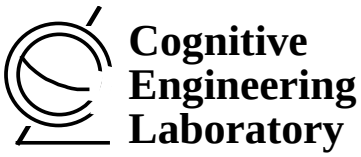

# Cognitive 'Dipsticks': Knowledge Elicitation Techniques for Cognitive Engineering Research

**Klaus Christoffersen, Christopher N. Hunter, and Kim J. Vicente**

**CEL 94-01**


Cognitive Engineering Laboratory University of Toronto Department of Industrial Engineering
4 Taddle Creek Rd. Toronto, Ontario, Canada M5S1A4
phone: (416) 978-0881 email: benfica@ie.utoronto.ca fax: (416) 978-3453




## COGNITIVE ENGINEERING LABORATORY

Director: Kim J. Vicente, B.A.Sc., M.S., Ph.D.

The Cognitive Engineering Laboratory (CEL) at the University of Toronto (U of T) is located in the Department of Industrial Engineering, and is one of three laboratories that comprise the U of T Human Factors Research Group. The CEL began in 1992 and is primarily concerned with conducting basic and applied research on how to introduce information technology into complex work environments, with a particular emphasis on power plant control rooms. Professor Vicente's areas of expertise include advanced interface design principles, the study of expertise, and cognitive work analysis. Thus, the general mission of the CEL is to conduct principled investigations of the impact of information technology on human work so as to develop research findings that are both relevant and useful to industries in which such issues arise.

### Current CEL Research Topics

The CEL has been funded by Atomic Energy of Canada, Ltd., Natural Sciences and Engineering Research Council, Defense and Civil Institute for Environmental Medicine, Japan Atomic Energy Research Institute, and Asea Brown Boveri Corporate Research - Heidelberg. The CEL also has collaborations and close contacts with the Westinghouse Electric Company, and Toshiba Nuclear Energy Laboratory. Current projects include:

- Studying the interaction between interface design and skill acquisition in process control systems.
- Understanding control strategy differences between people of various levels of expertise within the context of process control systems.
- Developing a better understanding of the design process so that human factors guidance can be presented in a way that will be effectively used by designers.
- Evaluating existing human factors handbooks.
- Developing advanced interfaces for computer-based anesthesiology equipment.

### CEL Technical Reports

For more information about CEL, CEL technical reports, or graduate school at the University of Toronto, please contact the address printed on the front of this technical report, or send email to Dr. Kim J. Vicente at <benfica@ie.utoronto.ca>.

ABSTRACT

Cognitive engineering research frequently requires one to investigate the content, structure, and form of what people know. This technical report presents a number of knowledge elicitation techniques which are believed to be potentially useful in examining the nature and extent of subjects' knowledge. Several procedures designed to act as process measures (how subjects do what they do) are presented followed in more detail by other procedures which are specifically designed to examine subject competencies (what subjects know). The techniques are presented in their original form and are then discussed in the context of cognitive engineering research. Specific examples are presented of how they have been applied within the context of research with a thermal-hydraulic process control simulation.

TABLE OF CONTENTS

## INTRODUCTION

One problem frequently encountered in cognitive engineering research is that of how to obtain information about the nature of what people know. This problem can be particularly difficult because the extent and structure of people's knowledge is often not apparent from their overt behaviour. Therefore, it is useful to have available a battery of methods which are specifically designed for eliciting subjects' knowledge. Many researchers in cognitive science, experimental psychology, and cognitive engineering have found it necessary to develop specialized procedures in order to gain insight into a particular aspect of the knowledge possessed by their subjects. This technical report represents a compilation of a number of these knowledge elicitation techniques. They were originally assembled by us for use in a long-term learning and adaptation experiment using the DURESS II system (see below). The procedures are explained and illustrated within the context of this experiment. However, we believe that they can be successfully adapted to a variety of other domains (e.g. interface design evaluations, training studies, etc.) for use in both longitudinal or cross-sectional studies in cognitive engineering.

### Overview

This paper will begin with a brief introduction to the DURESS II system in order to allow the reader to follow the examples of the methods presented. The remainder of the report will describe the rationale and background behind each of the procedures, as well as an example of its application.

## THE DURESS II SYSTEM

DURESS II (DUal REservoir System Simulation II) is an updated version of DURESS, a thermal-hydraulic process simulation that has served as a research vehicle for a number of studies on advanced interface design for process control systems (see Bisantz and Vicente, in press; Christoffersen, Pereklita, and Vicente, 1993; Vicente, 1992). The physical structure of DURESS II is illustrated in Figure 1. The system consists of two redundant feedwater streams that can be configured to supply water to two reservoirs. The goals are to keep each of the reservoirs at a prescribed temperature (40 $^{O}$C and 20 $^{O}$C), and to maintain enough water in each reservoir to satisfy each of the current demand flow rates, which are externally determined. To satisfy these system goals, there are eight valves, two pumps, and two heaters. DURESS II was modeled to be consistent with the laws of physics (e.g., the conservation laws), although several simplifying assumptions were made.

