analysis of the ANOVA results concerned firstly how subjects reacted to time pressure, followed by an examination of the effectiveness of the ES-based training. Interactions were studied where appropriate. In-depth analysis of the interaction was conducted by the method of simple main effects (Kirk, 1993).

The main effects of both time pressure and training on the confidence score were significant, $F(1, 38) = 10.96$, $p < .003$, and $F(1, 38) = 30.93$, $p < .0001$, respectively. However, a significant interaction, $F(1, 38) = 18.91$, $p < .0001$, called for further investigation of the main effects. The simple main effects analysis revealed that the confidence score varied greatly in response to the manipulation of time pressure, and that the source of the significant variation came from the subjects at the control level of the training treatment. This was evidenced by the fact that the subjects who did not receive the ES training showed a natural tendency to be overconfident. This tendency became significantly stronger with the imposition of time pressure, $M = 0.071$ versus $M = 0.248$, $F(1, 38) = 29.33$, $p < .0001$. However, for those who received the ES training, their response to the presence of time pressure in the confidence score was not sensitive, $M = 0.011$ versus $M = -0.013$, $F(1, 38) = 0.54$, *ns*.

On the other hand, the ES training was found to exhibit an impact on the subjects' calibration performance. The simple main effects analysis showed that the source of the significant difference came primarily from the training effect with the presence of time pressure. The result was evidenced by the significant change in the confidence score from being overconfident at the control level to being underconfident at the ES level when the rule-based task was performed under time pressure, $M = 0.248$ versus $M = -0.013$, $F(1, 76) = 49.83$, $p < .0001$. Under the normal (no-time-pressure) condition, the ES training caused a slight decrease in the confidence score, $M = 0.071$ versus $M = 0.011$, $F(1, 76) = 2.63$, $p < .10$.

With respect to the measure of confidence rating, the ANOVA results showed no significant interaction of time pressure and training,

$F(1, 38) = 1.53$, *ns*. The main effect of time pressure was significant, $F(1, 38) = 24.61$, $p < .0001$. The confidence rating decreased from 0.7315 at the no-time-pressure level to 0.5650 at the time-pressure level. The main effect of training was marginally significant, $F(1, 38) = 3.56$, $p < .07$. The confidence estimate was reduced from 0.6845 for the control condition to 0.6120 for the ES condition.

With regard to decision accuracy, the main effect of time pressure was significant, $F(1, 38) = 71.06$, $p < .0001$, but there was no significant effect of training $F(1, 38) = 2.21$, *ns*. However, explanation of the two main effects must be qualified since there was a significant interaction, $F(1, 38) = 4.44$, $p < .05$. The simple main effects analysis pointed out that the subjects in both training conditions suffered significant decrease of accuracy when confronting time pressure: $M = 0.68$ versus $M = 0.38$, $F(1, 38) = 55.52$, $p < .0001$ for the control group; and $M = 0.70$ versus $M = 0.52$, $F(1, 38) = 19.99$, $p < .0001$ for the ES group. However, under the presence of a time constraint, the feedback received from the ES training showed its competence in sustaining a reasonable level of rule-based performance, $M = 0.52$ versus $M = 0.38$, $F(1, 76) = 5.29$, $p < .05$.

The calibration behavior under the manipulation of the four treatment conditions was also demonstrated by the calibration curve (Lichtenstein et al., 1982) displayed in Figure 1. The identity line in Figure 1 represents perfect calibration, with the curves below and above the identity line meaning

