Critical Decision Method for Eliciting Knowledge

GARY A. KLEIN, ROBERTA CALDERWOOD, AND DONALD MACGREGOR

Abstract —**A critical decision method is described for modeling tasks in naturalistic environments characterized by high time pressure, high information content, and changing conditions. The method is a variant of Flanagan's critical incident technique extended to include probes that elicit aspects of expertise such as the basis for making perceptual discriminations, conceptual discriminations, typicality judgments, and critical cues. The method has been used to elicit domain knowledge from experienced personnel such as urban and wildland fireground commanders, tank platoon leaders, structural engineers, design engineers, paramedics, and computer programmers. A model of decisionmaking derived from these investigations is presented as the theoretical background to the methodology. Instruments and procedures for implementing the approach are described. Applications of the method include developing expert systems, evaluating expert systems' performance, identifying training requirements, and investigating basic decision research issues.**

Introduction

THE OBJECTIVE of this article is to describe a knowledge elicitation strategy, the critical decision method (CDM), and to show some of the ways it has been applied. The CDM builds on critical incident techniques [1] by using a set of cognitive probes to determine the bases for situation assessment and decisionmaking during nonroutine incidents. Rather than waiting for these incidents to occur, the CDM relies on interviews with proficient decisionmakers to examine recent cases of interest. The CDM can be used to study the cognitive bases of judgment and decisionmaking in naturalistic settings, with people at different levels of expertise. Throughout this paper the terms "expert" and "novice" are used as a convenience to refer to higher and lower levels of skills and experience. There is no attempt to define an expert in absolute terms, but "expert" in the studies to which we refer are generally individuals who have over ten years experience and would be recognized as having achieved proficiency in their domain. The term "novice" is used strictly in the relative sense. Although they have had significantly less experience than the experts, the novices we studied all had more domain experience than is typical of experimental studies using college sophomores. The operational definitions of these terms differ from study to study and are provided where appropriate. This approach is especially valuable for examining skilled performance under time pressure, where there is a limited opportunity for conscious deliberation.

To understand the choices made in designing the CDM, it is important to examine the need for knowledge elicitation techniques in general.

The Importance of Knowledge Elicitation

Extensive effort is being applied on a number of fronts to improve the quality of human performance in decisionmaking tasks. These efforts include a) the development and implementation of technologies to aid and support cognitive and behavioral components of human performance; b) the design of instructional curricula to speed the training of individuals to expert-level proficiency in task performance; and c) the design of systems to automate critical task functions that incorporate elements of human reasoning processes. A common element that exists in all efforts to improve human performance is a specification of the bases of skill performance that will enable task performance to be enhanced through training, aiding, or automation.

One approach for improving the overall level of human performance in a task is to understand how proficient individuals perform that task. By studying in detail the general knowledge, specific information, and reasoning processes an expert uses, a model of the task can be constructed that exhibits some of the properties of the expert being modeled. In addition, the model can be used to identify opportunities for improved training of nonexperts and for aiding and supporting decisionmaking.

Why is it that the behavioral sciences have not already provided knowledge elicitation tools? During most of this century, American researchers have emphasized *processes* of cognition. The information-processing paradigm addressed general heuristics related to thinking and problem solving. Methodology has focused on studying cognitive processes in context-restricted laboratory environments. With the emergence of expert systems and interest in naturalistic decisionmaking, researchers have become interested in the *content knowledge* of experts.

Until recently there have been very few methods for eliciting this content knowledge. Some of these, e.g., multi-

Manuscript received March 6, 1988; revised August 14, 1988. This work supported by the U.S. Army Research Institute for the Behavioral and Social Sciences under contracts MDA903-85-C-0099, MDA903-85-C-0327, and MDA903-C-0170.

G. A. Klein and R. Calderwood are with Klein Associates Inc., P.O. Box 264, Yellow Springs, OH 45387.

D. MacGregor is with Decision Research, 1201 Oak Street, Eugene, OR 98411.

IEEE Log Number 8927040.

0018-9472/89/0500-0462$01.00 ©1989 IEEE

dimensional scaling and network analysis (e.g., [2]) and repertory grid analysis [3], often require much time, effort, and control. Protocol analysis, for all its weaknesses, is generally applicable to naturalistic tasks, and the CDM described in the following relies on a type of protocol analysis for recalled events. These retrospective protocols are less disruptive than ongoing verbal protocols and are therefore applicable to a wider set of naturalistic tasks.

It would be a mistake to select a knowledge elicitation strategy without first developing a perspective on expert performance. Any technique will highlight some aspects of expertise and de-emphasize others. To interpret the results of a knowledge elicitation effort, it is necessary to appreciate the various aspects of proficient performance. The knowledge elicitor needs to understand both what is being captured and what is being missed.

This paper presents an approach for structuring knowledge elicitation for a class of behaviors that we term recognition-primed decisions (RPD), decisions for which action alternatives are directly derived from a recognition of critical information and prior knowledge.

The Nature of Expertise

One class of knowledge necessary for expertise is explicit and objective knowledge—*factual knowledge*, *if/then rules*, and *analytical procedures*. These types of knowledge are the stock in trade of knowledge engineers in building expert systems.

A second component of expertise has been described by Polanyi [4] as "tacit knowledge," since it is resistant to being articulated. Appreciation of *contextual* implications falls into this category. Contextual knowledge is viewed as the background of practices enabling experts to articulate if/then rules and apply analytical procedures. Aspects of *analogical inference* also belong in this category. Analogical reasoning is sometimes explicit, but tacit inferences may be drawn by the way the situation is recognized. *Judgments of typicality* are tacit since you don't have to analyze a situation to determine that you experienced similar cases in the past [5]–[7].

A third aspect of expertise involves *perceptual learning* and the development of a *perceptual-motor feel*. As skills are mastered, finer discriminations are made and tools come to be manipulated automatically. Not all tasks require perceptual learning. Flying an airplane, driving a car, playing tennis, and putting out fires in apartment buildings all require perceptual learning, whereas chess and computer programming do not.

