Supporting Operator Problem Solving Through Ecological Interface Design

Kim J. Vicente, *Member, IEEE*, Klaus Christoffersen, and Alex Pereklita

***Abstract*— This paper describes two experiments evaluating ecological interface design (EID), a novel theoretical framework for the design of interfaces for complex human-machine systems. According to EID, to properly support operator problem solving activities, an interface should display both the physical and functional properties of the work domain in the form of a multilevel representation based on the abstraction hierarchy. To evaluate this claim, two interfaces for a thermal-hydraulic process simulation were developed, one based on a traditional format containing only physical information (P) and another based on EID which also contained information about higher-order functional variables (P+F). The findings of Experiment 1 are consistent with the claim that an interface based on an abstraction hierarchy representation can provide more support for problem solving than an interface based on physical variables alone, thereby providing some initial support for the EID framework. There was also some evidence to indicate that theoretical expertise is required to enjoy the full benefits of the P+F interface. The findings of Experiment 2 indicate that subjects who exhibited effective diagnosis performance using the P+F interface tended to start their search at a high level of abstraction and gradually work their way down to more detailed levels, as predicted. Furthermore, previous experience with the DURESS system was found to be the most reliable background variable that predicted performance.**

I. Introduction

COMPLEX human-machine systems pose a strong set of constraints on effective interface design. Because such systems usually consist of many components that interact, effective interface design must go beyond specification of surface features and must focus on the domain semantics [1], [28]. This puts a great demand on designers to understand and represent work domain constraints, so that they can identify the information and relations that need to be included in the interface. For complex systems, this is an onerous demand. Moreover, such systems are also characterized by events which are unfamiliar to operators and that have not been anticipated by designers [16]. In fact, it is during these circumstances that the threat to system safety is at its peak. Thus, it is also important that an interface provide operators with the information needed to cope with such unanticipated events. Traditional interface design practices do not address these features of complex human-machine systems, and so an alternative is needed.

Ecological interface design (EID) is a theoretical framework for interface design that is intended to address some of the unique properties of complex human-machine systems [26]. The framework, based on Rasmussen's skills, rules, knowledge (SRK) taxonomy of levels of cognitive control [17], consists of three prescriptive design principles, each directed at providing the appropriate interface support for a specific level of cognitive control. In general terms, the goals of EID are twofold: to allow operators to effectively exploit their powerful perception and action capabilities, while also providing the support required for problem solving activities. This paper presents two experiments which evaluate how well the principles of EID achieve the latter goal.

There are several reasons to focus on support for problem solving, or knowledge-based behavior (KBB). The most challenging situations that operators face are those where the plant is in an unfamiliar and unanticipated abnormal state. Under these circumstances, operators are required to improvise a solution, typically by relying on KBB since the event is an unfamiliar one. Thus, it is critical that the interface provide the necessary support for such activities, particularly if one considers that: a) it is under these conditions that the potential for disaster is greatest, and b) that KBB is an effortful and error-prone activity under the best of circumstances [26].

The remainder of this section describes the theoretical framework guiding this research, the process simulation that served as a research vehicle, and the two interfaces that were used. This information will provide the background necessary to interpret the two experiments discussed in the remainder of the paper.

A. Theory

According to EID, an interface needs to represent the work domain as an *abstraction hierarchy* [18] if it is to properly support KBB. This multilevel representation format describes the various layers of constraint in the work domain. Each level represents a different model of the system. For many complex human-machine systems, five levels of constraint have been found to be of use: the purposes for which the system was designed (Functional Purpose); the intended causal structure of the process in terms of mass, energy, information, or value flows (Abstract Function); the basic functions that the plant is designed to achieve (Generalized Function); the characteristics of the components and the connections between them (Physical Function); and finally, the appearance and

Manuscript received January 21, 1994; revised June 3, 1994. This work was supported by a Connaught New Staff Grant from the University of Toronto and by a Research and an Equipment grant from the Natural Sciences and Engineering Research Council of Canada awarded to the first author.

The authors are with the Cognitive Engineering Laboratory, Department of Industrial Engineering, University of Toronto, Toronto, Ontario, Canada M5S 1A4.

IEEE Log Number 9406635.

spatial location of those components (Physical Form). Higher levels represent *functional* information about system purposes, whereas lower levels represent *physical* information about how those purposes are realized by plant components.

There are two advantages to adopting the abstraction hierarchy as a basis for interface design [26]. First, this approach allows one to identify, *a priori*, the information needed to cope with events which are unfamiliar to operators and which have not been anticipated by designers. This is a very important property since, as mentioned, such unanticipated events pose the greatest threat to system safety in complex systems. Whereas traditional approaches to interface design rarely make an attempt to deal with this problem, the abstraction hierarchy was explicitly designed to deal with unanticipated events. Second, the abstraction hierarchy is also a psychologically relevant problem representation. There is a significant body of empirical research from a number of quite diverse domains showing that problem solving protocols can be mapped onto an abstraction hierarchy representation (see [19] and [26] for reviews). Thus, in addition to satisfying the engineering requirement of containing the information needed to cope with unanticipated events, the abstraction hierarchy also satisfies the psychological requirement of providing a representation that is consistent with operators' problem solving processes.

B. Research Vehicle

The present research was conducted within the context of DURESS (DUal REservoir System Simulation), a thermal-hydraulic process simulation that was designed to be representative of complex human-machine systems [22]. The system, illustrated in Fig. 1, consists of two redundant feedwater streams, each consisting of a pump and three valves, which can be configured to supply water to two reservoirs. The system goals are to keep each of the reservoirs at a prescribed temperature (40°C and 20°C), and to maintain enough water in each reservoir to satisfy each of the current externally determined demand flow rates (D1, D2). The means available for control are six valves (VA, VA1, VA2, VB, VB1, VB2), two pumps (PA, PB), and two heaters (H1, H2). (The version of the simulation used here did not have the capability for real-time control). The temperature (T1, T2) and volume (V1, V2) of the two reservoirs are also displayed.

The representation in Fig. 1 is the Physical (P) interface that was included as a control condition in Experiment 1, below. It contains information at three levels of the abstraction hierarchy: Functional Purpose representing the goal variables (volume, demand, and temperature); Physical Function representing the state of all of the pumps, valves, and heaters; and Physical Form representing the topographic layout of the system components. Functional information defined by the levels of Generalized Function and Abstract Function is absent from this interface (see [25] for an abstraction hierarchy of DURESS).

To evaluate the utility of the EID framework, a second interface based on the principles of EID was constructed. This interface, illustrated in Fig. 2, will be referred to as the P+F interface since it contains the physical information in

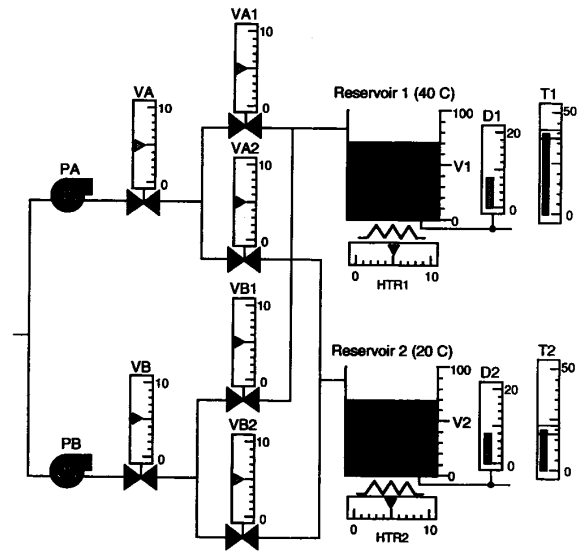


Fig. 1. Physical (P) interface for DURESS.

the P interface as well as higher-order functional information. The design rationale behind this int erface is described in detail in [25], and so only a brief overview will be provided here. The first step was to develop an abstraction hierarchy representation for DURESS. This analysis indicated that there were two levels of information missing from the P interface, Generalized Function (representing flow rates and heat transfer rates), and Abstract Function (representing mass and energy balances). The second step was to embed these added levels of information into the new interface. This was done by mapping the underlying domain relationships onto geometric forms that are easy to perceive. For example, a quadrangle was designed to represent the mass balance state equation. The key point is that there is an isomorphic mapping between the variables and relations in the state equation and the perceptual properties of the quadrangle (e.g., length, area, and slope). Similar mappings were developed for other domain relationships. *Note that the design of this interface is based on a representation of domain constraints that are relevant to the purposes for which the system was designed, not on operator tasks or specific events.* Again, for more detail on the design of the P+F interface, see [25].