There are five basic types of tasks that subjects can be asked to perform in the DURESS II system. The first is what is referred to as a start-up task. This consists of bringing the system

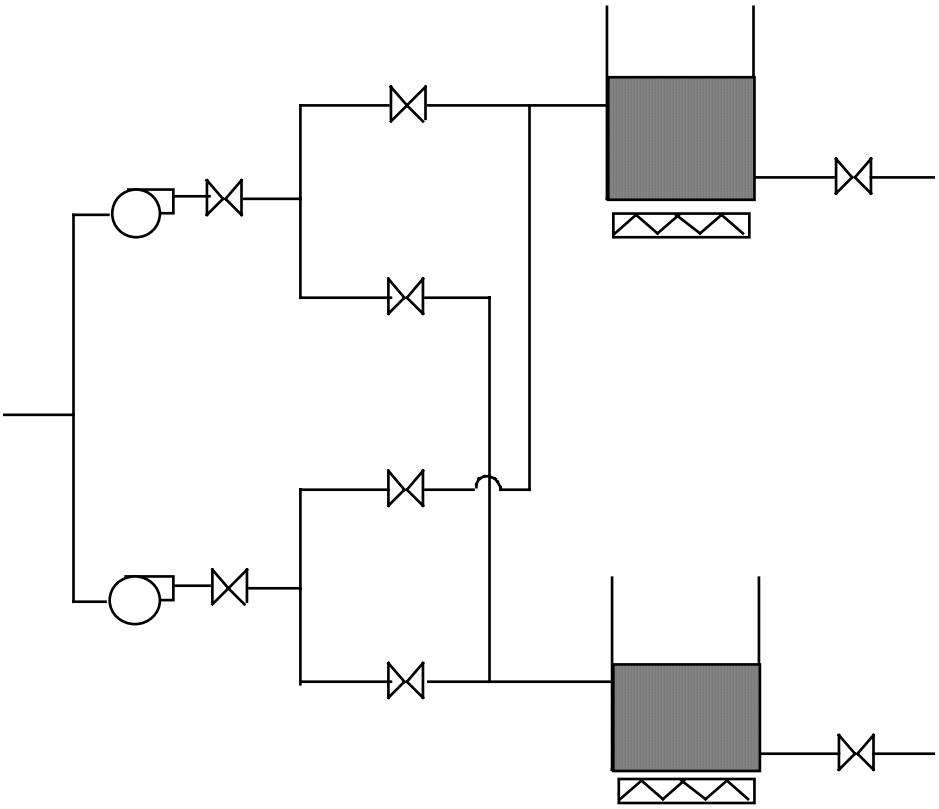


Figure 1. Physical structure of DURESS II.

from a shut-down state to a steady-state condition (operationally defined as a condition where the output flowrates and temperatures have been within the specified allowable ranges for five consecutive minutes). The second is to bring the system from this condition to a second steady-state condition given a new set of output flowrate setpoints (referred to as a tuning task). The third task is to bring the system from this second steady-state condition back to a shut-down condition. These three tasks together, in order, represent what is referred to as a standard trial.

The fourth type of task consists of a start-up task in which a routine fault is introduced into the system. Routine faults are faults which are relatively simple in nature and therefore comparatively easy to diagnose and compensate. They are meant to be analogous to recurring failures as might be observed in any industrial system where some of the system components are inherently less reliable than others. Three examples of such faults are a blocked valve, a heater failure, or a leak in one of the two reservoirs.

The fifth type of task is again a start-up task into which a non-routine fault is introduced. Non-routine faults are those which are more complex in nature. These can consist of an

occurrence of some combination of routine faults within the same trial, or of an unusual, single fault (e.g. a change in the temperature of the inlet water or the simulated presence of an external heat source). These faults are intended to be analogous to rare, unanticipated occurrences within a system which, although they can be compensated for, are more difficult to diagnose and in some cases more difficult to compensate for.

## PERFORMANCE MEASURES and KNOWLEDGE ELICITATION PROCEDURES

In any cognitive engineering study, it is important to have a framework with which to interpret subjects' behaviour. One such framework is based on the method of cognitive work analysis developed by Rasmussen (1986). This is an a priori analysis of the work domain, focusing on the constraints of the system and an understanding of the demands that these constraints place on operators' information processing capabilities. This analysis should be performed before any work is done on the experimental design since the constraints and properties of a given system must be fully understood before one can meaningfully decide what type of tasks to give subjects for the purpose of any given study. Moreover, a cognitive work analysis is essential because it provides a referent against which subjects' performance can be evaluated.