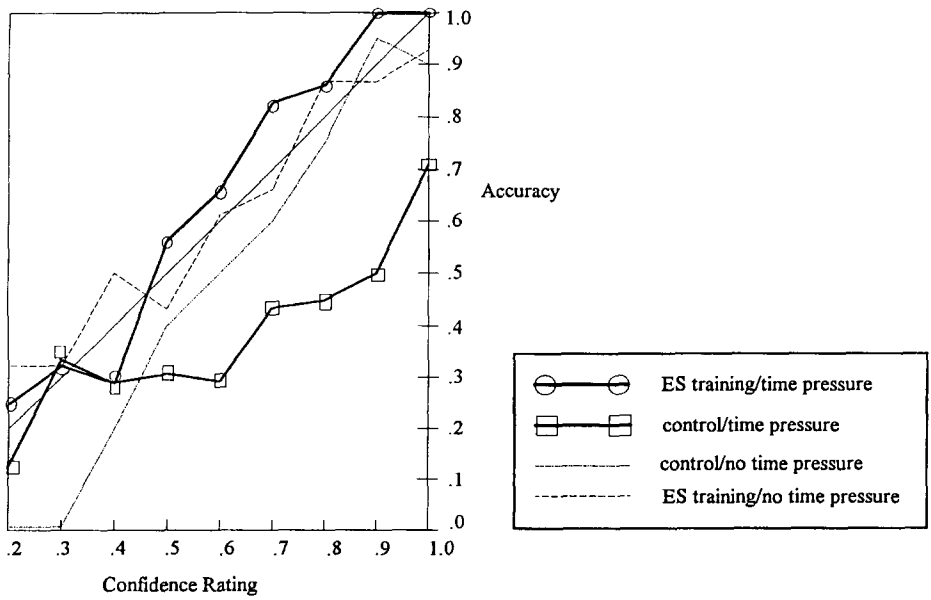


**Figure 1. Calibration curves for the four training and time-pressure manipulations. ES = expert systems.**

overconfidence and underconfidence, respectively. The calibration pattern found in the present study was quite similar to that in Arkes et al. (1987).

According to Figure 1, the subjects who were not exposed to the ES training tended to assign unjustifiably high confidence assessment, relative to the corresponding accuracy level. This bias was particularly strong in the time-pressure condition, as demonstrated by the most distant curve below the identity line. However, the confidence judgment of ES subjects acted in the opposite direction. The ES-trained calibration behavior was displayed by the curve close to the identity line for the no-time-pressure condition, and by the underconfidence curve for the time-pressure condition.

## DISCUSSION

Overall, the experiment results support our two primary hypotheses. The first hypothesis predicted the manifestation of overconfidence in the rule-based emergency domain and this prediction was confirmed. It appears that the increasing feelings of confidence in operators supervising emergencies are not a sure sign of increasing accuracy in their decisions. At this point, we will examine the mechanism that underlies the observed overconfidence. One plausible explanation is that, due to the limited capacity of working memory (Wickens, 1992), people tend to employ the heuristic of anchoring (Tversky & Kahneman, 1974) in the processing of rule-based information. We assume that this is done by putting unjustified weight on the rules whose symptom values are immediately available.

This tendency is very likely to cause the confirmation bias (Einhorn & Hogarth, 1978) in which people show an inertia to confirm the anchoring rules and to disregard the rules that need to be further inferred. The biased sense of confirmation may lead subjects into overestimating the accuracy of a decision that could eventually be overridden by the neglected rules, hence resulting in overconfidence. As cited earlier, people adapt to time pressure by relying more heavily on intuitive processing of information. The contingent behavior is expected to result in more frequent use of the hypothesized heuristics, which explains the much stronger phenomenon of overconfidence under time pressure for the nontrained individuals.

Given this speculation, any contingency rules or procedures that are readily available to the decision-maker's memory can be the origin of the possible overconfidence. These rules may include, for example, rules whose symptoms are directly provided in an incident, and the rules associated with high-frequency events. This implication may serve as a guideline for information display design. We suggest that rule-based knowledge for emergency management should be presented to human operators with equal

salience so that the anchoring rules bearing cognitive availability will not be processed with unwarranted precedence.