For many reasons there has been a tendency to emphasize explicit and objective knowledge. Yet the other aspects of expertise are important and, in many cases, necessary to proficient levels of performance. It is essential that knowledge elicitation methods include some means of representing the contribution made by tacit knowledge and by perceptual learning. It may not be possible to analyze tacit knowledge, but knowledge elicitation methods should describe the function served by tacit knowledge in proficient task performance so that it should not appear that explicit knowledge is sufficient for proficient performance. If the knowledge elicitation method is insensitive to tacit knowledge, then it is easy to draw the mistaken conclusion that explicit knowledge is sufficient for performing a task well. We want to avoid this mistake.

Our knowledge elicitation strategies grew out of efforts to model naturalistic decisionmaking. It may be helpful to describe this model, which serves as theoretical basis for the method.

Recognition-Primed Decisionmaking

For the past several years we have been investigating the decisionmaking strategies used by experienced personnel in operational settings, such as urban firefighters [8], [9], wildland firefighters [10], tank platoon commanders [11], paramedics [12], and design engineers [13]. These studies were designed to collect descriptive data and would complement studies of decisionmaking in laboratory settings that used well-defined tasks and naive subjects.

A major finding of our research is that experts rarely report considering more than one option (e.g., [8]). Instead, their ability to handle decision points appears to depend on their skill at recognizing situations as typical and familiar. This recognition suggests feasible goals, sensitizes the decisionmaker to important cues, provides an understanding of the causal dynamics associated with a decision problem, suggests promising courses of action, and generates expectancies. Fig. 1 presents the model of what we have termed RPD's. (See [14], for a fuller account of the RPD model.)

One way of distinguishing this conceptualization of decision processes from more normative approaches is to use serial and concurrent option evaluation. According to a concurrent evaluation model, the decisionmaker considers several options at the same time. This can be done by performing pair-wise comparisons of options in terms of the evaluation criteria, and making relative evaluations of the strengths and weaknesses of each option. Options are then ordered according to their performance on the evaluation criteria. The option selected for implementation is the one achieving the highest multiattribute evaluation score. Though decision models of this general type are widely prescribed for making decisions in many contexts (e.g., [15]), their validity as descriptors of human decisionmaking is extremely weak. In time-sensitive contexts, for example, the tempo of decisionmaking is frequently much too fast for the performance of an all-inclusive generation and evaluation procedure (e.g., [16], [17]).

In contrast, in a serial evaluation model an option is generated, then tested for feasibility, and then either implemented or rejected. If it is rejected, a second option is considered, and so forth, until a suitable option is found. All options are not necessarily evaluated on all dimensions or on the same dimensions. Although one or more options may be considered, only one option is examined at a time.

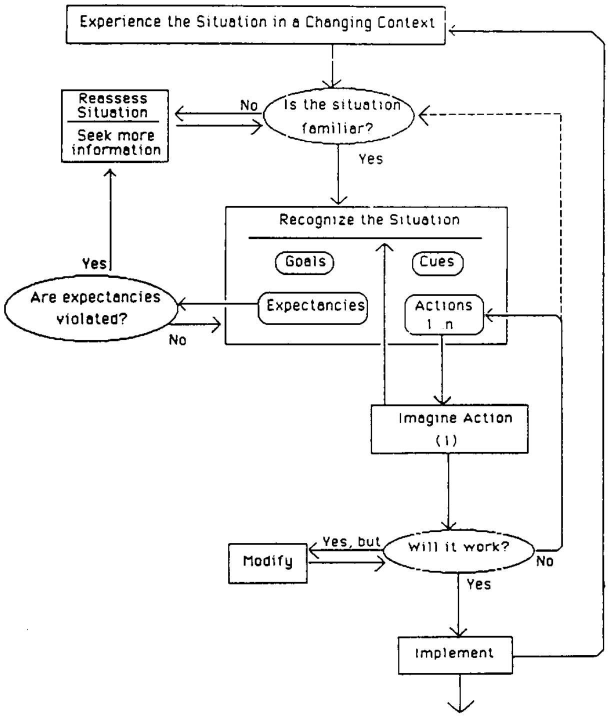

Fig. 1. Recognition-primed decision (RPD) model.

Serial evaluation has clear benefits for most applied time-sensitive contexts.

Because the RPD model assumes that an acceptable course of action may be chosen without conscious generation and evaluation of alternatives, the emphasis in this model is on what we have called *situational awareness*. This view is similar to the concept of *coup d'oeil* described by von Clausewitz [18], the skill of making a quick assessment of a situation and its requirements. It should also be clear that the RPD model includes aspects of behavior that are typically considered as problem solving. The focus of the model is on how people commit to options even though alternatives exist. However, in naturalistic settings it is rarely possible to keep problem solving and decisionmaking distinct. Rather than attempting to preserve such a distinction, the RPD model reflects both types of processes.

Practical Knowledge Elicitation Issues

It is important to understand some of the constraints on performing knowledge elicitation. Little will be accomplished by using techniques that are well-grounded in theory or are impressively validated unless the application of those techniques is feasible within the context of the task. We have identified several practical criteria for judging the effectiveness of knowledge elicitation methods.

First is the time needed to apply the methods. In basic research studies it is rare that the knowledge elicitor will have more than two hours at a stretch with a domain expert, and we have been in situations where we had 15 min or less. Therefore it is essential that the knowledge elicitor be able to prioritize the prepared questions and probes and to minimize extraneous material.

Second is the cost-effectiveness of data collection and analysis. On only rare occasions have we been able to videotape sessions and then perform microanalyses of the verbal and nonverbal behavior. However, working under budgeting restrictions, it is important to restrict the type of data collected to that which can be reasonably analyzed. There must also be efficient techniques for encoding and analyzing these data.

Third is the timeliness of the results. Basic research projects may not be completed until years after the data were collected. However, sponsors of applied projects tend to be less patient. In some training situations it is valuable to provide direct feedback of the knowledge elicitation data. In expert systems development it is useful to quickly program the results of the sessions to evaluate the results.

Fourth is the level of training needed for the knowledge elicitors. For most projects, highly trained elicitors are necessary. It is feasible to use personnel with less training by providing highly structured exercises and interviews, which have other drawbacks. Their developers must know so much about the domain they are studying that heavy front-end effort is required. The results of the subsequent elicitation are therefore less informative.