The P+F interface contains all levels of the abstraction hierarchy. Beginning at the top level of Functional Purpose, the demand (D1, D2) and temperature (T1, T2) setpoints are represented in the interface. For the temperature settings, the upper and lower limits around the setpoints (40°C and 20°C) are shown as vertical lines on the two temperature scales (T1 and T2, respectively). The level of Abstract Function is represented by the group of graphics on the right. This portion of the display will be described below. At the level of Generalized Function, the flowrates in each feedwater stream (e.g., FVA, FPA, FA1, FA2) and the heating rates (e.g., HTR1) are displayed as bar scales. At the level of Physical Function, the valve settings (e.g., VB) and heater settings

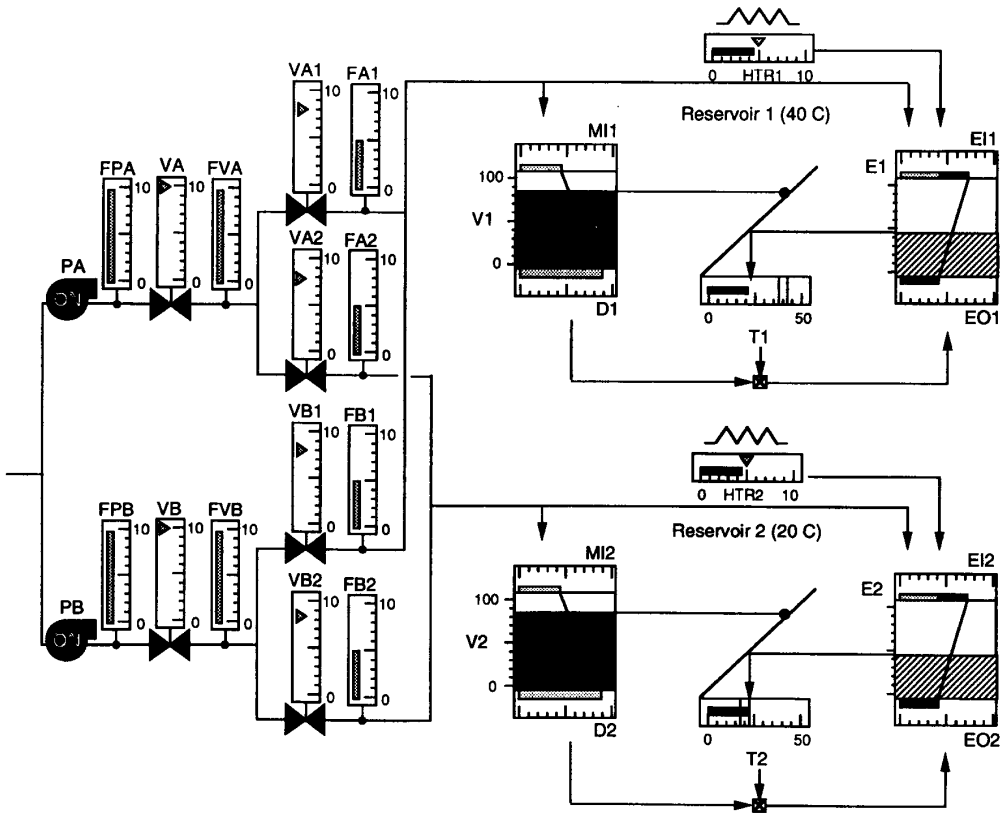


Fig. 2. Physical/Functional (P+F) interface for DURESS.

(e.g., HTR2) are indicated by the small triangular pointers on the respective scales. Since the pump settings (e.g., PB) are discrete (either ON or OFF), they are directly labeled on the pumps themselves. Finally, at the level of Physical Form, the location of the components and the connections between them are also represented.

With regard to the level of Abstract Function, the rectangular graphic on the left represents the mass balance for the reservoir, while the graphic on the right represents the energy balance. Both operate in a similar manner. Referring to Reservoir 1, the inputs are shown at the top (e.g., MI1 for the mass and EI1 for the energy), the inventories on the side (e.g., V1 for volume, or mass, and E1 for energy), and the outputs at the bottom (e.g., D1 for demand, or mass, and EO1 for energy). The energy inputs (EI1 and EI2) are partialled out according to the two contributors. The energy added by the feedwater is shown as the lightly shaded bar, while the energy added by the heater is shown as the dark bar. Intuitively, these energy and mass graphics rely on a funnel metaphor. Thus, if the bottom is wider than the top (i.e., output greater than input, as with the mass balance for Reservoir 1 in Fig. 2), then it is easy to visualize the consequence, namely that the volume should be decreasing. Thus, the slope of the line represents the rate at which the mass (or energy) inventory is changing. If input equals output, then the line would be perpendicular, indicating that the inventory should not change.

The graphic in the middle, between the mass and energy balances, illustrates the structure of the relationship between volume, energy, and temperature. The horizontal bar with a ball on the end that emanates from the current volume level is of fixed length. The height of this bar always accompanies any change in volume (i.e., the bar will always be at the same height as the water level, V1 or V2). The thick diagonal line in the center display is always tangent to the ball on the edge of the horizontal bar. Thus, a change in the vertical position of the horizontal bar serves to change the slope of the line in the center display. For example, if volume increases, the horizontal bar goes up, causing the diagonal to rotate counterclockwise, and thereby increasing the slope of the diagonal line. The slope of the diagonal represents the function that maps the amount of energy onto temperature. This mapping is indicated by the line from the energy inventory (E1, E2) that comes across and reflects off the diagonal and down onto temperature (T1, T2).

C. Overview

The remainder of this paper describes two experiments evaluating the extent to which an interface based on the principles of EID can support KBB. Experiment 1 compared the P and P+F interfaces for DURESS in the context of a combined diagnosis and memory task. In addition, the interaction between interface type and expertise was also

investigated. Experiment 2 was designed to investigate the reasons for the advantages that the P+F interface exhibited in Experiment 1. Thus, process tracing measures [27] were used to understand the cognitive processes involved in using the P+F interface during a diagnosis task. Furthermore, the effects of various dimensions of subjects' expertise were investigated to determine which dimensions were reliable predictors of performance, both in terms of product and process.

II. Experiment 1

Experiment 1 was designed to serve two purposes, one relevant to basic psychological issues (the relationships between memory recall performance and expertise and interface design) and the other relevant to more pragmatic cognitive engineering issues (the effectiveness of the EID framework). The findings and discussion pertaining to the former issue were presented in [23]. The presentation here focuses solely on the findings relevant to the latter, more pragmatic set of issues. Since the method of the experiment has already been presented at great length in [23], only an abbreviated description will be provided here.

A. Hypotheses

The primary theoretical proposition evaluated in Experiment 1 was that an interface should reveal the structure of the work domain in the form of an abstraction hierarchy if it is to properly support KBB. If this is so, then one would predict that the P+F interface should result in superior diagnosis performance when compared to the P inte rface. In addition, it is also expected that the predicted interface advantage should be mediated by subjects' level of theoretical expertise, with expert subjects showing a larger advantage. The rationale here is that an understanding of the physical principles governing the behavior of the process is required in order for subjects to be able to interpret, and fully exploit the benefits of, the added levels of information in the P+F interface (see [21] for a more detailed justification).

B. Method

1) Experimental design: A 2 × 2 × 2 × 2 × 2 mixed factorial design with two within-subjects factors (Interface and Trial Type) and three between-subjects factors (Expertise, Order, and Sequence) was adopted for this experiment. There were two levels of Expertise: Experts and Novices. The Order factor, which refers to the order in which the subjects were exposed to the two interfaces, had two levels: P interface first, and P interface second. The Sequence factor refers to the order in which the 4 blocks of 10 scenarios were presented to subjects. There were two levels: Forward (Block 1 to Block 4) and Backward (Block 4 to Block 1). The two within-subjects factors were factorially crossed and nested within each of the eight subject groups. As mentioned above, Interface had two levels: P and P+F. The P interface contained 16 variables corresponding to the states of the physical components (see Fig. 1), whereas the P+F interface contained 34 variables representing both physical and functional variables (see Fig. 2). There were also two Trial Types: Semantic and Random (see description of Trial Types, below). Subjects used each interface for two successive sessions. Each session consisted of 10 trials, with 5 replications of each Trial Type.

2) Experimental task: On each trial, a dynamic, real-time event sequence of the behavior of DURESS was presented for a duration varying from 25 to 30 seconds. These brief exposure times made the task a challenging one so as to maximize the chances of detecting interface effects. Thus, the important question is the relative performance differences between interface groups, not the absolute level of performance achieved by either group. Subjects viewed the scenario and tried to understand and remember as much as they could of what took place. In the instructions, subjects were told to concentrate on understanding rather than rote recall. While the event was being presented, no response was required. Once the event ended, the screen went blank and then a recall screen was automatically displayed.

The recall screen contained the 34 process variables which were represented in the P+F interface. Subjects were required to estimate the final value each of these variables had at the end of the preceding scenario[1]. The procedure was the same regardless of which interface the subject was using. This means that, for the P interface, subjects were asked to estimate the values of variables that were not displayed. The reason for asking subjects to do this was to determine whether it was possible for them to derive the higher order functional variables from the physical variables that were displayed in the P interface. See [23] for more details about the memory recall procedure.

Once the recall procedure was terminated, subjects answered a set of structured questions evaluating their diagnosis of the previous event[2]. The following questions were posed:

1) Was the scenario consistent with your understanding of the functional principles governing DURESS' behavior (admitting the possibility of a fault)? (If NO, then stop).
2) Did a fault or disturbance occur in the system during this scenario? (If NO, goto 4).
3) Describe the fault in as much detail as you can. Where was the fault? What did it consist of? (Stop here.)

[1] The rationale for adopting the memory recall measure is discussed at length in [21] and [23]. The basic idea is as follows. Previous research on expertise has shown that memory recall performance is a measure of domain understanding, with experts remembering more than novices after a brief presentation of the stimulus material. In this study, this finding was applied to evaluate displays. A poor display should impede subjects' understanding of system state, whereas a good display should enhance such understanding. As a result, one would expect that the better display would lead to better memory recall performance. For the data bearing on this hypothesis, see [23].

[2] In both of the experiments presented here, subjects were diagnosing canned scenarios, and not actively interacting with the system in a closed loop fashion. It is important to distinguish between these two situations since closed loop control provides subjects with a much richer set of means to manage a fault (e.g., acting on a component and observing the results). As a result, our findings cannot be easily generalized to closed-loop fault management. However, this does not mean that diagnosis, as studied here, is meaningless and not worthwhile studying. Diagnosis is a subset of the cognitive activities that make up closed-loop control and can play a significant role in fault management. Thus, the task conditions investigated in this paper, while certainly restricted compared to an interactive control task, are meaningful and worthwhile investigating on their own, as long as one does not blindly generalize from one situation to the other.