The second hypothesis on the effectiveness of the ES calibration training was also confirmed. Obviously, through the feedback concerning the ES/AI search strategies (e.g., forward/depth-first chaining), the subjects have an opportunity to correct their intuitive decision errors and learn the normative, cognitive-demanding process of rule-based reasoning. It appears that it is the comprehension which makes the subject recognize the complexity required for deriving successful solutions and, therefore, assign more realistic (i.e., downward) estimates of hit rates. This interpretation may imply that inappropriate realism of confidence judgment can be prevented if decision problems are resolved based on normative algorithms. This finding is analogous to that of Jiang, Muhanna, and Pick (1996), which suggested the use of normative methods (e.g., Bayes' Rule) for justified confidence in model selection in a decision support system. In addition, the adjustment was found to be particularly apparent under time pressure. This may be due to the ES subjects' realization that the normative completion of the rule-based task is less likely with such a time stringency, resulting in a more cautious confidence evaluation of their performance.

Note that the effects of the ES training were built on simultaneous improvements on both dimensions of the confidence–accuracy relationship. The performance capability in terms of accuracy level was also improved by the ES training. This result is in line with that of Sharit, Chen, and Lin (1993). The concurrent (but different in magnitude) enhancement of performance ability and confidence judgment is considered the cause of the observed underconfidence for the ES subjects in the time-pressure condition. Underconfidence is more desirable than overconfidence, since the consequences of underconfidence are usually less detrimental (Sen & Boe, 1991). Therefore, in emergency situations, it seems justified to encourage trained individuals to be conservative on the decisions made, even when their performance capabilities have actually been strengthened. However, the extent of the downward adjustment for confidence still needs to be controlled within an acceptable range.

Another interesting finding is that, in the present study, a significant improvement of calibration was achieved through a small amount of training. The investigation by Lichtenstein and Fischhoff (1980), perhaps the most ambitious effort in the literature of calibration training, found 200 items followed by feedback were sufficient to teach people to be well calibrated. The feedback in their study comprised summary statistics of calibration performance. Arkes et al. (1987) also achieved a significant improvement in calibration, but with only 5 feedback items where the feedback consisted of specific answers to individual questions. The present

study used even fewer training trials (4) in which more detailed feedback concerning the exhaustive, noncompensatory process of the rule-based search was provided. The improvement may primarily relate to the subjects' learning of logical reasoning, rather than their direct learning of calibration. That is, by observing specific feedback demonstrating how the rules were logically processed, ES-trained individuals become better reasoners and also more aware of their own fallibility in the manipulation of rule-based knowledge. The implication of this finding is that feedback with a higher degree in specificity may represent a more efficient debiasing technique for overconfidence. It is suggested that, in practice, more informative feedback should be given higher priority in situations where a substantial amount of training is not possible.

## CONCLUSIONS

Although ES have been thought of as a powerful decision support system (Hayes-Roth et al., 1983), the present study demonstrates a new paradigm of ES as a training aid. The significance of utilizing ES to train stems from the need for proper calibration of decision confidence during emergency situations. The qualification of the ES-based calibration training is justified by the evidence that trained individuals do lower their undue confidence toward desirable realism of their performance levels. Despite the supportive findings, there are limitations to the present research that must be addressed.

First, the knowledge base for the selected application domain in the experiment was kept monotonic and reliable. Although this simplicity was needed to control the scale of the tasks appropriate for the subjects, the problems arising in real-world emergency situations often bear data that are uncertain and incomplete. Future research calls for the need to incorporate uncertainty in domain knowledge in order to better understand the calibration issue with real-life implications. This can be done by training people using an ES built with models that handle uncertainty, such as fuzzy logic (Zadeh & Kacprzyk, 1992) and the certainty theory (Buchanan & Shortliffe, 1984). It would then be interesting to compare people's calibration behavior with the ES confidence models for unreliable rule-based information.

Second, most emergency incidents differ in the degree of processing difficulty. Previous studies (Lichtenstein et al., 1982; Sen & Boe, 1991) indicated that task difficulty plays an important role in influencing humans' calibration behavior. Therefore, it is necessary to take into account the hard–easy effect to further investigate the issues examined in the present study.

*Acknowledgments* — The authors are grateful to the anonymous reviewers for their helpful comments. This study received support from the National Science Council of Taiwan under grant NSC85-2213-E-009-059.