Fifth is the packaging of the knowledge elicitation results. When the goal is to develop an expert system, the format is clear cut — a software program. In other applications, such as system design or training, the format is less clear. One typical mistake is to obtain a large amount of information without having a concise means of presenting the knowledge to users.

These issues have been raised to put the problem of developing a knowledge elicitation strategy into perspective. Our criteria for a knowledge elicitation method were straightforward. First, the method had to address the basis of proficiency for highly skilled personnel. A technique such as task analysis was not satisfactory because it did not differentiate experts from novices when they carried out procedures in the same fashion. Second, the method had to be applicable to field conditions as opposed to well-controlled laboratory environments. Sometimes it is necessary to elicit information as a part of observing task performance; usually we traveled to the domain expert's place of work to conduct the interviews. Third, the products of the knowledge elicitation should have applied value in terms of training, system design, or development.

Description of the Critical Decision Method

The CDM is a retrospective interview strategy that applies a set of cognitive probes to actual nonroutine incidents that required expert judgment or decisionmaking. Once the incident is selected, the interviewer asks for a brief description. Then a semistructured format is used to probe different aspects of the decisionmaking process. Specific procedures have also been developed for analyzing the data.

Although the CDM shares many features with other interview methods, especially those related to Flanagan's critical incident technique [1], taken as a whole it offers some specific features that distinguish it from these and other knowledge elicitation strategies.

Focus on Nonroutine Cases

The CDM, like all critical incident techniques, focuses on nonroutine cases. Incidents that are nonroutine or difficult are usually the richest source of data about the capabilities of highly skilled personnel. This procedure increases the efficiency of the data collection sessions and allows certain aspects of expertise to emerge that would not be apparent in routine incidents. Although there are any number of methods available for performing a task analysis or establishing the procedural rules that apply to a given problem domain, an examination of nonroutine decisions is still essential for understanding the limiting cases. The focus on nonroutine cases also makes the method most appropriate to eliciting tacit knowledge that is not part of the formalized procedures for a domain.

Case-Based Approach

In a critical decision interview, questions always refer to a specifically recalled incident. We usually obtain more specific and useful information when we probe concrete and non-routine events than when we ask about general rules and procedures. This is one of the reasons for the success of Flanagan's [1] critical incident method. An interesting side effect of this approach is that the cases themselves become an important source of data for suggesting future research or testing system capabilities.

Cognitive Probes

Unlike Flanagan's critical incident technique, probing in the CDM is not limited to responses that can be objectively anchored and verified. Questions sometimes require the decisionmakers to reflect on their own strategies and bases for decisions. This is what makes the method so appropriate to a variety of knowledge elicitation needs. Granting that self-report methods are imperfect, they are still a rich source of data for generating hypotheses and testing proposed models.

Semistructured Probing

The CDM attempts to strike a balance between a totally unstructured approach, such as an ongoing verbal protocol [19], and one completely structured, such as an interview. Verbal protocols can fail to produce certain kinds of information, such as perceptually based cues, that are difficult or unnatural for people to articulate. In keeping with Flanagan's approach, the probes are designed to obtain information at its most specific and meaningful level. As currently implemented, a significant amount of the interviewing time is spent in uncovering these cues.

The method was also designed to avoid some of the problems faced by fully structured interview methods. Although specific questions are asked for each decision point, the order and wording can still follow the natural flow of a dialogue. Because the interviewers have "heard the whole story" before probing begins, they are in a better position to adapt the timing and wording of questions to the specific case. When the answers to some questions have already been given, there is no need to repeat or bore the interviewee by going over this material. We have found the dialogue format to be essential in maintaining full cooperation and interest from participants.

Development of the CDM

The CDM was first formally applied in two related studies of fireground command decisionmaking [8], [9], and these studies will serve as a basis for illustrating the essential features. Fireground commanders are charged with allocating personnel and equipment at the scene of a fire or other emergency incident. Many of their decisions are guided by standard operating procedures that must be implemented or modified based on their own evaluation of the situation. Information is gained from observations and reports at the scene as well as through prior knowledge and preplanning. This task closely parallels military command-and-control environments that are characterized by time pressure, risk, and dynamically changing events.

Our choice of data-gathering methods was designed to strike a balance between a host of research objectives and practical constraints. For example, direct observations of command decisions coupled with an ongoing verbal protocol of a commander's thought processes was first considered (see [21] method of tough cases). However, such an approach was deemed impractical in this case. Not only are challenging incidents relatively rare in any single location and expensive to cover because of the extreme time pressure, but the nature of the task makes any risk of outside interference untenable. We have used on-site observations to develop requisite domain knowledge prior to performing the actual elicitation task, and whenever possible to augment the data gathering. Thus on-site observations were made during simulated force-on-force tank platoon field exercises [11] and when time permitted, as in the wildland fire incident [10].

At another extreme, simply asking fireground commanders to describe their decisions would have resulted in little more than unrelated "war stories." Our goal was to focus the expert on those elements of an incident that most affected decisionmaking and to structure responses in a way that could be summarized along a specified set of dimensions while still allowing the details to emerge with the commander's own perspective and emphasis intact.

Core Procedures

The procedures adopted for the critical decision interviews represent our solutions to meeting these goals and practical considerations. The basic procedure can be summarized in the following steps.

Step 1 — Select Incident: Incidents are selected that can illustrate nonroutine aspects of a domain. The idea is to

TABLE I
CRITICAL DECISION INTERVIEW PROBES

Probe Type	Probe Content
Cues	What were you seeing, hearing, smelling ...?
Knowledge	What information did you use in making this decision, and how was it obtained?
Analogues	Were you reminded of any previous experience?
Goals	What were your specific goals at this time?
Options	What other courses of action were considered by or available to you?
Basis	How was this option selected/other options rejected? What rule was being followed?
Experience	What specific training or experience was necessary or helpful in making this decision?
Aiding	If the decision was not the best, what training, knowledge, or information could have helped?
Time Pressure	How much time pressure was involved in making this decision? (Scales varied.)
Situation Assessment	Imagine that you were asked to describe the situation to a relief officer at this point, how would you summarize the situation?
Hypotheticals	If a key feature of the situation had been different, what difference would it have made in your decision?

probe for components that go beyond the general knowledge and procedures that enable a competent individual to perform routine tasks; we want to study those components that might discriminate the true expert. In doing this, it makes sense to select cases that presented a unique level of challenge for the individual. Thus we ask the decisionmaker to select an incident that was challenging and that, in his or her decisionmaking, might have differed from someone with less experience. This selection criterion is common to critical methods (e.g., [22]).