4) Given that there was no fault, provide a detailed functional description of what you observed.

In the instructions, subjects were only told that there would be 3 types of trials: scenarios 1) that exhibit a normal pattern of behavior according to physical principles; 2) that have a single fault or disturbance; and, 3) where the process variables would not be driven by a simulation of DURESS (on these trials, the behavior of the system would not obey physical laws). They were also told that there would be no trials with multiple faults and/or disturbances, and no trials with sensor failures. Note that subjects were not told what types of faults could appear, nor what the ratio of fault to normal to random trials were. Knowledge of results was not provided at any point during the experiment.

3) Subjects: Expert subjects were graduate students in either Mechanical or Nuclear Engineering. Following the rationale discussed earlier, these subjects were theoretical experts, not experts at controlling the system. In this study, theoretical experts are conceived as having a good conceptual understanding of the functional structure of the system. This construct was operationally defined and tested using a thermal-hydraulics pre-test to be described below. Novices were graduate students who had never been enrolled in a science or engineering major. Expert subjects had taken an average of 5.73 graduate or undergraduate physics courses (range of 3 to 16) and 5.09 graduate or undergraduate thermodynamic or thermal-hydraulic courses (range of 3 to 9). In contrast, Novices averaged 0.75 physics courses (range of 0 to 2). No Novice had ever taken a thermodynamic or thermal-hydraulic course. The two subject groups were roughly equivalent on other demographic attributes. There were 12 subjects in each group, 2 females and 10 males. Subjects were paid a total of $24 for participating in the experiment.

4) Apparatus: The presentation of the scenarios and the subsequent recall procedure was conducted on a Zenith PC compatible microcomputer equipped with a Motorola 80386 CPU, a math coprocessor, a PC mouse, an EGA graphics card, and a NEC Multisync II color monitor. Both the P and P+F interfaces were in color. The scenarios were generated off-line on a simulation of DURESS developed at Risø National Laboratory. The simulation was written in PC-DYSIM, a software package developed at Risø for the simulation of continuous dynamic processes [11].

5) Trial types: Each trial consisted of a dynamic event sequence illustrating DURESS' behavior. The settings of the components did not change during the trial. Thus, the trajectory followed by the process variables was determined solely by the initial conditions and the particular fault (if any) introduced into the simulation, and not by any action taken on the system components. When present, faults were injected at an arbitrary point within the first 10 seconds of the scenario. During all of the trials, the pumps and valves in the two feedwater streams were configured in such a way that each stream was supplying water to both reservoirs.

There were five different types of trials in the Semantic condition, each occurring once within a session. In the *steady state* trial type, there were no changes in any of the system variables. In the second condition, there was a *change in reservoir volume* (either an increase or a decrease) caused by a difference between the mass input flow rate and the current demand. With the third trial type, a *reservoir leak* was introduced in one of the two reservoirs. This means that the volume gradient would be less than it should be, given the current input and output flow rates for the reservoir in question. The *blocked valve* trial type resulted in a complete blockage of one of the six valves. The effect was to reduce the flow through the affected valve to zero, thereby decreasing the supply of water to the reservoir(s) to which the failed valve was connected. The displayed state of the valve did not change. The fifth trial type consisted of a *change in inlet water temperature* (either an increase or a decrease). This caused a corresponding change in the temperature of both reservoirs. The first two trial types were normal scenarios, whereas the last three types were fault scenarios.

The Random scenarios contained the same set of trajectories as the Semantic scenarios except that trajectories that once belonged together were placed in different scenarios. In this way, the average temporal distribution properties of the variables displayed in the Random scenarios were identical to those of the Semantic scenarios. The primary difference between the two trial types was that the Random scenarios did not obey the laws of physics, whereas the Semantic scenarios did. Consequently, random trials exhibited nonsensical patterns that were characterized by multiple, simultaneous violations of system constraints during a single trial[3].

6) Procedure: The experiment consisted of one introductory session followed by four data collection sessions, each conducted on a different day. Subjects performed the task with one interface for two sessions, and then with the other interface for another two sessions. The order in which the two interfaces were presented was counterbalanced. The entire experiment lasted from 4 to 6 hours.

During the first session, subjects were given a general introduction to the experiment and read a brief description of the physical properties of the DURESS simulation. They then took a pre-test of thermal-hydraulics knowledge, consisting of 20 multiple-choice questions couched within the context of DURESS, to evaluate their theoretical knowledge of thermal-hydraulics. A maximum of 30 minutes was allotted for taking the test. Finally, subjects were given descriptions of the 34 variables in DURESS, the labels that were used to identify each variable throughout the experiment, and the procedure for recalling the state of these variables. Subjects also practiced the recall procedure.

[3]The random trial manipulation was included to determine if performance on normal trials was solely due to the differences in form between the two interfaces (as opposed to content). According to EID, the P+F interface should lead to better performance than the P interface, in part, because it contains added levels of abstraction not found in the P interface. Making the trials random removes this content because the higher-order relationships that usually exist between variables are violated. Thus, the benefit of added content is removed. However, making the trials random does not affect the form of the two interfaces in any way. Thus, if performance differences between the two interfaces on normal trials are merely due to differences in form, then making the trials random should not affect performance. In fact, memory recall performance on random trials was significantly worse than on normal trials, thereby supporting the claim that the differences in content between the two interfaces are important [23].

At the beginning of the first session with each interface, subjects were introduced to the interface they would be using for the next two sessions. For the first data collection session only, the experimental task was explained to subjects. Each session consisted of 10 trials. Since the primary comparison of interest for trial types was between meaningful and non-meaningful trials (see footnote 2), and not between the specific trial types listed above, half of the trials were Semantic and half were Random. The order of the trials was randomized within a session with the added constraint that no more than three Semantic or Random trials appear successively. Each event sequence was presented only once to each subject. All subjects received the same 40 scenarios in the two sequences described earlier.

7) Performance measures: The primary performance measure reported here is diagnosis accuracy. Three levels of analytic resolution were adopted evaluating how well subjects could discriminate: 1) Random from Semantic trials; 2) Random from Normal from Fault trials; and finally, 3) the exact trial type (see description of Trial Types, above). The categories in Levels 1 and 2 are exhaustive and mutually exclusive so there is no ambiguity. At Level 3, subjects had to give a correct description of the specific event similar to the labels for the trial types provided above if their diagnosis was to be scored as correct. It was not sufficient to merely describe the variable values observed on the screen. Scoring was conducted by the first author. The various methods used to measure memory recall performance are described in [23].

B. Results

The analyses presented here are based on data collected from the second session with each interface only, since the first session with each interface served as practice. Only the results pertinent to the hypotheses presented earlier are described here. For additional analyses and results, including traditional analysis of variance, see [22], [23]. In cases where within-subject comparisons were of interest, data were analyzed according to the method recommended by Hammond, Hamm, Grassia, and Pearson [9]. This involves evaluating the predicted result for each and every subject, and then aggregating over subjects to determine how many conformed to the prediction. This procedure provides a more meaningful analysis than the traditional comparison of potentially statistically "mythical" group means.

This section is divided into 3 subsections: expertise effects, diagnosis accuracy, and correlation between diagnosis and memory.

1) Expertise effects: The results of the thermal hydraulics pre-test will be described first. Experts' pre-test scores ranged from 10 to 18 with a mean of 14.67. Novices' test scores ranged from 4 to 12 with a mean of 9.58 (out of a maximum of 20). A Mann-Whitney U test [20] indicated that this difference in means is statistically signific ant ($U = 10$, $p < 0.002$).

Two conclusions can be derived from this analysis. First, the experts clearly outperformed the novices, thereby validating the selection criterion that was adopted for defining the two subject groups. Second, the knowledge of experts is not completely accurate, as evidenced by the fact that no subject attained a perfect score.

TABLE I
RESULTS FROM SIGN TEST ANALYSIS OF DIAGNOSIS ACCURACY FOR EXPERTS AND NOVICES

LEVEL	P+F > P	TIES	$p<$
a) EXPERTS ($n = 12$)			
1	9	1	0.066
2	9	3	0.004
3	11	1	0.001
b) NOVICES ($n = 12$)			
LEVEL	P+F > P	TIES	$p<$
1	7	3	n.s.
2	6	1	n.s.
3	7	1	n.s.

Several other informal observations shed further light on the nature of the expertise effects in Experiment 1. It seemed that there were substantial differences between subjects in the Expert condition. The best subject, for instance, derived the state equations describing the system dynamics from first principles upon reading the verbal description of DURESS! He then used these equations to answer the pre-test questions. On some trials, this subject's memory recall performance was on the order of 2% error. This is a remarkable achievement when one considers that subjects were being asked to remember the final values of 34 variables that were changing and only visible for 25 to 30 seconds. One or two other expert subjects exhibited this level of proficiency on some trials as well, but most of the other "experts" were not nearly as accurate. Thus, it seems that our definition of theoretical expertise can be improved.

2) Diagnosis: As mentioned earlier, three levels of analytic resolution were adopted for scoring diagnosis accuracy, Level 1 being the coarsest and Level 3 being the finest. The predicted superiority of the P+F over the P interface for diagnosis was evaluated for each individual subject. Aggregation over subjects was accomplished by counting the number of subjects whose behavior conformed to this prediction. A statistical test was then performed using a Sign Test [20]. Accordingly, there were actually 24 individual experiments testing the theoretical prediction, one for each subject.

The results from this analysis are presented in Table I. For the Experts, the results indicate that the P+F interface was clearly better than the P interface at Levels 2 and 3 ($p < 0.004$ and $p < 0.001$, respectively) and approaching significance at Level 1 ($p < 0.066$). In contrast, the effect of Interface was not significant at any of the levels of analysis for Novices. These results indicate that the P+F interface resulted in superior diagnosis when compared to the P interface for Experts, but that there was no statistically significant difference between interfaces for Novices, as predicted.