Rasmussen's framework suggests several possible levels of analysis. They include:

i) State of the work domain

ii) Control tasks (**product** measures - **what** the subjects do)

iii) Strategies and heuristics (**process** measures - **how** they do what they do)

iv) Operator competencies (what subjects **know** - mental models, rules, cues, skills, meta-knowledge)

Several studies performed in the context of DURESS and DURESS II have focused on levels i, ii, and iii. With respect to level i, data have been collected by recording a data log of all system variables over the course of each trial. With respect to levels ii and iii, the following techniques have been, or will be, employed:

I. Control Actions/Transitions

II. Action Transition Graphs

III. State-Space Diagrams

IV. Secondary Tasks

V. Grouping of Control Actions

VI. Manipulation of Required Task Performance

These are described briefly below. In this report, we are concerned primarily with level iv, for which a number of data-gathering techniques have been adapted in order to give insight into the

acquisition and evolution of subjects' skills and knowledge. These techniques are described in more detail below and include:

VII. Categorization
VIII. Domain Knowledge Test
IX. Interface Transfer
X. Verbal protocols
XI. Post-hoc Explanation
XII. Control Recipes

I. Control Actions/Transitions

Moray, Lootsteen, and Pajak (1986) suggest that tables comparing total number of operator actions and total number of control transitions are useful for providing information about subject performance. As well, time-history graphs of error scores (in terms of system states around the goal state or setpoint) can also provide meaningful data.

II. Action Transition Graphs

Moray et al. (1986) also suggest that action transition graphs are particularly useful for revealing changes in patterns of skill or performance in multivariate (complex) systems. These graphs are used to show an operator's control movements in a manner that readily indicates open or closed-loop activity. Moray et al. (1986) argue that these types of graphs, similar to the method of control actions/transitions above, are a revealing summary of the operators' understanding of the system because one can better understand the changes from closed to open-loop behavior as they learn about the system. Closed-loop behavior is indicated by multiple visits to a particular node, and as the behavior of the subject changes (improved performance), one sees less and less of these recurring visits to a particular node.

III. State-Space Diagrams

Sanderson, Verhage, and Fuld (1989) suggest using a state-space diagram for analysis. They define this as an *a priori* method of describing a system as "entering a discrete set of states that are similar to or different from goal states. The operator then has to choose efficient control actions to achieve the current goal state" (p. 1349). The operators are asked to move from point to point within this state-space, while their actions are shaped by the constraints of the system and the space itself. This type of analysis provides the number of moves made, the history of the subjects' control actions, and most importantly, the subjects' control actions in the context of the history of the system states experienced. It can be useful for describing simple performance

measures, and used retrospectively for describing operators' strategies as they move through the space.

## IV. Secondary Tasks

Wickens' (1992) multiple resource theory model emphasizes the relative independence of humans' verbal and spatial information processing resources. Many studies have exploited this distinction in order to better understand what type of cognitive resources are tapped by certain tasks. By intentionally loading one of these resources with a secondary task and examining the resulting effect on performance of the primary task, it is possible to infer which type of cognitive processing is being most heavily utilized by the primary task. For example, if one adds a spatial secondary task to the primary task and performance is found to degrade significantly, one might conclude that the primary task draws heavily on spatial resources. This technique can also be used to compare interfaces in terms of the extent to which they comparatively load on either verbal or spatial resources.

Pawlak (in press), for instance, has employed both verbal and spatial secondary tasks as loading tasks to evaluate the cognitive demands associated with two interfaces for DURESS II. The tasks chosen have been frequently used in basic experimental psychology research and are known to load on either verbal or spatial resources. Brook's (1968) mental imagery task was used as the spatial loading task. Brooks had subjects imagine that they were navigating around the outside of certain letters of the alphabet and asked them to state which direction (right or left) they were turning each time they came to the end of a straight edge. A verbal repetition task used by Saariluoma (1992) was used as the verbal loading task. Saariluoma asked subjects to continuously repeat a nonsense word while they performed the experimental task. Pawlak (in press) found these tasks to be sensitive measures of resource utilization, as a function of interface type.

It is important to use both types of secondary tasks since diametrically opposed results can potentially be observed if only one type of secondary task is adopted. This will happen if one interface loads more on spatial resources whereas the other loads more on verbal resources, a result obtained by Pawlak (in press).