The participant usually met this criterion easily. If no case immediately came to mind, several cases might be briefly screened so that the participant could pick one that seemed most interesting. We particularly learned to avoid cases in which a death or unusual episode made the incident a memorable one but one in which the commander may not have played a key decisionmaking role.

Step 2 — Obtain Unstructured Incident Account: The officer was asked to describe the incident from the time he received the alarm to the time when the incident was judged to be under control. For the most part this account proceeded without interruption by the interviewers, except for minor points of clarification. The procedure accomplished several goals. First, it created a context for understanding on the part of the interviewer. It was important to us to have a sense of the individual's phenomenological perspective of the event as a bulwark against our own biases and the contamination of the questioning procedure itself. Second, the account served to activate the officer's memory of the event as a context for questioning. In addition, we judged that the procedure helped us achieve a high level of cooperation from the officers by establishing us as listeners rather than interrogators. Obtaining cooperation is absolutely essential to any knowledge elicitation effort, yet failure here is frequently cited as a major obstacle to success [23].

Step 3 — Construct Incident Timeline: After the incident had been related, the interviewer proceeded to reconstruct the account in the form of a timeline that established the sequence and duration of each event reported by the officer. Events included both objectively verifiable occurrences (e.g., "the second alarm equipment arrived two minutes later") and thoughts and perceptions reported by the officer (e.g., "the color of the smoke indicated the presence of a toxic substance. I thought I might have to call a second alarm at this point"). The timeline served to establish a shared awareness of the "facts of the case" from the officer's perspective. Many times inconsistencies in the account could be detected and corrected on the basis of the timeline, and missing facts filled in. In addition, questions about the timeline focused the officer's attention on events from more than a single time perspective, an approach having demonstrated utility for obtaining accurate eyewitness testimony [24].

Step 4 — Decision Point Identification: During the timeline construction, specific decisions were identified for further probing. In some cases the verbal cues marking a decision were obvious (e.g., "I had to decide whether it was safe enough to send my crews inside"), but this was not always the case. In other cases, it would be clear that an officer was taking one of several possible courses of action or was making a judgment that affected the outcome, but there was no clear indication that the officer saw himself as "making a decision" at this point. A decision point was probed if the officer would agree that other reasonable courses of action were possible or that another officer (perhaps one with less or greater expertise) might have chosen differently.

Step 5—Decision Point Probing: Different studies have used different probes, depending on the objectives of the projects. Table I summarizes the probe types that have been routinely used.

Questions to elicit the details of cue usage were almost always asked first as part of the timeline construction, and represented the current information that was likely to have been heeded at each event time. Prior knowledge was also probed. We had a special interest in eliciting any recall of prior experiences that influenced the officer's size-up or expectancies about a situation. Such specific remindings were coded as analogues.

Goals are an important part of situation assessment. However, it was important to elicit specific goals rather than accept the participant's tendency to state goals in a very general form. For example, "putting out the fire" is a higher-level goal that is always present but does not drive specific behaviors. Specific goals have reasonably stated alternatives such as "protect exposures" versus "attack the seat of the fire."

Probes about options were asked for each decision, both those that were actually deliberated and those that existed only hypothetically. We have found that the reasons for taking a particular action are most frequently illuminated through understanding choices that were not made or were rejected. The probe for hypotheticals was developed in order to identify the key features that determined action and to elaborate the scale along which the values of these features could be varied. The basis for selecting an option was probed extensively, and if a rule was used, it was stated.

Additional Procedures

Whenever possible we use a pair of interviewers for each session rather than a single interviewer. It can sometimes be difficult for a single interviewer to pay close attention to the unstructured account, take notes, and begin the timeline restructuring without missing something. Duties for notetaking and asking questions can frequently be allocated between the interviewers in such a way that eye contact and rapport are maximized. Having interview pairs also serves to allow a less-experienced interviewer to gain experience without having to risk valuable interview time if mistakes are made.

Interviews are routinely tape recorded so that verbatim transcripts can be made. However, this has not always been possible in particular cases. The importance of verbatim transcripts will depend on the nature of the elicitation goals.

Most interview sessions are planned to last for about two hours, but length can be adjusted to meet the complexity and time constraints of a given application. In the study of tank platoon exercises [11], for example, the need to reconstruct the timeline was obviated by the fact that the interviewers were observing the scene. Decision-point probing was undertaken in under 15 min during breaks in the field exercises.

We have almost always found it valuable to allow the interviewee to draw a diagram during the recounting of an incident and in response to specific probes. For many individuals the diagram serves as a necessary memory aid in reconstructing the key features of the incident. It also provides a common source of reference in communicating the participant's perspective to the interviewers.

Reliability and Validity

It is difficult to assess the reliability of a semistructured interview method because the exact circumstances can never be recreated and, once interviewed, the interviewee's memory for the event will alter to some unknown degree. However, we have attempted to assess the intercoder reliability of the method by having subsets of the verbatim transcripts (representing from 1 to 2.5 h of interviewing) coded by different researchers.

One question was how reliably a "decision point" could be identified from the unstructured portion of the interview. If these could not be reliably located, then CDM interviews would be expected to produce highly variable results depending on the particular interviewers who were present. Four transcripts were selected at random from the urban fireground command study [9] and were coded both by the developer of the original coding scheme and by a new coder who was not present during any of the original interviews. Twenty-nine decision points were identified by the criterion coder in these transcripts, and the agreement between the two coders ranged from 81 to 100 percent. All of the decision points initially identified were also identified by the second coder. Disagreements reflected exclusively the tendency of the newer coder to call something a decision point that the experienced coder treated as simply a "standard response." Our own experience tells us that the interview methods are sensitive to the domain experience of the interviewer. Someone who knows very little about firefighting will tend to probe more than someone who has "heard the answer to that one" before.