The preceding analysis only provides a test of the effect of Interface for each of the two Expertise groups. A more direct assessment of the predicted interaction between Interface and Expertise can be performed using a one-tailed exact probability test. For each of the three levels of analysis, a 2×3 contingency table was derived with one dimension representing Expertise and the other dimension representing the ordinal performance relationship between interfaces (see

TABLE II
RESULTS FROM CONTINGENCY TABLE ANALYSIS OF DIAGNOSIS ACCURACY (ONE-TAILED)

a) LEVEL 1 ($p > 0.10$)

	P+F > P	P > P+F	TIES
EXPERTS	9	2	1
NOVICES	7	2	3

b) LEVEL 2 ($p = 0.0211$)

	P+F > P	P > P+F	TIES
EXPERTS	9	0	3
NOVICES	6	5	1

c) LEVEL 3 ($p = 0.0466$)

	P+F > P	P > P+F	TIES
EXPERTS	11	0	1
NOVICES	7	4	1

Table II). The Expertise dimension had two levels, Novice and Expert, whereas the performance dimension had three levels, P+F better than P, P better than P+F, and P+F equals P. The results are illustrated in Table II . At Level 1, the result failed to reach significance ($p > 0.10$). However, at Level 2, there was a statistically significant interaction between Interface and Expertise ($p = 0.0211$). A similar pattern was obtained for Level 3 ($p = 0.0466$). The results obtained from Levels 2 and 3 therefore suggest that the Experts benefitted more from the P+F interface than did the Novices, as predicted.

3) Correlation between diagnosis and memory: Another way to evaluate the importance of presenting higher order functional variables is to examine the correlation between diagnosis performance and memory for physical and functional variables. If the functional information in the P+F interface is indeed critical to diagnosis as predicted, then one would expect that diagnosis performance would be better correlated with memory for functional variables than with memory for physical variables. An analysis of these data over all subjects [23] reveals that there is a marginally significant correlation between diagnosis and memory for Functional variables ($r(22) = -0.400, p < 0.0528$), whereas the correlation with memory for Physical variables is not significant ($r(22) = -0.258$, n. s.). Thus, the better the diagnosis, the lower the memory error for Functional variables. This result supports the claim that higher order Functional variables are critical to diagnosing the system state, including faults.

C. Discussion

How consistent are these results with the two hypotheses put forth earlier? It was predicted that the P+F interface would result in more accurate diagnosis than the P interface, particularly for Experts. There was strong evidence indicating that the P+F interface was in fact superior to the P interface in terms of diagnosis. The importance of higher levels of abstraction was also supported by the significant correlation between diagnosis performance and memory for Functional variables, a relationship that did not exist with memory for Physical variables.

As for the interaction between Interface and Expertise, several results were consistent with this prediction. A Sign Test analysis of the diagnosis data indicated that the P+F interface was significantly superior to the P for Experts but not for Novices. In addition, a contingency table analysis of those same data revealed a statistically significant interaction between Interface and Expertise at Levels 2 and 3. However, these results must be interpreted with caution, since these effects were not statistically significant in an ANOVA of the same data [23].

There are several reasons to believe that the equivocal results on the interaction between Interface and Expertise may have been due to the definition of theoretical expertise that was adopted in this experiment. First, although there was a statistically significant difference between Experts and Novices on the pre-test (and in memory recall performance as well, [23]), there was some overlap in the two distributions. This is unexpected given the significant difference in educational background between the two groups. Second, no subject received a perfect score on the pre-test. This is also surprising since the Experts had taken several undergraduate and graduate level courses on physical principles similar to, and much more complex than, those governing DURESS' behavior. Third, there were rather noticeable differences between subjects in the Expert group. Some were obviously "more expert" than others (e.g., the subject who derived the system equations from first principles). Although it was not possible to confirm this, it seems possible that the graduate students who performed best were probably involved in research that relied on the same physical principles that are relevant to DURESS. In contrast, the other graduate students in the Expert group had taken the same number and type of courses but probably were not using these physical principles on a daily basis. Thus, while the selection criterion adopted in this experiment for Expertise was valid enough to lead to some statistically significant differences in performance, it is clear that a more refined definition of expertise would be desirable. This issue was addressed in Experiment 2.

Finally, Experiment 1 is also limited in that it did not investigate how subjects used the information in the P+F interface. Now that it has been demonstrated that an interface based on EID can lead to more effective performance for tasks requiring KBB, it would be useful to know more about the cognitive processes that are elicited by such an interface. Knowledge of cognitive processes would also elucidate in more detail why the P+F interface exhibits the performance advantages revealed in Experiment 1. This issue was also investigated in Experiment 2.

III. EXPERIMENT 2

Experiment 2 compared the performance of subjects of different backgrounds on a diagnosis task (see footnote 2) using only the P+F interface for DURESS. The study was designed with two distinct purposes in mind: a) to examine in detail the problem solving strategies employed by subjects in diagnosis scenarios with the P+F interface for DURESS, and b) to explore alternate ways of classifying subjects' backgrounds to

determine which factors are predictive measures of expertise at this task. The rationale behind each of these questions will be addressed next.

A. Rationale

1) Problem solving strategies: EID predicts that an interface that contains information at all levels of the abstraction hierarchy will provide better support for KBB than an interface that contains only a subset of that information. Experiment 1 generated empirical support for this claim. However, in addition to making claims about product, EID can also be used to generate claims about process. That is, the advantages of an EID interface for KBB can be explained in terms of the cognitive strategies that are intended to be induced by such an interface. As a result, it is possible to make specific predictions relating cognitive strategies and the efficacy of task performance.

As Vicente and Rasmussen [26] have pointed out, effective reasoning within an abstraction hierarchy representation should reveal itself in several empirically observable ways. First and foremost, it should be possible to meaningfully map problem solving protocols onto an abstraction hierarchy representation of the domain. Fortunately, Bisantz and Vicente [2] have developed a formal abstraction hierarchy model of DURESS in LISP that can be used in Experiment 2 as a problem space for interpreting subjects' problem solving strategies. An overview of this problem space is provided in Fig. 3. Each "cell" in this problem space consists of a representation of DURESS (at a specific level of abstraction and aggregation). The space consists of three types of relations. Means-end links connect the levels of abstraction shown vertically in Fig. 3. Part-whole links connect levels of aggregation shown horizontally in Fig. 3. Finally, topological links connect objects within any cell at a given level of abstraction and aggregation. For a detailed discussion of the representations of DURESS summarized in Fig. 3, see [2].

Second, one would also expect that subjects' problem solving trajectories would begin at a high level of abstraction and gradually focus in on lower levels, thereby exploiting the goal-relevant constraint provided by the hierarchy. This type of "zooming in" behavior, generally moving from the upper left to the lower right corner of Fig. 3, has indeed been observed in other studies of practical problem solving (see [19] and [26] for reviews). These observations are also consistent with recent research on problem solving expertise which has consistently shown that experts spend a great deal of their time analyzing the functional structure of a problem at a high level of abstraction before narrowing in on more concrete details [8].

In contrast to these effective strategies, ineffective problem solving performance should be characterized by thinking predominantly at lower levels of abstraction, i.e., in the lower right corner of Fig. 3. Problem solving solely within lower levels is difficult because of the high dimensionality of the problem representation at lower levels in the hierarchy and because of the large number of interactions between components. In summary, subjects who do not exploit the hierarchical nature of the abstraction hierarchy representation and the

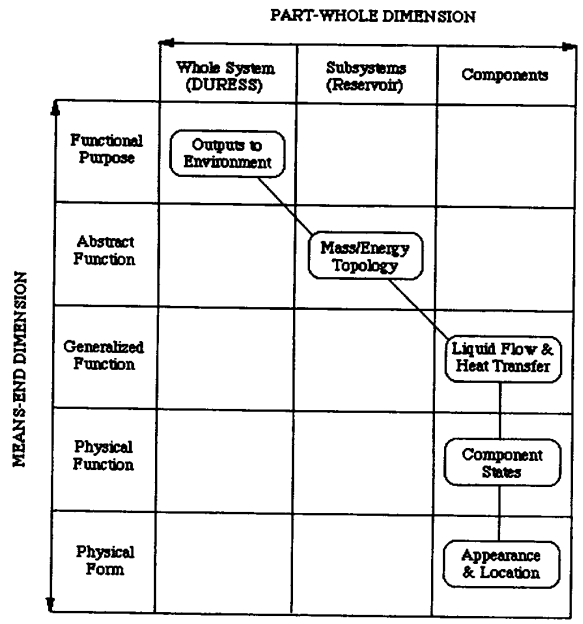


Fig. 3. Two dimensional problem space used to analyze the verbal protocols (adapted from [22]). The space is defined by a means-end dimension and a part-whole dimension. Each cell that is labeled in the space contains a representation of DURESS consisting of various objects connected by topological links. See [2] for more details on the knowledge representation scheme and its formalization.

functional relationships embedded in the P+F interface are expected to not do as well as those who do. These hypotheses were tested in Experiment 2 by adopting process tracing techniques and correlating process measures with measures of performance, or product.