## V. Grouping of Control Actions

Roenker, Thompson, and Brown (1971) describe several measures that have been developed to estimate clustering. Clustering refers to the degree to which a subject performs actions that are designated to be of a particular category in a consecutive fashion. For example, if a subject always performs actions of each category together, then one would say that perfect

clustering is present. The context in which these procedures are presented is that of basic psychological research (free recall tasks). However, the methods are relatively generic in terms of their potential applications. Roenker et al. (1971) present their own clustering index (the Adjusted Ratio of Clustering (ARC) score) which they claim is free of some of the limitations of other such measures in that it is normalized with respect to the number of categories used and the distribution of actions across categories. Using the ARC method, chance clustering always receives a score of zero and perfect clustering receives a score of one.

To use this method in cognitive engineering research, one must define an independent but exhaustive set of categories which can be used to classify actions. Although the method can be used with only one set of categories, more information can be gathered by defining several complementary ways of categorizing actions. For example, in DURESS II, one can categorize actions physically (e.g., all actions on any valve belonging to the same category) or functionally (e.g., all actions that affect a given reservoir belonging to the same category). Given a number of systematic ways of categorizing actions, one can then use ARC scores to examine how much clustering or organization is present in subjects' behaviour using each criterion adopted for categorization. This can then give clues about the strategies subjects are using and the implicit methods of categorization present in subjects' patterns of behaviour. All that one needs to perform such an analysis is a string representing the order in which subjects acted on the system components, with each action defined in terms of the category to which it belongs. Note, however, that this leads to a state-independent analysis, which of course, has its limits. For an application of this technique to cognitive engineering research, see Pawlak (in press).

## VI. Manipulation of Required Task Performance

In studies of human behaviour in complex systems, one often finds that performance measures such as completion time and total number of actions are rarely sufficient to discriminate between subject groups (Bainbridge, Beishon, Hemming, and Splaine, 1974), especially under non-fault conditions (Spencer, 1974). For instance, it is possible that two subject groups exhibit the same level of performance but that one group is exerting more effort to do so than the other. This phenomenon is well known in mental workload research (O'Donnell and Eggemeier, 1986).

As a result, it would be useful to have a measure which can help one identify such situations. One can develop such a measure by borrowing from the logic behind the use of primary task measures in mental workload assessment (O'Donnell and Eggemeier, 1986; Wierwille and Eggemeier, 1993). This literature shows that increasing task demands can sometimes lead to a decrease in task performance. As a result, one way to try to identify the situation described in the previous paragraph is to increase task demands. If one finds that one

subject group's performance degrades more than another group's, then there is evidence to show that there are in fact differences between the two groups in terms of the effort required to perform the task.

This technique will be used in a study with DURESS II by narrowing the tolerance on the temperature goals. Currently, subjects must maintain temperature within a 2 degree boundary centred around the setpoint temperature for 5 minutes before they are deemed to have reached steady state. With this set of task demands, performance measures do not show any difference between subjects using a traditional interface and those using an advanced interface under non-fault conditions (Pawlak, in press). To see if there is in fact a difference between interface groups, the task demands will be increased by reducing the acceptable tolerance on the temperature goals. Another way to do this would be to extend the amount of time that subjects had to maintain the temperature in the goal region. In other work domains, task demands may be increased by speeding up the tempo of events. In the study on DURESS II, we expect that making the task more stringent will allow us to uncover differences between interfaces that are not currently visible.

There are at least two points that need to be taken into account in using this measure. First, if task demands are made too stringent, then the task will become too difficult for all subjects, regardless of what group they are in, and performance will degrade uniformly. Second, it is possible that subjects can compensate for the increase in task demands by increasing the effort they invest, in which case one would still find no performance difference between groups. More detail on these considerations can be found in the discussion of primary task measures of workload in O'Donnell and Eggemeier (1986).

The measures discussed to this point are product and process measures of performance (see levels ii and iii, above). The remainder of the report will be devoted to describing various knowledge elicitation techniques designed to evaluate subjects' competencies (see level iv, above).

## VII. Categorization

The use of categorization as a performance measure was inspired primarily by the work of Chi, Feltovich, and Glaser (1981). In their study, groups of novices and experts were each given the task of sorting a number of physics problems into distinct categories (without solving the problems). Their hypothesis was that, when confronted with a problem, subjects construct a problem representation based on their domain knowledge. This representation is then compared with and classified according to a subject's available problem schemata which provides guidelines for solving that particular category of problems. Their results showed that there is a

fundamental difference in the ways that novices and experts classify physics problems. Novices tended to categorize problems on the basis of surface features ( i.e., superficial similarities between problems), whereas experts typically based their categorizations on the deep structure of a problem (i.e., the major physics principle required to solve the problem). In their study, they also included a subject of intermediate expertise who was found to exhibit a combination of the features of the novices' and experts' strategies. In the description of their EUREKA model, Elio and Scharf (1990) describe a similar shift from surface features to deep features in the method in which the model indexes and organizes information as it learns.