Besides reliably identifying the decision points, we were interested in how reliably the decision could be classified in terms of the identified strategies. To answer this question, the same 29 decision points were independently coded by the same coders based on the interview transcripts. The exact agreement between the coders for classifying the decision points into the five categories defined in this study was 66.5 percent, based on a weighted average. Although this rate of agreement was significantly higher than would be predicted by chance (chi-square = 10.83; $df = 1$; $p < 0.001$), it indicated that coders had difficulty in making the discriminations at this level of detail. Essential agreement was calculated based on codes that were directly adjacent in the coding continuum and this was substantially higher (87.8 percent). Based on these findings we have abandoned the more fine-grained analysis in favor of the more meaningful distinctions that can be made.

A second source of reliability data was obtained in the study of wildland firefighting [10]. In this study the issue of decision-point identification was not present because the interviewers were present during the incident itself. Consequential decisions could be pinpointed immediately by mutual agreement between the interviewers and commanders. Decision strategy reliability was assessed on 18 probed decision points coded by the criterion coder and by a new coder who had not been involved in the initial study. The coding scheme was a modification of the one used in the fireground study [9] and the reliability results were similar. Overall, for five coding categories, exact agreement was 74.4 percent, but essential agreement calculated on the basis of adjacent categories raised this index to 88.9 percent.

Because interviews often take place months after an event has occurred, it is of interest to establish to what extent the time between the event and the interview influences the nature of the critical decision protocols. To examine this issue, a subset decision points probed immediately after a decision event occurred were compared to protocols obtained five months after the wildland fire incident [10]. A rater who was not present at the initial study conducted a content analysis of all material obtained for the decision points, both on-site and follow-up. The correspondence between on-site and follow-up was found to vary substantially for individual decisionmakers, ranging from 56 to 100 percent, with an average rate of agreement of 82.5 percent.

Two additional studies have examined the degree to which the CDM aids in the elicitation of situation assessment information over and above more unstructured methods [25] and whether the details so obtained can be used to structure effective training modules [26].

Analytic Procedures and Products

There is no single coding procedure for the critical decision protocols. The needs of the sponsors and specific research questions define the nature of the coding effort. The following discussion will briefly describe the types of products that we have generated thus far.

Descriptive Decision Model

All of the critical decision interviews that we have obtained (over 480 decision points) have been used to support and refine the previously described model of recognition-primed decisions. All of these studies were of individual decisionmaking involving various degrees of time pressure. The decisionmakers were all highly experienced, with the exception of the tank platoon leaders [11] and novice fireground commanders [9].

As an example of the type of findings reported in these studies, the study of fireground command decisionmaking [9] investigated the nature of the differences in decision strategies between more experienced (11.1 years of command experience) and less experienced (1.5 years of command experience) commanders ("experts" and "novices"). The decision point coding, therefore, concentrated on developing a system that could distinguish the decision strategies employed by these commanders. Each decision point was coded for whether or not concurrent or serial (RPD) evaluation was primarily used; more specific codes that attempted to capture the nature of the evaluation process were employed as well. One code characterized the decision point as primarily involving an answer to one of two distinct decision "strategies": deliberation about the situation (situation assessment) or deliberation about the reaction (option evaluation). The results of this coding suggested an interesting interaction between the strategies used by the 12 expert and the 12 novice commanders who were interviewed. Both groups primarily relied on recognitional decisions, but when deliberation was reported the novices were more likely to concurrently deliberate on the option evaluation dimension.

Critical Cue Inventory

The critical cue inventory (CCI) is a collection of all of the informational and perceptual cues that are pinpointed in the protocols. Many of the probes in the critical decision interviews are directed at gaining specific cues that were used in formulating a situation assessment or considering options. Many of these cues are not spontaneously mentioned by decisionmakers and do not result from asking very general questions like "Why did you make this particular decision?" This is why cognitive probes are needed.

The most direct use of a CCI to date has resulted from a study of paramedics [27]. The CCI consisted of the cues actually used by paramedics and others to recognize heart attack victims during and prior to their showing standard symptoms. Table II shows the set of cues reported by medical personnel for recognizing early signs of cardiopulmonary distress. This CCI served as the basis for the design of training materials to teach these perceptual discriminations.

Situation Assessment Record

Because the RPD model treats decisionmaking as a form of complex pattern matching, much of the expertise elicited appears as situational assessment that we see as the expert's understanding of the dynamics of a particular case. We have operationalized this assessment in terms of specific changes in cue usage and goals that we term a situation assessment record (SAR). For each decision point, critical cues and current goals are probed. Often an initial situation assessment is maintained throughout an incident, with new information serving to elaborate on what was originally known (SA-elaboration). In this case goals do not change, but may be refined or made more explicit. More extreme changes in situational assessment (SA-shifts) result when there is a perceived change in the nature of observed cues causing the decisionmaker to modify or replace earlier goals.

Table III shows an SAR for a fireground incident in which an oil truck overturned and burst into flames on a highway. The operations proceeded well until a storm drain blew out, signaling that the fuel had leaked into the

TABLE II
EXAMPLE CRITICAL CUE INVENTORY: EARLY WARNING SIGNS OF CARDIOPULMONARY DISTRESS [26]

Cue Category	Description
Skin tone	changes in skin color (skin losing pinkness and becoming blue/grey), especially at extremities (fingers, toes) and where blood vessels are close to the skin surface (face, lips)
Eyes	glazed, unfocused look; pupils may be dilated
Skin	cold, clammy feel; sweaty, greasy
Breathing	may be rapid, shallow breathing may show "air hunger"; struggling to get air into lungs; a crackling, bubbling noise at both inhale and exhale
Mental state	reduced awareness of surroundings; confusion; anxious and agitated

In the days/weeks prior to an attack, additional signs may be seen: complaints of fatigue and lethargy, nausea or indigestion, tightness in chest, muscle or "bone" aches in left arm/shoulder. These may be interpreted as symptoms of flu.