2) Expertise: The expertise aspect of this study is more exploratory in nature. It was prompted by the observation that, in retrospect, many studies of expertise have employed oversimplified classifications of expertise, placing subjects into one of two broad and vaguely defined groups: novice or expert. Often, little consideration is given either to the precise set of skills possessed by each individual or to the level of ability in the specific context of the task presented to subjects (cf. [14]). As a result, such studies do not address the fact that reasoning strategies and performance may differ significantly from subject to subject, as a function of the relationship between each individual's unique background and the specific demands of the task they are asked to perform. The limitations associated with a simplistic novice/expert classification were evident in Experiment 1 where there were marked differences in performance even among "Experts". Experiment 2 sought to discover a more valid and refined definition of expertise. This was accomplished in two ways: first, by classifying each subjects' background along several dimensions which may be relevant to task performance; second, through deliberate sampling for heterogeneity [4]. Thus, we sought out subjects who varied along the various experience dimensions just mentioned. No specific hypotheses were tested. Rather, this aspect of the study was exploratory, attempting to determine

TABLE III
ACADEMIC EXPERTISE RATINGS

SUBJECT	FORMAL EDUCATION	RATING
8	PhD in Electrical Engineering	5.0
2	PhD in Industrial Engineering	5.0
4	BASc Mechanical Engineering	3.5
6	BASc Mechanical Engineering	3.5
1	BASc Industrial Engineering	3.0
9	BASc Civil Engineering	2.5
7	Diploma Mech. Eng. Technology	2.5
10	Stationary Engineering (1st class)	2.5
11	3rd Year Electrical Engineering	2.0
12	1st Year Engineering Science	1.5
3	BSc Physiology	1.0
5	BSc Psychology	0.5

TABLE IV
ANALYTICAL WORK EXPERIENCE RATINGS

SUBJECT	ANALYTICAL WORK EXPERIENCE	RATING
8	Modeling of thermal/fluid systems	10
6	Analysis of flows in reactor piping	6
4	HVAC* analysis	5
7	Superficial analytical experience**	2
10	Superficial analytical experience**	2
2	Marginal analytical experience**	1
9	Marginal analytical experience**	1
All others	None	0

*HVAC = Heating/Ventilating/Air Conditioning

**Superficial analytical experience was the rating given to subjects who had secondary experience in analyzing thermal/fluid systems. For example, subject 7 had some basic experience in piping analysis. Marginal analytical experience was the rating given to subjects who had been exposed to analysis of thermal-hydraulic systems in the context of other work.

which dimensions of experience would be good predictors of task performance, both in terms of product and process.

B. Method

1) Subjects: The subjects for this second experiment were selected with the intention that they should represent a broad variety of expertise, both in terms of type and amount. Five different dimensions were included. The first two dimensions were orthogonal: academic and practical experience. Practical expertise was further subdivided into analytical work experience and operating work experience. Subjects were also classified according to their previous knowledge of the DURESS system, of the P+F interface, and their experience in controlling a real-time interactive version of DURESS (see [15]). The assignment of values to subjects along these five dimensions is admittedly subjective but is justifiable given the exploratory nature of the questions being asked.

Academic expertise was judged on the basis of amount of formal education and relevance to the subject area. On this dimension, the subjects ranged from a psychology student with no background in either thermal-hydraulics or physics up to a professor of biomedical engineering holding a Ph.D. in electrical engineering who had taken several thermal-hydraulics/physics courses. The rating scheme for subjects assigned a mark ranging from 0 (lowest) to 5.0 (highest). Each score was adjusted plus or minus 0.5 according to the details of the subject's education. The ratings assigned to the various subjects for this dimension are shown in Table III.

Analytical work experience was defined as the amount of professional work experience in analyzing thermal and/or fluid systems. Subjects were rated on the basis of the amount and degree of relevance of experience that was specific to the task presented in this experiment. As shown in Table IV, subjects ranged from no analytical work experience (a rating of zero) to experience in modeling both thermal and fluid systems (a rating of 10) on this dimension.

Operating work experience in process control systems was judged solely on the basis of amount. The only subject with any practical operating experience (twenty-seven years as a power plant operator) was given a rating of 1. All others were assigned a rating of zero.

Finally, subjects were assessed on three related measures regarding their experience with the DURESS system. The first was prior knowledge of the structure and functionality of DURESS (subsequently referred to as "DURESS system knowledge"). The second was prior experience with the P+F interface. Thirdly, a score was given based on subjects' experience at controlling DURESS. This measure arose as a result of two subjects (11 and 12) who had participated in a separate experiment (see [15]) which involved the use of an updated, interactive version of DURESS (known as DURESS II). These were the only two subjects with any experience in directly controlling DURESS. They had controlled DURESS II each weekday for approximately one hour for a total of 6 weeks. A special note should be made with respect to subject 11. In the experiment involving DURESS II, he did not use the P+F interface used in this experiment. Instead, he had control experience with the P interface used in Experiment 1, which has some overlap with the P+F inter face. As such, he was given a lower score on experience with the P+F interface than subject 12, who did use the P+F interface in the experiment with DURESS II. Note also that subject number 2 was the designer of the DURESS system and the P+F interface. The ratings for DURESS experience are summarized in Table V.

TABLE V
DURESS EXPERIENCE RATINGS (SYSTEM KNOWLEDGE, P+F EXPERIENCE, CONTROL EXPERIENCE)

SUBJECT	DURESS SYSTEM KNOWLEDGE	RATING
2	Extensive	10.0
11	Moderate	5.0
12	Moderate	5.0
5	Marginal	1.0
1	Marginal	0.5
All others	None	0.0
SUBJECT	**P+F EXPERIENCE**	**RATING**
2	Extensive	10.0
12	Substantial	7.5
11	Moderate	3.0
5	Limited	1.5
1	Marginal	0.5
All others	None	0.0
SUBJECT	**DURESS CONTROL EXPERIENCE**	**RATING**
11	Extensive	1.0
12	Extensive	1.0
All others	None	0.0

A total of twelve subjects were tested. Sex differences were (inadvertently) controlled for, as all the subjects were male. The subjects ranged in age from 19 to 55 years old. The subjects were each paid $14 for their participation in the experiment.

2) Apparatus: DURESS was run on a clone PC-AT with a math coprocessor and a VGA color display. Subjects' verbal protocols were recorded with a Sony model CCD-V101 video camera. The DURESS simulation was the same as that used in Experiment 1.

3) Scenario types: There were five distinct scenario types, each of which was presented twice. These were the same scenario types used in Experiment 1 except that no Random scenarios were used this time. Of the five types, two were classified as normal system behavior: steady-state, and a change in volume taking place in one of the reservoirs. The remaining three scenario types were fault situations: a reservoir leak, a blocked valve, and a change in the temperature of the inlet water.

4) Procedure: When subjects arrived, they were first asked to fill out a demographic questionnaire which was later used as a basis for rating each subject's level of expertise. They were then asked to read a brief introduction to the experiment, followed by a technical description of DURESS. Once they had completed reading the technical description, the subjects were asked to complete a multiple choice test consisting of twenty questions to evaluate their knowledge of thermal-hydraulics. When the test was completed, the subjects were shown a static version of the P+F interface on the PC, as it would appear during the experiment. The experimenter then described the functioning of the interface. Following this, the subjects viewed ten scenarios of DURESS' behavior, each lasting for about 60–70 seconds. The scenarios consisted of two of each of the five types mentioned previously. All subjects viewed the same ten scenarios in the same order.

Subjects were told that their task was to diagnose the behavior of the system. They were told that the simulation represented a "real life" system, and that it was vulnerable to the same types of failures as any such system would be. Subjects were asked to think out loud as they reasoned about what they thought was happening. They were also asked to point to the part of the system they were thinking about whenever possible, in order to clarify what it was that they were referring to as they verbalized their thoughts.

Based in part on the findings of Ericsson and Simon [6] that retrospective analysis sometimes produces data that are inconsistent with the recorded behavior of subjects, it was decided to use concurrent rather than retrospective protocols. However, some retrospective data were accepted at the conclusion of each scenario. If the subject was in the middle of a verbalization, or series of verbalizations as the trial ended (at which point the screen went blank), they were given the option of using a photocopy of the interface to aid them in completing their analysis. All of this was recorded with the video camera.

5) Performance measures: The verbal protocols given by each subject were transcribed and mapped onto the two dimensional problem space illustrated in Fig. 3. A description of the mapping procedure and an example of its application are presented in the Appendix. Both product and process measures were then extracted for analysis.

There were two primary product measures, diagnosis accuracy and understanding. For diagnosis, one point was awarded if the subject directly identified the correct system state (for normal scenarios) or fault (for fault scenarios). A half point was given for indirect or partial identification of the state/fault. For understanding, one point was awarded if the subject demonstrated a good understanding of the system through their description/reasoning. A half point was given if the subject made only a minor slip, or if it could not reasonably be shown that the subject did not understand the system. These measures were then aggregated across the 10 trials, resulting in a total diagnosis score (out of 10) and a total understanding score (out of 10) for each subject. An overall performance measure (out of 20) was obtained for each subject by adding the diagnosis and understanding scores. The overall score was also subdivided to obtain scores for the four normal scenarios (out of 8), and for the six fault scenarios (out of 12).

Four process measures were extracted from the mapped protocols. First, the proportion of verbalizations expressed by subjects in each separate cell of the means-end/part-whole space (see Fig. 3) was determined. Second, the frequency with which subjects started their reasoning trajectories at each level of abstraction was also obtained. Third, the ratio of links (means-end, part-whole, or topological) which the subjects explicitly traversed in navigating through the problem space to the total number of distinct verbalizations expressed was also calculated. This was intended to act as a measure of the level of logical coherence in the subjects' reasoning trajectories (from the frame of reference provided by the abstraction hierarchy). Thus, subjects with a high score exhibited a great deal of connected verbalizations, whereas those with a low score typically voiced thoughts that were unrelated to each other. Fourth, the number of abstraction changes subjects made (both implicitly and explicitly) as a proportion of the total verbalizations expressed per trial was also calculated.

C. Results

Because of the small sample size and the exploratory nature of the study, the analyses consisted of an examination of the Pearson product-moment correlations existing between the various variables of interest. Expertise effects were investigated by analyzing the relationship between the various background measures and the product and process measures of performance. Strategy effects were investigated by analyzing the relationship between process and product measures.