## REFERENCES

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes, 39*, 133–144.

Brown, B. G., & Murphy, A. H. (1987). Quantification of uncertainty in fire-weather forecasts: Some results of operational and experimental forecasting programs. *Weather and Forecasting, 2*, 190–205.

Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule-based expert systems: The MYCIN experiments of the Stanford heuristics programming project*. Reading, MA: Addison-Wesley.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85*, 395–416.

Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting, 1*, 155–172.

Hayes-Roth, F., Waterman, D. A., & Lenat, D. B. (1983). *Building expert systems*. Reading, MA: Addison-Wesley.

Heath, F., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty, 4*, 5–28.

Jiang, J. J., Muhanna, W. A., & Pick, R. A. (1996). The impact of model performance history information on users' confidence in decision models: An experimental examination. *Computers in Human Behavior, 12*, 193–207.

Johnson, C. K., & Jordan, S. R. (1983). Emergency management of island oil and hazardous chemical spills: A case study in knowledge engineering. In F. Hayes-Roth, D. A. Waterman, & D. B. Lenat (Eds.), *Building expert systems* (pp. 349–375). Reading, MA: Addison-Wesley.

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

Kirk, R. E. (1993). *Experimental design: Procedures for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107*, 296–310.

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

Kulhavy, R. W. (1977). Feedback in written instructions. *Review of Educational Research, 47*, 211–232.

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

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior and Human Performance, 20*, 159–183.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*, 149–171.

Luger, G. F., & Stubblefield, W. A. (1989). *Artificial intelligence and the design of expert systems*. Redwood City, CA: Benjamin/Cummings.

Mehle, T. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychologica, 52*, 87–116.

Moray, N. (1988). *Can decision aids help to reduce risk and human error? A program for research on the management of risk* (Report EPRL-88-20). Urbana–Champaign: Engineering Psychology Research Laboratory, Department of Mechanical and Industrial Engineering, University of Illinois at Urbana–Champaign.

Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology, 29*, 261–265.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 534–552.

Rasmussen, J. (1986). *Information processing and human–machine interactions: An approach to cognitive engineering*. New York: North-Holland.

Sen, T., & Boe, W. J. (1991). Confidence and accuracy in judgments using computer displayed information. *Behaviour and Information Technology, 10*, 53–64.

Sharit, J., Chen, S., & Lin, D.-Y. M. (1993). Expert system based training for emergency management. *Journal of Computing in Civil Engineering, 7*, 6–22.

Svenson, O., & Mauley, A. J. (1993). *Time pressure and stress in human judgment and decision making*. New York: Plenum Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Wickens, C. D. (1992). *Engineering psychology and human performance*. New York: Harper Collins.

Zadeh, L., & Kacprzyk, J. (1992). *Fuzzy logic for the management of uncertainty*. New York: Wiley.

## APPENDIX A

### *Some Sample Rules of the ES Knowledge Base*

Rule 4:

| | |
|---|---|
| IF | (victims are contaminated) |
| and | (victims have blood circulation problems) |
| THEN | (perform treatment S on the victims) |

Rule 7:

| | |
|---|---|
| IF | (the spill area is > 10 mm) |
| and | (the spill is classified as type A) |
| THEN | (establish command post C2) |

Rule 24:

| | |
|---|---|
| IF | (the spill substance is chlorine) |
| and | (the spill density is > 5 ppm) |
| THEN | (classify the spill as type A) |

Rule 37:

| | |
|---|---|
| IF | (perform treatment S on the victims) |
| and | (take evacuation route X) |
| THEN | (assign the victims to the RED first aid zone) |

## APPENDIX B

### *A Sample Spill Scenario/Query System*

[Query]: Given the following facts, please identify the emergency level of the spill incident ... A, B, C, D, or E?

[Facts]: The spill is taking place in the chip production zone; spill substance is chlorine; night working shift is on duty; spill area is $> 10$ mm; victims are contaminated; there is on-scene explosion; spill density is $> 5$ ppm; victims have breathing problems.