TABLE III
EXAMPLE SITUATIONAL ASSESSMENT RECORD FROM TANKER INCIDENT

SA-1	
Cues/knowledge:	Overturned truck on highway; ruptured fuel tank; engulfed in flames; intense heat (highway signs melted); another truck 50 feet away; citizen rescuing driver.
Expectations:	Potential explosion; life hazard.
Goals:	Complete the rescue; extinguish fire; block traffic.
DP-1:	Aid in driver rescue.
DP-2:	Call for additional units: rescue unit, police, foam.
SA-2 (Elaboration)	
Cues/knowledge:	Additional equipment arrives; fire more involved.
Expectations:	Chance of explosion is less.
Goals:	Protect firefighters; gain needed resources (water).
DP-3:	Set up covering streams.
DP-4:	Hook up pumper hoses.
SA-3 (Elaboration)	
Cues/knowledge:	Protective streams functioning; foam trucks arrive. Fire banking down.
Goals:	Optimal truck placement; set up foam operations.
DP-5:	Directed truck and foam placement (using angles, impact, and wind direction cues).
SA-4 (Shift)	
Cues/knowledge:	Storm drain behind operations blows up; determines jet fuel has leaked into storm sewers.
Expectations:	Fire will span out of control.
Goals:	Check fire and prevent spread.
DP-6:	Call second alarm.

storm sewer system and a need to change the way operations had to be conducted.

Thordsen and Klein [28] have developed alternative formats for representing SAR's that can be used to communicate situation assessments during group operational planning exercises.

Case Studies

Whereas traditional laboratory-based approaches tend to remove context in order to produce quantitative summaries, we have found that our sponsors frequently found the "messy" data to be the most useful. For example, we recently completed a study of system analysts and programmers in a large company [29]. The critical decision method was used to elicit details of how the programmers solved specific programming problems. Managers of the management information systems (MIS) department found the excerpts from these cases to be useful in targeting specific training deficits. The cases were also used to highlight the differences between the approaches taken by their best and less successful programmers and suggested means for improving performance, particularly in the area of user interactions.

APPLICATIONS OF THE CRITICAL DECISION METHOD

Three types of applications have been made thus far for the CDM: as a knowledge engineering approach for an expert system data base, as a method for evaluating expert systems, and as a method of analyzing skilled performance for identifying training requirements.

Knowledge Engineering for Building Expert Systems

The CDM has been used for knowledge engineering in the development of a case-based reasoning system prepared for the Air Force Weapons Laboratory. The objective of the system is to assist structural engineers in making assessments of the survivability and vulnerability of concrete structures, hence the name of the system—Surver, which stands for survivability/vulnerability analyses through experiential reasoning. The most helpful probes in this project have dealt with cues, knowledge, analogues, goals, options, bases for decisions, and hypotheticals. The probes on critical experiences, potential aiding, time pressure, and situation assessment were not relevant.

Hoffman [21] has evaluated several knowledge engineering strategies: analysis of familiar tasks, interviews, limited information tasks, constrained processing tasks, and the analysis of tough cases. The analysis of familiar tasks covers routine events and is comfortable for domain experts but is not efficient for generating sophisticated aspects of knowledge. Interviews, especially unstructured interviews, were very time consuming, and Hoffman's measure of the efficiency of the different methods showed these to have the lowest efficiency. Limited information tasks placed more strain on domain experts and, while they generated some interesting results, they ran into the barrier of lower cooperation. The same was true of constrained processing tasks. The last category was what Hoffman [21] called the "analysis of tough cases," which overlaps with the nonroutine cases we have addressed with the CDM. Hoffman found that these yielded good amounts of information about refined reasoning, namely, the more sophisticated knowledge that differentiates personnel at the higher levels of expertise. However, Hoffman noted that these tough cases occur unpredictably, often when the knowledge engineer is not present. That is why the CDM is useful as a technique to probe such events retrospectively.

The CDM may have an additional function for knowledge engineering. The data collected can provide an overview of the expertise involved in performing a specific task so that the designers can appreciate which aspects of the task depend on explicit and objective knowledge (and area appropriate for rule elicitation interviews) and which aspects depend on tacit knowledge, requiring either extensive knowledge engineering or possible redesign of the system to avoid the difficulties associated with probing for tacit knowledge. It is important to formulate an overall strategy for efficient knowledge engineering in view of the costs that accrue to the domain expert and the knowledge engineers.

Evaluating Expert Systems

Now that expert system technology is moving from research development to the status of widespread applications, it has become important to evaluate the impact of these systems. Sponsors must evaluate impact prior to initiating system development. Users must evaluate impact in order to help guide the development and anticipate changes resulting from the system. Developers must be able to define the system performance objectives in order to work from functional specifications.

TABLE IV
COMPARISON OF AIQ^{sm} RATINGS FOR AALPS VERSUS CALM
(1 = LOW, 5 = HIGH, N/A = NOT APPLICABLE) [13]

			Aalps	Calm
A. System effectiveness:				
	1)	Content coverage:		
		a) # cases	4	2
		b) % time	4	2
	2)	Power:		
		a) speed	5	5
		b) success rate	3	2
		c) quality	3	2
	3)	HOE:	2	1
	4)	Flexibility:	3.5	1.5
	5)	Extendability:		
		a) external maintenance	3.5	1
		b) internal maintenance	N/A	N/A
B. User effectiveness:				
	1)	Explanatory:	1	1
	2)	Error handling:	3	3
	3)	Tutoring:	1	1
	4)	Metaknowledge:	3	1
C. Organizational effectiveness:			N/A	N/A

We have used the CDM in several studies assessing expert system feasibility and performance. The first application [30] was to probe the expertise of data analysts to determine whether it was feasible to build an expert system to handle a portion of their job. The study showed that only a portion of their job involved routine rules-of-thumb, and that most of their job required complex and deep domain knowledge. Plans for expert system development were subsequently dropped.