1) Thermal hydraulic pre-test: Scores on the thermal hydraulic pre-test ranged from 9 to a perfect 20, with a mean of 13.5. The scores were significantly correlated with previous knowledge of DURESS ($r(10) = 0.686$, $p < 0.05$) and the P+F interface ($r(10) = 0.608$, $p < 0.05$), both of which are experience measures. Pre-test scores were also correlated with logical coherence ($r(10) = 0.733$, $p < 0.01$) and the proportion of explicit abstraction changes in relation to the total number of verbalizations expressed ($r(10) = 0.805$, $p < 0.01$) both of which are process measures. However, performance on the

pre-test was not significantly correlated with any of the product measures of performance, suggesting that the test did not have a great deal of predictive power.

2) DURESS system knowledge: Previous knowledge of DURESS emerged as the one most strongly correlated with performance. It was significantly correlated with diagnosis ($r(10) = 0.688$, $p < 0.05$), understanding ($r(10) = 0.719$, $p < 0.01$), and overall performance ($r(10) = 0.756$, $p < 0.01$). It was significantly correlated with performance on fault trials ($r(10) = 0.672$, $p < 0.05$), but not with performance on normal trials $r(10) = -0.059$, n.s.). It was also highly correlated with logical coherence ($r(10) = 0.843$, $p < 0.01$), the proportion of explicit abstraction changes ($r(10) = 0.787$, $p < 0.01$), and the proportion of verbalizations expressed at the Abstract Function level ($r(10) = 0.613$, $p < 0.05$). Obviously, domain specific knowledge is an important prerequisite for successful performance.

3) P+F experience: Experience with the P+F interface was also significantly correlated with diagnosis ($r(10) = 0.673$, $p < 0.05$), understanding ($r(10) = 0.688$, $p < 0.05$), and overall performance ($r(10) = 0.732$, $p < 0.01$). It was also significantly correlated with performance on fault trials ($r(10) = 0.601$, $p < 0.05$), but again, not on normal trials ($r(10) = 0.042$, n.s.). It too, was highly correlated with logical coherence ($r(10) = 0.768$, $p < 0.01$) and with the proportion of explicit abstraction changes ($r(10) = 0.636$, $p < 0.05$). These findings mirror those presented above because DURESS system knowledge and P+F experience were correlated in our sample of subjects (see Table V).

4) DURESS control experience: DURESS control experience was not correlated with any performance measures.

5) Education: The level of formal education was significantly correlated with understanding ($r(10) = 0.618$, $p < 0.05$), but not with diagnosis ($r(10) = 0.487$ n.s.). It was still, however, significantly correlated with overall performance ($r(10) = 0.592$, $p < 0.05$). Education was correlated strongly with performance on fault trials ($r(10) = 0.713$, $p < 0.01$), but did not correlate significantly with performance on normal trials ($r(10) = 0.522$ n.s.).

6) Analytical work experience: Analytical work experience was significantly correlated with education ($r(10) = 0.672$, $p < 0.05$), but was not significantly correlated with any performance measures.

7) Operating work experience: Operating work experience was not significantly correlated with any of the performance measures.

8) Problem space analysis: For each valid cell in the part-whole/means-end space (see Fig. 3, or Appendix) the proportion of each subject's verbalizations in that cell was averaged over the ten trials. It should be noted that the values varied fairly widely for each subject depending on the trial type. This is to be expected since indivi dual problem solving trajectories are constrained by the specific context of each scenario. However, the purpose of this analysis was to determine if there any trends that were invariant across the noted differences in scenarios. The results are presented in Table VI.

The proportion of verbalizations in the Generalized-Function/Component cell had a significant negative correlation with understanding ($r(10) = -0.621$, $p < 0.05$) and overall performance ($r(10) = -0.601$, $p < 0.05$). It was also negatively correlated with experience with the P+F interface ($r(10) = -0.616$, $p < 0.05$). The proportion of verbalizations in the Physical-Form/Component cell, on the other hand, was positively correlated with overall performance ($r(10) = 0.575$, $p < 0.05$) and diagnosis ($r(10) = 0.611$, $p < 0.05$). The proportion of verbalizations in this cell was also well correlated with all three of the predictor variables surrounding experience with DURESS: DURESS system knowledge ($r(10) = 0.654$, $p < 0.05$), P+F interface experience ($r(10) = 0.708$, $p < 0.01$), and DURESS control experience ($r(10) = 0.785$, $p < 0.01$).

TABLE VI
PROBLEM SPACE ANALYSIS RESULTS

PROBLEM SPACE CELL	MEAN PROPORTION	RANGE
Functional-Purpose/Whole-System	9%	2%–17%
Abstract-Function/Subsystem	16%	2%–44%
Generalized-Function/Subsystem	16%	8%–29%
Generalized-Function/Component	32%	20%–43%
Physical-Function/Component	22%	11%–34%
Physical-Form/Component	5%	0%–15%

9) Logical coherence: The ratio of links explicitly traversed as part of a sequence of connected verbalizations to the total number of verbalizations expressed was adopted as a measure of the logical coherence of subjects' reasoning trajectories. The mean value for this variable was 0.13 with a range of 0.03 to 0.29. It was significantly correlated with diagnosis ($r(10) = 0.833$, $p < 0.01$), understanding ($r(10) = 0.656$, $p < 0.05$), and overall performance ($r(10) = 0.804$, $p < 0.01$), as well as with performance on fault trials ($r(10) = 0.630$, $p < 0.05$). As noted above, it was also strongly correlated with three of the predictor variables: DURESS system knowledge ($r(10) = 0.843$, $p < 0.01$), experience with the P+F interface ($r(10) = 0.768$, $p < 0.01$), and the pre-test score ($r(10) = 0.773$, $p < 0.01$).

10) Trajectory initialization: The level of abstraction at which subjects began their trajectories was expected to provide information regarding the type of strategy that the subjects were utilizing (e.g., top-down vs. bottom-up). The proportion of trials begun at the Abstract Function level was positively correlated with diagnosis ($r(10) = 0.614$, $p < 0.05$), understanding ($r(10) = 0.697$, $p < 0.01$), and overall performance ($r(10) = 0.704$, $p < 0.01$). This measure was also strongly correlated with DURESS system knowledge ($r(10) = 0.803$, $p < 0.01$) and experience with the P+F interface ($r(10) = 0.754$, $p < 0.01$), as well as with the level of logical coherence exhibited ($r(10) = 0.736$, $p < 0.01$). It was also found that the proportion of trials begun at the Physical Function level, a lower level of abstraction, was negatively correlated with understanding ($r(10) = -0.567$, $p < 0.05$).

D. Discussion

1) Expertise effects: The results concerning the predictor variables will be discussed first. As mentioned previously, the thermal-hydraulic pre-test was not significantly correlated with

any of the primary performance measures. It will therefore not be discussed further.

The predictors most strongly correlated with performance were previous knowledge of the DURESS system and the P+F interface. The five subjects who had prior knowledge of DURESS and the P+F interface sometimes seemed to make their diagnoses of the system in a "recognition-based" manner (cf. [10]), resembling rule-based behavior. They often seemed to make the connection between the symptoms exhibited by the system and their causes without proceeding through any intermediate steps. This can likely be attributed to the fact that they had previous exposure to the typical behavior of the system. On the occasions where these subjects could not recognize the situation, they were nonetheless able to utilize their understanding of the interface along with their mental models of DURESS to aid them in their diagnosis.

Experience at directly controlling DURESS was not significant in enhancing subjects' ability to perform diagnosis tasks. In part, this may be due to the fact that the previous experiment, in which subjects 11 and 12 had been required to control DURESS [15], exposed them to only three fault situations, whereas 60% of the trials in this study were classified as fault trials. It is also possible that this result was due to the fact that the present study required diagnosis of "canned" scenarios, whereas the previous experiment required subjects to compensate on-line for any faults which might occur, so that the system goals were satisfied. Because of this difference, the fault management task used in [15] did not necessarily require a root cause diagnosis of the state of the system (see footnote 2). This may explain why control experience was not a strong predictor of diagnosis performance.

The remaining subjects, being unfamiliar with DURESS, were forced to rely on a strictly knowledge-based approach supported only by their general knowledge, the (unfamiliar) P+F interface, and whatever mental model of DURESS they had been able to form from the pre-experimental material. In light of these factors, the reasons for the correlations which emerged become relatively clear. It is worth noting however, that if the subjects had been allowed a number of practice trials before beginning the actual experiment, the effects of prior knowledge may have been reduced to some degree.

Education was also well correlated with two of the primary performance measures, including overall performance. The level of formal education (with relevance to the experimental task) was likely a factor influencing the ease with which subjects could understand the constraints of the system and integrate information about the system into a coherent mental model.

DURESS knowledge, experience with the P+F interface, and education all showed significant correlations with performance on fault trials but not normal trials. This may have been due to the fact that, when faced with a scenario in which some of the constraints governing the normal operation of the system were violated, those subjects with a fuller understanding of those constraints were better able to interpret and integrate the information made available by the P+F interface. On normal trials it was simply not necessary to know as much about the system in order to understand its behavior.

The fact that seems to emerge from all of these results is that prior knowledge of DURESS and the P+F interface, and not practical or general experience, appears to be the most relevant measure of expertise in the context of this study. This finding is consistent with recent research pointing to the highly context-specific nature of skill (e.g., [12], [14]).

2) Effects of strategy: As already mentioned, previous research has shown that in practical problem solving situations, experienced operators tend to begin their reasoning trajectories at a high level of abstraction and gradually focus in on lower levels [26]. Starting at a higher level allows one to "see the forest through the trees", providing a means for deciding which part of the system warrants a more detailed inspection. This results in a directed, coherent, and economical mode of problem solving, as many irrelevant details are readily eliminated from consideration.