Within a longitudinal study, categorization can be used as a measure of subjects' evolving methods of problem representation, and the interaction between this evolution and a given type of experimental manipulation, e.g. the interface subjects are using. For instance, it may be shown that as subjects' expertise develops, their representations move from a strictly physical basis to a functional basis, reflecting the changes in their understanding/knowledge of the problem domain.

Two implementations of this technique within DURESS II are: categorization of components, and categorization of demand pairs. In classifying the components of DURESS II (2 pumps, 8 valves, 2 heaters, and 2 reservoirs), our expectation was that novices would initially form categories based on physical characteristics of the components (i.e. placing all the valves together, both pumps together, etc.). As they progress, we expect that they will begin to employ functional categories (e.g. placing all the components belonging to a particular reservoir subsystem together).

The second type of categorization problem was categorization of demand pairs, meaning the combination of particular pairs of demand setpoints in the upper and lower reservoirs of DURESS II. Our expectation was that novices would classify the demand pairs based on one of many possible surface features (e.g., both demands are even numbers). As their level of expertise increases, we expect that they might begin to classify the demand pairs based on the sort of valve configuration strategy that must be used in order to meet those setpoints. For example, some demand pairs can be satisfied by only one feedwater stream, others can be satisfied by having each feedwater stream supply water to only one reservoir (decoupling), whereas other demand pairs require subjects to adopt a configuration that results in a many to many mapping between feedwater streams and reservoirs (see Pawlak, in press).

Each type of categorization test was conducted using cards on which the components or demand pairs are drawn / written. The subjects were asked to iteratively sort the cards into piles and then, at the end of each test, to describe what criteria they used to define their categories.

## VIII. Domain Knowledge Test

One procedure which seems to suggest itself in many studies is to try to determine the extent of subjects' prior knowledge with respect to the domain of the experimental task. It may be desirable to control for this factor or it may be desired to choose subjects in such a way that this dimension is systematically varied as one of the variables of interest in the experiment. A third possibility is that it may be of interest to separately examine how much subjects' domain knowledge increases or decreases as a result of performing the experimental task. In any case, a measure is needed to give an idea of how much subjects know about the domain at any given point in time. Although interviews and questionnaires can provide valuable information about this, a somewhat less subjective measure is to administer a test of some sort.

Various studies in the context of DURESS and DURESS II (Vicente, 1992; Christoffersen, Pereklita, and Vicente, 1993; Pawlak, in press) have used a 20 question multiple choice questionnaire to test subjects' knowledge of thermal-hydraulics. Each question has five possible answers, only one of which is correct. The questions are all presented within the domain of the DURESS II system. It has normally been administered at the outset of the experiment immediately after subjects have been given a detailed technical description of DURESS II. Another study currently in progress is employing the pre-test at various stages during the experiment to see if test performance changes with experience at controlling the system. Pawlak (in press) used this test in a pre-test, post-test design and found that subjects who had experience at controlling DURESS II with an advanced interface improved their test scores more often that subjects who used a traditional interface. This result suggests that the advanced interface was more effective in allowing subjects to improve their knowledge of the system with control experience.

## IX. Interface Transfer

Robertson (1990) points out that transfer problems "have long been used as a measure of understanding." (p. 253). It seems to be largely accepted that subjects with a deeper understanding of a given domain will perform better on transfer problems than subjects with a less comprehensive understanding who are forced to rely on memorized algorithms to solve problems. Frensch and Sternberg (1989), dissent somewhat from this view. They argue that people with high levels of expertise in a given domain tend to become relatively inflexible in their modes of thinking, which can negatively affect their ability to deal with novel situations where some form of restructuring of their knowledge base is required to effectively solve the presented problem. However, this type of result is highly dependent on the type of transfer problem used.

Kossack (1992) used transfer as a measure in a study comparing two displays of the same system, one designed with the benefit of an ecological task analysis and one without. As the last

two trials in the experiment, subjects switched displays and were asked to perform the experimental task as usual. His expectation was that the people who had used the original display would do better when presented with the extra information in the enhanced display, and that the subjects who switched from the enhanced display to the original would get worse or at least remain constant in their level of performance. His results however, showed that both groups actually improved in their performance when they switched displays. The reasons for this effect are not clear. It may have been an example of a Hawthorne effect brought on by the switching of the displays. A second possibility is that the subjects who switched from the enhanced display to the unenhanced version were able to transfer the knowledge they had accumulated through the use of the enhanced display and use it even in the context of the other display. Also, there may have been the possibility that switching displays had no effect and that subjects were just continuing to improve with continued practice. Finally, it is possible that the two transfer trials were simply easier than the previous trials.