Klein and Brezovic [13] used the CDM to develop a strategy for evaluating task performance of expert systems. Since this project was akin to building an IQ test, the technique is labeled "AIQ^{sm}" (artificial intelligence quotient) method. This method was successfully used to evaluate the automated air load planning system (AALPS) developed by Stanford Research Institute to support U.S. Army and Air Force personnel responsible for load configurations for cargo missions. Table IV shows an AIQ assessment of AALPS versus a predecessor system (computer air load manifest, or Calm) on two scales: system performance and human–computer interface.

The AIQ method consists of three steps. First, the bases for expert performance within a domain are specified using the CDM. Second, these bases are contrasted to the expert system approach taken by the system developers, and mismatches are flagged to show aspects of expertise that will not be covered as well as new capabilities that did not previously exist. (Since AALPS was a knowledge-based system that included algorithms for rapidly computing center of gravity, it could outperform the experts on certain important subtasks.) Third, using the performance level of the skilled operator as an anchor, the system performance is specified on a set of four scales: system performance, performance of the operator/system dyad, adequacy of user–computer interface, and system impact on the organization. Each of these scales has several subscales. For example, performance of the operator/system dyad depends on the expertise of the user and the level of difficulty of the task. Using a three-point scale for each of these, we needed a 3×3 matrix for each dimension of system effectiveness, which itself included speed, proportion of cases handled, and quality of output.

The AIQ method has since been applied to a second expert system early in the conceptual developmental cycle and to a decision support system developed by the Army for operational brigade-level planning. The value of the AIQ method is multiple. It clearly specifies the strengths and limitations of knowledge-based and decision support systems. It uses a variety of performance dimensions. It anchors these evaluations in terms of the expertise of the decisionmaker and highlights the effect of the system on enhancing or restricting that expertise.

Identifying Training Requirements

Because the CDM can generate an inventory of critical cues, these can be defined as training requirements. An example of this use for medical training was presented earlier. In addition, Crandall and Klein [30] used the CDM to study computer programmers at a large corporation. The results highlighted some key differences between effective and ineffective programmers, and these were used by the company to identify a new training requirement. Finally, the CDM also generates a set of nonroutine incidents, each with important decision points. These materials can be used for training program design.

Conclusion

The CDM was designed to efficiently gather data on the bases for proficient performance of naturalistic tasks. It is a theory-driven strategy that is based on the assumption that expertise emerges most clearly during nonroutine events and focuses on these as the prime source of information. The events have actually occurred, so there is no need to develop artificial simulations that are limited in contextual richness and are time-consuming in preparation and validation. The interviews cover prior events, so there is no need to wait for nonroutine events to occur. The interviews are semistructured so that they take less time to gather relevant information. They are easier to code than unstructured interviews, while still retaining the freedom of exploring promising avenues of investigation. The use of a standardized set of probes results in reliable data. In short, the method is an effective means of gathering knowledge elicitation material with relatively little effort.

The CDM has been employed in a range of field applications. It has been used in 2-h interviews with fireground commanders and in 15-min interviews with tank platoon leaders, just after an exercise while they were getting ready to attend an after-action review. It has been used in lengthy interviews with design engineers, computer programmers, and paramedics; and in briefer interviews with corporate-level battle commanders.

The method has served a variety of functions. These include providing knowledge engineering for expert system development, identifying training requirements and generating training materials, evaluating the task performance impact of expert systems and decision support systems, and performing basic research on decisionmaking.

Another limitation is that the CDM relies on verbal report methods. It is important to acknowledge that verbal reports are not a direct window into people's cognitive processes and the people can misrepresent their own decisionmaking strategies and goals. Each type of data collection method has its own limitations and potential biases. Rather than pretending that a perfect method can be developed, it is preferable to try to understand the biases that are possible and work to reduce these. For example, we have designed the CDM procedures to minimize sources of bias by asking for initial uninterrupted incident descriptions and refraining from directly asking why certain actions were taken.

Our experience with the CDM, covering a range of domains and applications, has suggested a more general discipline of *knowledge engineering*. Whereas the term has usually been applied to methods for eliciting information for expert systems, there are a variety of functions served by knowledge elicitation techniques—including communications, training, and documentation.

Organizations may be learning to treat knowledge as a resource, on a par with capital, equipment, manpower, and good will. Organizations accrue knowledge in a variety of ways, including technical skill development and corporate memory of how to handle difficult tasks and the lessons learned from previous incidents. Organizations suffer when they do not properly value their own expertise and when they lose skilled personnel without a chance to retain, share, or preserve the knowledge of people who retire or leave for other positions (including promotions into different departments).

Just as we have petroleum engineers and structural engineers, we may soon see the development of knowledge engineers. Petroleum engineers work on the tasks of identifying petroleum sources, extracting the petroleum, processing it, and applying it. Similarly, we may find knowledge engineers working to find sources of expertise in organizations, using methods for eliciting that expertise, and for codifying the expertise so that it can be applied. Expert systems are one means of codifying and applying expertise, but there will obviously be other techniques. The excitement about expert systems has helped us recognize the importance of knowledge engineering as a new speciality. This special issue includes a variety of methods for eliciting and codifying expertise, methods that may coalesce into the discipline of knowledge engineering.

Acknowledgment

Special thanks are offered to Stan Halpin, Rex Michel, Dennis Leedom, and Judith Orasanu for their encouragement and support.

References

[1] J. C. Flanagan, "The critical incident technique," *Psych. Bull.*, vol. 51, pp. 327–358, 1954.

[2] R. W. Schvaneveldt, F. T. Durso, T. E. Goldsmith, T. J. Breen, N. M. Cook, R. G. Tucker, and J. C. De Maio, "Measuring the structure of expertise," *Int. J. Man–Machine Studies*, vol. 23, pp. 699–728, 1986.

[3] J. H. Boose, "A knowledge acquisition program for expert systems based on personal construct psychology," *Int. J. Man–Machine Studies*, vol. 23, pp. 495–525, 1985.

[4] M. Polanyi, *The Tacit Dimension.* Garden City, NY: Doubleday, 1986.