Having established that prior knowledge of DURESS and the P+F interface seem to be the most appropriate measures of expertise in the context of this experiment, one might expect that the degree of knowledge along these dimensions would correlate well with the proportion of trajectories initiated at higher levels of abstraction. This is indeed the case, as shown by the strong correlations between the number of trajectories starting at the Abstract Function level and both DURESS knowledge and experience with the P+F interface.

In further support of the rationale outlined above, it was found that the proportion of trajectories initiated at the Abstract Function level correlated well with the performance measures and with the average level of logical coherence of the verbalization sequences. Conversely, it was also found that initiating trajectories at a lower level, examining individual components that are all highly interconnected, had a negative effect on performance. Specifically, the finding was that there was a significant negative correlation between the proportion of trajectories initiated at the Physical Function level and understanding.

The logical coherence of subjects' reasoning trajectories is a good indication of how directed their strategies are. Indeed, it seems possible that logical coherence is an intermediate factor between expertise and performance. This seems to be borne out by the fact that the logical coherence measure was strongly correlated with knowledge of DURESS and the P+F interface. These findings seem to naturally suggest that logical coherence would be related to performance. Our results support this suggestion, as logical coherence was significantly correlated with the primary performance measures.

It was observed that many subjects' strategies seemed to consist mainly of a serial scan of the settings and states of system components, without any observable focus. These subjects' verbalizations were concentrated mainly at the lower levels of abstraction, leading to a generally inefficient strategy. This was indicated by the significant negative correlations between the proportion of verbalizations in the Generalized-Function/Component cell of the means-end/part-whole space and both understanding and overall performance.

The seemingly contradictory fact that the proportion of verbalizations in the Physical-Form/Component cell showed a significant positive correlation with overall performance and

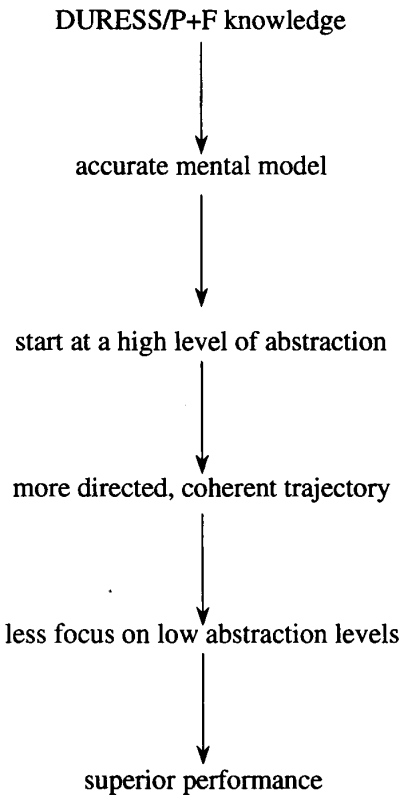


Fig. 4. Integrative summary of the main findings from Experiment 2. Since the empirical evidence is based on correlations alone, the directions of the arrows are merely hypotheses regarding causation. Also, the second factor, mental model, was not directly measured in this experiment, and thus its influence is also conjectural.

diagnosis is explained by the fact that two of the three types of fault trials required at least one judgment in this cell for proper diagnosis.

A summary of the results from Experiment 2 is illustrated in Fig. 4. In summary, effective performance was characterized by initiating search at a high level of abstraction, spending time overviewing the problem at a high level of abstraction, and searching through the means-end, part-whole, and topological links defined by the problem space in Fig. 3 in a coherent and connected manner. These findings provide clear support for the hypotheses generated by the rationale behind the psychological validity of the abstraction hierarchy.

IV. General Discussion

The two experiments described here represent the first empirical tests of the EID framework. Experiment 1 showed that an interface based on the principles of EID can provide more support for KBB than a more traditional interface based on physical information alone. There are two reasons why this result is the most important obtained in Experiment 1. First, it provides evidence that an interface based on the principles of EID can indeed result in improved performance compared to a more traditional interface. Second, this result is important because it is the first time that an interface based on the abstraction hierarchy has been empirically demonstrated to result in superior performance as compared to another type of interface [24].

Experiment 2 tested the claim that an EID interface supports KBB by allowing subjects to initiate their search at a high level of abstraction and then gradually focus on the relevant system components at lower levels of abstraction. The results indicate that effective diagnosis performance is indeed associated with this "zooming in" pattern, as predicted. This result is very significant since it is, as far as we know, the first time that the psychological validity of an abstraction hierarchy representation has been tested in a quantitative and statistically reliable manner. Previous studies (see [19], and [26] for reviews) had shown that problem solving protocols could be mapped onto an abstraction hierarchy representation, but not that effective problem solving performance is associated with the "zooming in" strategy described earlier. Therefore, taken together, Experiments 1 and 2 provide tests of EID in terms of both product and process with very encouraging results.

It is important to point out that these findings are intended to be specific to demands requiring problem solving, or KBB. The experimental conditions were intentionally set up to evaluate how well the two interfaces support performance in novel situations and to minimize the influence of rote perceptual activity. In both experiments, subjects were faced with unfamiliar events and were not given any external aids except the interface itself.

The two experiments presented here also speak to the issue of expertise. Briefly, Experiment 1 uncovered some evidence that theoretical expertise is required to take full advantage of an EID interface. Experiment 2 tried to improve upon the definition of expertise adopted in Experiment 1. The results indicate that system-specific knowledge, and not generic analytical or control experience, is the best predictor of performance in a task requiring KBB. Note, however, that these results may be closely tied to the conditions under which Experiment 2 was conducted. The context-specific nature of expertise suggests that the predictive power of the background variables examined here may vary as a function of the task. Thus, the generalizability of this result needs to be explored in future studies.

V. Conclusions

The findings presented here allow one to derive the following defensible conclusion: An interface based on an abstraction hierarchy representation of the work domain can provide knowledgeable subjects with more support for KBB than an interface based on physical variables alone. This conclusion is expected to generalize to other situations where operators are engaged in KBB, and to domains that are structurally similar to DURESS [see 22]. These results provide some initial support for the EID framework. This study is the first to demonstrate the advantage of an interface based on an abstraction hierarchy representation over a more traditional interface format [24].

Are these results really informative? Playing the role of devil's advocate for a moment, one could argue that all

that has been demonstrated is that an interface with more information is better than an interface with less information, and that this finding is self-evident and need not have required any empirical evaluation. This may seem to be a reasonable criticism to some readers so it is important to clearly point out why this conclusion is incorrect.

The advantage of the P+F interface cannot be solely attributed to more information. While this contention has not been empirically tested here, a simple thought experiment should be sufficient to convince most readers of the validity of the claim. One could easily design an interface that had the information that was in the P interface and some extra information as well. To take a ludicrous example, one could also display the current temperatures in major cities around the world. But of course this added information would be of no use since it is completely unrelated to system goals. Therefore, it is not the case that the experimental results are merely due to the P+F interface containing more information than the P interface. The key is that the added levels of information are *goal-relevant*. EID provides a principled approach to identifying, *a priori*, the goal-relevant information that needs to be included in the interface.

A. Limitations

While this initial empirical evaluation of the EID framework has provided some encouraging results, it is evident that many questions remain unanswered. The most obvious is whether the differences between interfaces observed in this study will still hold with subjects who, like most operators, are expert at control but not theory. Also, we still do not know to what extent the benefits of the P+F interface are due to the added levels of information (content) or to the way in which that information is presented (form), except that both are important (see footnote 3).

The conclusions from the present study are also limited in terms of defensible generalizability. It remains to be seen how well these findings generalize to other settings, both within and outside of process control, and whether interfaces as effective as the P+F interface for DURESS could be developed by designers who are not intimately familiar with the theoretical principles of EID. There are also various design issues associated with visual form (see [1]) that are not addressed in detail by EID, and that therefore have not been investigated in this research. All of these factors can potentially affect the generalizability of the results presented here. Only when the effects of these manipulations have been studied will it be possible to defensibly argue that the principles of EID have been evaluated in a comprehensive and sound manner. The experiments presented here, like most others of this sort, fall well short of living up to such stringent criteria. These broad issues can only be settled by a long-term research program.

The final class of limitations of this research arise from the fact that there are a variety of important interface design issues which have simply not been addressed at all. The primary reason for this is that all of the information was displayed on a single display page. In large-scale systems, many display pages will be required to disp lay all of the data and this introduces new design issues (e.g., context sensitivity, dialogue structure, and visual momentum). The important point to note, therefore, is that adherence to the principles of EID, as they are currently formulated [26], is necessary but by no means sufficient to guarantee effective performance in larger scale systems. These other design issues also need to be considered.

B. Contributions

While limited by these factors, this research has made several contributions to the study of interface design for complex human-machine systems. More specifically, the research program centered around EID addresses several of the research issues that were identified by a U. S. Nuclear Regulatory Commission workshop on human-machine interfaces [5]. For instance, the workshop participants recommended that interfaces should structure information at various levels of abstraction. Clearly, the idea of displaying multilevel information is a significant component of the EID framework, as evidenced by the prominence of the abstraction hierarchy. Moreover, EID is a principled approach to designing such interfaces because it has foundations in basic research in cognitive engineering and psychology, as evidenced by the influence of the SRK taxonomy and the abstraction hierarchy [19] and the concepts of ecological psychology [3], [7], respectively. (See [25] for a discussion of the relationship between ecological psychology and EID). Another outcome of the workshop was a call for evaluating whether an interface that represents process relationships in a manner that is consistent with a subject's mental model actually results in improved performance. According to the workshop report, this proposition had not been formulated as a testable hypothesis [5, p. 11]. The experiments described here are an initial attempt at dealing with this important research question. A further step in this direction would be to directly elicit subjects' mental models. In summary, the work presented here addresses several recognized research needs pertinent to interface design for complex systems.