In contrast, two experiments cited by Frensch and Sternberg (1989) showed that performance on transfer tasks was negatively related to the amount of practice on the original task. In fact, subjects not only got worse, but took significantly longer to reach their previous level of performance once presented with the transfer task. Thus, the results of this measure can help to determine the degree to which subjects' actions have become proceduralized, and the degree to which the implementation of these procedures depends on the specific nature of the presented tasks. One can adapt this logic to evaluate subjects' knowledge by presenting them with transfer problems that consist of faults of various levels of difficulty (see above).

In studies such as that of Kossack (1992) where, for instance, two interfaces are compared, it is possible to use transfer problems to investigate how the interfaces affect subjects' knowledge, strategies, and performance. A couple of basic rules should be kept in mind when using this technique. First, in any long term study it is important to attempt to maintain subjects' motivation throughout the course of the experiment so as to avoid the appearance of a Hawthorne effect when transfer trials are presented. A potential solution is to offer a (non-trivial) monetary bonus for sustained good performance. While this may not completely eliminate a Hawthorne effect, it will hopefully make some strides toward minimizing it. Second, an attempt should be made to ensure that the transfer trials are of an approximately equal difficulty as the trials against which they are being compared.

## X. Verbal Protocols

In examining good vs. poor students using verbal protocols, Chi, Bassok, Lewis, Reimann, and Glaser (1989) found that good students tended to verbalize refined conditions for action and related these conditions to general principles, whereas poor students relied on

examples instead of applying general principles when justifying actions. Good students were also able to monitor themselves so that they could tell when they made a mistake (part of this was done by explaining encountered problems aloud to themselves); poor students could rarely tell when they had committed an error. Quantitatively more explanations were also given by the good students indicating an understanding of most facets of the task at hand. It was thus concluded that "explicit verbalization is the strictest criterion for assessing the understanding of a principle" (Chi et. al., 1989, pp. 145). Therefore verbal protocol could be used as a method for measuring expertise and to better understand cognitive processes.

In order to gain insight into expert knowledge during verbal protocol, a new situation or problem must be encountered (i.e. problems which are structurally but not conceptually unfamiliar) (Robertson, 1990, p. 253). By changing the structure of a problem, subjects cannot use memorized algorithms to solve the problem; they must use higher order knowledge (exhibit expert behaviour). When poor subjects are presented with a new problem then they tend to persist with their old methods instead of inventing new ones to accommodate the new situation (Frensch and Sternberg, 1989, p. 170). By observing verbal protocols when new problems are introduced, it can be determined if subjects are following some memorized set of rules to perform tasks or if they actually have a deep understanding of concepts. Using these conclusions, subjects' verbal protocols should be investigated only after a novel situation and analyzed for: self-monitoring of errors, fully explicated reasons for control actions relating to general principles, self-explanation of problems, and a greater verbalization of actions. Thus a deeper insight into cognitive processes in order to better understand expertise may be gained (Collins, Brown, and Newman, 1989).

In the context of DURESS II verbal protocols have been used to evaluate subjects' mental models of the system and how these changed over time. The protocols allowed the observation of how subjects apply their knowledge in order to find a solution. They also allowed comparisons of protocols for consistency with the post-hoc explanation (see following section).

The verbal protocol procedure required subjects to verbalize what they were thinking during the trial. They were asked to keep the following questions in mind ad hoc as they performed tasks on DURESS II (Pawlak, 1993):

1. How is the system behaving? Why?
2. What are you thinking about during the session?
3. When you make a control input, why did you do it?
4. What effect do you think a certain control input will have on the system?

These sessions were recorded using a video camera so that what was verbalized could be compared to the actual action being performed. Using this information, comparisons could be made between what subjects actually did and what they described in the post-hoc explanation test described next.

XI. Post-Hoc Explanation

As already mentioned, product measures alone cannot be reliably used to determine whether differences between subject groups exist. The example given was that subjects can be expending different amounts of effort to achieve the same level of performance. It is also possible that subjects achieve the identical level of performance using completely different cognitive processes. More specifically, the performance of one group of subjects may be driven by shallow, rote knowledge and that of the other group may be driven by a deep, theoretical understanding of the domain (see Ericsson and Harris, 1990 for a poignant example). Clearly, it is important to be able to distinguish between these two cases.