[5] L. L. Jacoby and L. R. Brooks, Nonanalytic cognition: Memory, perception, and concept learning, in *The Psychology of Learning and Motivation: Advances in Research and Theory*, vol. 18, G. H. Bower, Ed. New York: Academic Press, 1984, pp. 1–47.

[6] D. Kahneman and D. T. Miller, "Norm theory: Comparing reality to its alternatives," *Psych. Rev.*, vol. 93, pp. 136–153, 1986.

[7] D. L. Hintzman, "'Schema abstraction' in a multiple-trace memory model," *Psych. Rev.*, vol. 93, pp. 411–428, 1986.

[8] G. A. Klein, R. Calderwood, and A. Clinton-Cirocco, "Rapid decision making on the fire ground," Rep. KA-TR-84-41-7, in *Proc. Human Factors Society 30th Ann. Meeting*, vol. 1, 1986, pp. 576–580.

[9] R. Calderwood, B. Crandall, and G. A. Klein, "Expert and novice fire ground command decisions," Rep. KATR-858(D)-87-02F, Yellow Springs, OH, Klein Associates Inc.; prepared under contract MDA903-85-C-0327 for U.S. Army Research Institute, Alexandria, VA, 1987.

[10] J. Taynor, G. A. Klein, and M. Thordsen, "Distributed decision making in wildland firefighting," Rep. KATR-858(A)-04F, Yellow Springs, OH, Klein Associates Inc.; prepared under contract MDA903-85-C-0327 for U.S. Army Research Institute, Alexandria, VA, 1987.

[11] C. P. Brezovic, G. A. Klein, and M. Thordsen, "Decision making in armored platoon command," Rep. KATR-858-(B)-87-05F, Yellow Springs, OH, Klein Associates Inc.; prepared under contract MDA903-85-C-0327 for U.S. Army Research Institute, Alexandria, VA, 1987.

[12] H. A. Klein and G. A. Klein, "Perceptual/cognitive analysis of proficient cardio-pulmonary resuscitation (CPR) performance," presented at Midwestern Psychological Assn. Meeting, Detroit, MI, 1981.

[13] G. A. Klein and C. P. Brezovic, "Evaluation of expert systems," in *Defense Applications of Artificial Intelligence*, S. J. Andriole and G. W. Hopple, Eds. Lexington, MA: D.C. Heath, 1988.

[14] G. A. Klein, "Recognition-primed decisions," in *Advances in Man-Machine Systems Research, Vol. 5*, W. Rouse, Ed. Greenwich, CT: JAI Press, 1989.

[15] H. Raiffa, *Decision Analysis: Introductory Lectures on Choices Under Uncertainty.* Reading, MA: Addison Wesley, 1986.

[16] K. R. Hammond, R. M. Hamm, J. Grassia, and T. Pearson, *The Relative Efficacy of Intuitive and Analytical Cognition.* Boulder, CO: Center for Research on Judgment and Policy, 1984.

[17] D. Zakay and S. Wooler, "Time pressure, training, and decision effectiveness," *Ergonomics*, vol. 27, pp. 273–284, 1984.

[18] C. von Clausewitz, *On War.* Princeton, NJ: Princeton University Press, 1833/1943.

[19] K. A. Ericsson and H. A. Simon, "Verbal reports as data," *Psych. Rev.,Z* vol. 87, pp. 215–251, 1980.

[20] G. A. Klein, "Knowledge engineering analysis of data extraction strategies," Technical Report for Perceptronics, Yellow Springs, OH, Klein Associates Inc., 1985.

[21] R. R. Hoffman, "The problem of extracting the knowledge of experts from the perspective of experimental psychology," *AI Mag.*, vol. 8, pp. 53–67, 1987.

[22] G. O. Klemp and D. C. McClelland, "Executive competence: What characterizes intelligent functioning among senior managers?" in *Practical Intelligence: Nature and Origins of Competence in the Everyday World*, R. J. Sternberg and R. K. Wagner, Eds. MA: Cambridge University Press, 1986.

[23] F. Hayes-Roth, D. A. Waterman, and D. B. Lenat, *Building Expert Systems.* MA: Addison-Wesley, 1983.

[24] R. E. Geiselman, R. P. Fisher, D. P. MacKinnon, and H. L. Holland, "Eyewitness memory enhancement in the police interview: Cognitive retrieval mnemonics versus hypnosis," *J. Appl. Psych.*, vol. 70, pp. 401–412, 1985.

[25] B. Crandall, "Acomparative study of think-aloud and critical decision elicitation," *SIGART*, May, 1988.

[26] L. A. Whitaker and T. H. Baynes, "Situation assessment training via firefighting study," paper presented at 29th annual meeting of the Psychonomic Sec., 1988.

[27] B. Crandall and G. A. Klein, "Critical cues for MI and cardiogenic shock symptoms," working paper, Yellow Springs, OH., Klein Associates, 1987.

[28] G. A. Klein and M. Thordsen, "Use of progressive deepening in battle management," in *Proc. 11th Biennial D. D Psych. Conf.*, Colorado Springs, CO, 1988.

[29] B. Crandall and G. A. Klein, "Key components of MIS performance," Report to the Standard Register Company, Yellow Springs, OH, Klein Associates Inc., January 1988.

Gary Klein is the Chief Scientist of Klein Associates Inc., an R&D company founded in 1978 to perform work in applied cognitive psychology. His principal area of research is in methods of knowledge elicitation that reflect the perceptual–cognitive aspects of expertise, with applications to training, displays, and expert systems. Previously he had been a Research Psychologist with the Air Force Human Resources Laboratory.

Roberta Calderwood, is a Senior Research Associates at Klein Associates. Her research has primarily involved developing methods to study operational decision making, especially realtime command and control. She has compared the decisionmaking strategies of more- and less-experienced fire ground commanders in order to investigate the development of expertise in stressful and complex environments, and is currently developing training applications for team and distributed decision environments.

Donald MacGregor is the President of MacGregor-Bates, Inc., and a Research Associate with Decision Research in Eugene, OR. His research interests are in human judgment and decision-making, decision aiding, and human–system interaction. Most recently he has been involved in the development of graphic interfaces for decision support and aiding systems, including artificial intelligence applications and workstations for database systems.