Some research on this topic has been conducted in the medical domain, where informal experiential knowledge is referred to as clinical knowledge and formal causal knowledge is referred to as biomedical knowledge. Boshuizen and Schmidt (1992) investigated the relationship between these two types of knowledge as a function of expertise by asking medical students, interns, and doctors to read a medical case study and make a diagnosis. They were asked to verbalize their thought processes as they performed the task and asked to give a written post-hoc explanation of their diagnosis after they had verbally reasoned through the problem. Boshuizen and Schmidt found that the interns used many detailed biomedical terms in their explanations whereas the doctors used less biomedical knowledge and more clinical knowledge instead. Based on these data alone, one would conclude that the interns had a better theoretical understanding than the doctors, perhaps because the deep knowledge of the latter had degraded over time. However, when subjects were asked to give a post-hoc explanation of the case, it was found that the doctors did in fact have more biomedical knowledge than the interns, although the doctors did not overtly refer to it during the on-line diagnosis. Thus, using both concurrent think-aloud and post-hoc explanation methodologies was instrumental in determining that experts had a deeper understanding of the domain than the interns. Thus, the post-hoc explanation is an efficient method of learning what type of understanding a subject has in a given domain. These findings indicate that the post-hoc test should be administered in addition to verbal protocols in order to elicit further information about experts' mental models not gained from verbal protocols alone.

A similar method for knowledge elicitation was investigated while researching the teaching of reading, writing and arithmetic (Collins, et al., 1989). It was found that a good

general test of student comprehension is to ask the student to summarize past learning. A strong correlation between summary quality and performance was observed. The results of this test thus formed a basis for comprehension monitoring. This further strengthens the rationale behind using the post-hoc explanation as a knowledge elicitation measure.

Using DURESS II the post-hoc test has been used (in addition to the verbal protocol test) to further understand subjects' mental models of the system and how these models changed over time. This test will be compared with the results of the verbal protocols to gain a better understanding of these mental models. It will also be used to contrast knowledge of subjects using the different interfaces.

The post-hoc test was given immediately following a fault trial. Subjects were asked to explain how the system performed during the trial and what they did to control the system. From this, the manner in which a subject detected, diagnosed, and compensated for the fault was observed. The post-hoc test was intended to measure if the subject saw the system as just a bunch of gauges that need adjusting (surface knowledge) or if s/he understood the underlying causes of the fault, how the fault affected in the system, and how to remedy the fault (deep knowledge).

## XII. Control Recipes

Experts tend to represent knowledge procedurally whereas novices have declarative knowledge representation. This was investigated by Hayes and Simon and by Anderson (VanLehn, 1988). They found that novices combined declarative knowledge with generic procedural knowledge representation in order to solve problems. After performing this laborious combination a more efficient procedural knowledge representation evolved. Given this finding, it would be useful to have an idea of the types of procedures subjects develop for controlling a system. Irmer and Reason (1991) asked subjects to give a written description of the procedure they use for controlling the system. The type of control 'recipe' subjects provide and the time it takes them to do so may provide some insight into subjects' knowledge.

The goal of this test is to gain as much information as possible about subjects' strategies and learning processes over the course of an experiment. It is intended to measure how well the subject understands the task by looking at what type of procedure they give. In general, it can be expected that the more knowledgeable subjects will give more precise recipes whereas poorer subjects will give imprecise descriptions as well as verbalizing problems they had with the system and other extraneous information (Irmer and Reason, 1991, p. 18). Therefore, this test attempts to understand the rules and strategies that the subject is using to perform a given task.

In the context of DURESS II subjects were asked to describe what procedure or recipe they would use to control the system for a given operation. Unlike the post-hoc test, the recipe test required the subject to give a procedure rather than explanation. This was made very clear

to the subject . For example they were told: "You must write a detailed procedure of how you operate the system (for a given task) so that this procedure could be used by someone who has never seen the system".

The recipe test will also be performed before the experiment started (though after the initial introduction to DURESS II) to try and elicit how the subject perceived the system a priori. It will also be given after the last trial to compare subjects' perception of DURESS II a posteriori. The subject's use of his/her recipe will be validated by comparing it with his/her trial on the following day. The subjects were asked to write down their recipe, and it was then analyzed for understanding and depth of knowledge of the DURESS II system. No trials will be given on the day of the recipe test in order to minimize reinforcement effects. As a rough control condition, a general copying exercise (e.g. "Write out the following: ÔThe quick brown fox jumped over the lazy dog'") was timed for each subject to allow the times taken to write out the recipes to be normalized with respect to individual writing speed.

## CONCLUSION

This paper has presented a collection of some of the methods which are available for use in cognitive engineering research where it is desired to gain an insight into the extent and structure of people's knowledge. Using the aforementioned techniques, we believe it is possible to examine the state of subjects' knowledge at a particular point in time or, by using these techniques repeatedly, to examine the process of skill acquisition and evolution of subjects' knowledge over an extended period of performing a given experimental task. Although they were originally compiled for use in a long term learning and adaptation study using the DURESS II system, we firmly believe that these techniques could prove effective in many other environments and studies where a detailed understanding of subjects' knowledge and mental processes is sought.

## ACKNOWLEDGEMENTS

We would like to thank Randy Mumaw for helpful discussions. This work was sponsored by a contract from the Japan Atomic Energy Research Institute (Dr. Fumiya Tanabe, Contract Monitor), a research grant from the Natural Science and Engineering Research Council of Canada, both awarded to the last author, and a graduate scholarship from the Natural Science and Engineering Research Council of Canada awarded to the second author.

REFERENCES


Bainbridge, L., Beishon, J., Hemming, J. H., and Splaine, M. (1974). A study of real-time human decision-making using a plant simulator. In E. Edwards and F. P. Lees (Eds.), The human operator in process control (pp. 67-78). London: Taylor Francis.

Bisantz, A. M., and Vicente, K. J. (in press). Making the abstraction hierarchy concrete. International Journal of Human-Computer Studies.

Boshuizen, H. P. A., and Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. Cognitive Science, 16, 153-184.

Brooks, L. R., (1968). Spatial and verbal components of the act of recall. Canadian Journal of Psychology, 22, 349-368.

Chi, M. T. H., Feltovich, P. J., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Chi, M. T. H., Glaser, R. , and Farr, M. J. (1988). The nature of expertise. Hillsdale, NJ: Lawrence Erlbaum Associates.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., and Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science, 13, 145-182.

Christoffersen, K., Pereklita, A. J., and Vicente, K. J. (1993). Effects of expertise on reasoning trajectories in an abstraction hierarchy: Fault diagnosis in a process control system (Tech. Report CEL 93-02). Toronto, Canada: University of Toronto, Cognitive Engineering Laboratory.

Collins, A., Brown, J. S., and Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resvich (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser (pp. 453-495). Hillsdale, NJ: Lawrence Erlbaum Associates.

Elio, R., and Scharf, P. B. (1990). Modeling novice-to-expert shifts in problem-solving strategy and knowledge organization. Cognitive Science, 14, 579-639.

Ericsson, K. A., and Harris, M. S. (1990). Expert chess memory without chess knowledge: A training study. Paper presented at the 31st Annual Meeting of the Psychonomic Society. New Orleans, LA.

Frensch, P. A., and Sternberg, R. J. (1989). Expertise and intelligent thinking: When is it worse to know better? In R. Sternberg (Ed.), Advances in the psychology of human intelligence, vol. 5, (pp. 157-188). Hillsdale, NJ: Lawrence Erlbaum Associates.

Irmer, C., and Reason, J. T. (1991). Early learning in simulated forest fire-fighting. In Simulations, Evaluations, and Models: Proceedings of the Fourth MOHAWC Workshop (pp. 1-20). Roskilde, Denmark: Risø National Laboratory.

Kossack, M. F. (1992). Ecological task analysis: a method for display enhancement, (unpublished master's thesis). Georgia Institute of Technology, Atlanta, Georgia.

Moray, N., Lootsteen, P., and Pajak, J. (1986). Acquisition of process control skills. IEEE Transactions on Systems, Man, and Cybernetics, SMC-16, 497-504.

O'Donnell, R. D., and Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, and J. P. Thomas (Eds.), Handbook of human perception and performance, volume II: Cognitive processes and performance (pp. 42-1 - 42-29). New York: Wiley.

Pawlak, W. S. (1993). Complex system simulations: Experimentation and analysis. Unpublished manuscript. University of Toronto, Cognitive Engineering Laboratory, Toronto, Canada.

Pawlak, W. S. (in press). Inducing effective control strategies through ecological interface design. Unpublished M. A. Sc. thesis, University of Toronto, Toronto, Canada.

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

Robertson, W. (1990). Detection of cognitive structure with protocol data: Predicting performance on physics transfer problems. Cognitive Science, 14, 253-280.

Roenker, D. L., Thompson, C. P., and Brown S. C. (1971). Comparison of measures for the estimation of clustering in free recall. Psychological Bulletin, 76, 45-48.

Saariluoma, P. (1992). Visuospatial and articulatory interference in chess players' information intake. Applied Cognitive Psychology, 6, 77-89.

Sanderson, P. M., Verhage, A. G., and Fuld, R. B. (1989). State-space and verbal protocol methods for studying the human operator in process control. Ergonomics, 32, 1343-1372.

Spencer, J. (1974). An investigation of process control skill. In E. Edwards and F. P. Lees (Eds.), The human operator in process control (pp. 67-78). London: Taylor Francis.

VanLehn, K. (1988). Problem solving and cognitive skill acquisition. In M. I. Posner (Ed.), Foundations of cognitive science (pp. 527-573). Cambridge, MA: MIT Press.

Vicente, K. J. (1992). Memory recall in a process control system: A measure of expertise and display effectiveness. Memory & Cognition, 20, 356-373.

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

Wierwille, W. W., and Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. Human Factors, 35, 263-281.