Perhaps even more important, however, is the fact that the research generated by EID has been recognized by industry as being relevant to real design problems. More specifically, this research has influenced the design of Toshiba's prototype advanced control room for their next generation of NPP's [13]. Toshiba has explicitly adopted the EID framework, as well as some of the specific features of the P+F interface in Fig. 2. Several other companies have also incorporated portions of the P+F interface into smaller-scale demonstration prototypes as examples of principle-driven advanced control room design. This contribution to technology transfer indicates that, contrary to the opinion of some, it is in fact possible to have a close tie between fundamental basic research and significant applied problems.

VI. Appendix

This appendix describes the coding scheme that was used to analyze the verbal protocols. An example of a raw protocol and how that transcription was mapped onto an abstraction hierarchy representation are also presented.

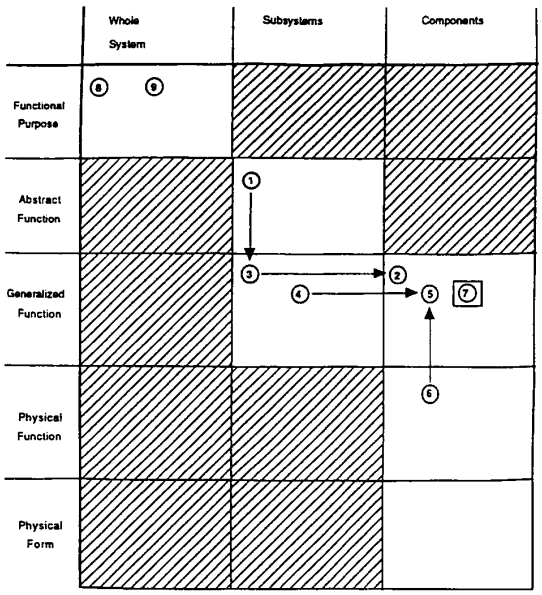


Fig. 5. Example showing how the verbal protocol can be mapped onto the two dimensional problem space to graphically represent a problem solving trajectory. Note the similarity to the trajectories illustrated in [19, p. 119] and [2].

Verbal Protocol Coding Scheme

The mapping of subjects' verbalizations onto the abstraction hierarchy representation of DURESS was done following, as closely as possible, the descriptions of the means-end/part-whole space developed by Bisantz and Vicente [2] (see Fig. 3). Subjects' verbalizations were illustrated as circled numbers in the appropriate cell of the matrix, as shown in Fig. 5. The mapping procedure became quite subjective whenever there was a one-to-one mapping between levels in the space. In these situations, the same verbalization could be mapped onto more than one cell in the space. To address this problem, rules were developed to ensure that the coding was consistent (see below). For the most part, however, the procedure was straightforward and objective, although extremely time consuming.

Judgments about the system goals, or the state of the system as a whole, were mapped in the Functional-Purpose/Whole-System cell. Statements about subsystems referring explicitly to mass or energy were mapped in the Abstract-Function/Subsystem cell. Statements about flows of heat and water were mapped at the Generalized Function level under either the Subsystem category (e.g., a statement about the flow being provided by feedwater stream 1) or under the Component category (e.g., a statement about the flow through a specific valve). Serial scans of all the valves in a feedwater stream were mapped as separate verbalizations under the Component category. Statements concerning the states or settings of system components (e.g., valve or heater settings) were mapped at the Physical Function level under the Component category. In general, static concepts such as component settings, volume levels, and energy levels were mapped at the Physical Function level while dynamic concepts such as flows and rates of change were mapped at the Generalized Function level. Statements about the physical condition of components (e.g., a leak in a reservoir) were mapped in the Physical-Form/Component cell.

Links between verbalizations were mapped when subjects explicitly traversed a means-end, part-whole, or topological link [2]. For example, referring to Fig. 5, verbalizations 1, 2, and 3 are all connected since they are related by links in the problem space. Unrelated verbalizations or related verbalizations in non-adjacent cells of the means-end/part-whole space were not linked. Verbalizations 7, 8, and 9 in Fig. 5 are examples of unconnected verbalizations. When subjects traversed a link using a statement such as "X because Y", the directionality of the link was shown as proceeding from verbalization Y to verbalization X in order to change the sense of the statement to "Y therefore X" (e.g., verbalizations 2 and 3 in Fig. 5). Redundant statements were eliminated unless they were judged to be an important part of a series of verbalizations other than the one in which they were first expressed. Statements given after a scenario had ended were mapped separately (at the bottom of the cells) in order to distinguish them. Statements were no longer mapped after a subject correctly diagnosed the system unless they were relevant to the thought process which led to the diagnosis.

As mentioned, there was an inherent ambiguity concerning where to map thoughts referring to certain system components. The reservoirs in particular could be represented at the Generalized Function level under either the Subsystem or the Component categories, or at the Physical Function level under the Component category. Statements concern ing the reservoirs which included references to the flows into and/or out of the reservoir were mapped in the Generalized-Function/Subsystem cell because they were felt to represent thoughts about the reservoir's interactions with its environment. Statements about the rate of change of volume in the reservoir were mapped in the Generalized-Function/Component cell because they were interpreted as referring to an internal change of state in the reservoir. Finally, statements about the (static) volume level were mapped in the Physical-Function/Component cell. References to the line connecting the input and output flowrates in the mass and energy graphics (see Fig. 2) were mapped in the Abstract-Function/Subsystem cell. Though the judgments being made were essentially ones about the relative input and output flowrates, it was felt that to consider the slope of the connecting line represented a step up in abstraction.

Transcription Example

The following is a typical example of a subject's verbalizations during a scenario. Bracketed numbers represent distinct thoughts which were mapped according to the rules just presented. The completed mapping is illustrated in Fig. 5. The bold number in the transcription signifies a correct diagnosis. Note the similarity between the problem solving trajectory presented in Fig. 5 and the trajectories produced by professional electronic troubleshooting technicians presented in [19, p. 119] and the simulated problem solving trajectories generated by a computer model presented in [2].

Subject X. Trial Y (change in volume - reservoir 2)

"Okay, the first thing I'm looking at is, there's a deviation here (1)." (subject points to reservoir 2, focusing on the 'mass-slope' line (MS-2) connecting input and demand flowrates)

"Because this line is sloped the way it is, the volume level should be going down, at a constant, relatively slow rate (2), since the output flow is slightly larger than the input flow rate (3)."

"Looking over here," (4) (subject scans the whole feedwater system) "... the flows seem to be okay (5), and what I mean by that is that, given the valve settings (6), the flows are what they should be (R5)."

"Okay, the water level is going down in reservoir 2, that is, the volume is decreasing (7)."

"So overall, there is enough water in reservoir 1 to satisfy the demand (8), but pretty soon there won't be in reservoir 2 ... there's going to be a shortage of water (9)."

ACKNOWLEDGMENT

Part of this paper was prepared during the first author's stay as a visiting scientist in the Cognitive Systems Group of Risø National Laboratory, Roskilde, Denmark. In addition to those acknowledged in [23], we would like to thank Mark Chignell, Mogens Christoffersen, Scott Walters, and JoAnne Wang for their help and for providing comments. We are also indebted to the reviewers, whose thorough suggestions helped to improve the paper considerably.

REFERENCES

[1] K. B. Bennett and J. M. Flach, "Graphical displays: Implications for divided attention, focused attention, and problem solving," *Human Factors*, vol. 34, pp. 513–533, 1992.

[2] A. M. Bisantz and K. J. Vicente, "Making the abstraction hierarchy concrete," *International Journal of Human-Computer Studies*, vol. 40, pp. 83–117, 1994.

[3] E. Brunswik, *Perception and the Representative Design of Psychological Experiments*, (2nd ed.). Berkeley, CA: University of California Press, 1956.

[4] T. D. Cook and D. T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton-Mifflin, 1979.

[5] J. DeBor, and R. Swezey, *Man-Machine Interface Issues in Nuclear Power Plants* (NUREG/CR-5348), Washington, DC: USNRC, 1989.

[6] K. A. Ericsson, and H. A. Simon, *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press, 1984.

[7] J. J. Gibson, *The Ecological Approach to Visual Perception*. Boston: Houghton-Mifflin, 1979.

[8] R. Glaser and M. T. H. Chi, "Overview," in M. T. H. Chi, R. Glaser, and M. J. Farr, Eds., *The Nature of Expertise*, pp. xv–xxviii. Hillsdale, NJ: Erlbaum, 1988.

[9] K. R. Hammond, R. M. Hamm, J. Grassia, and T. Pearson, "Direct comparison of intuitive and analytical cognition in expert judgement," *IEEE Trans. Syst. Man Cyber.*, vol. SMC-18, pp. 753–770, 1987.

[10] G. A. Klein, "Recognition-primed decisions," in W. B. Rouse (Ed.), *Advances in Man-Machine Systems Research*, vol. 5, pp. 47–92, Greenwich, CT: JAI Press, 1989.

[11] P. La Cour Christensen, J. Kofoed, and N. Larsen, *PC-DYSIM–A Program Package for Simulation of Continuous Dynamic Processes. User's Manual* (Risø-I-342), Roskilde, Denmark: Risø National Laboratory, Department of Information Technology, 1988.

[12] J. Lave, *Cognition in Practice*. Cambridge, UK: Cambridge University Press, 1988.

[13] K. Monta, Y. Takizawa, Y. Hattori, T. Hayashi, N. Sato, J. Itoh, A. Sakuma, and E. Yoshikawa, "An intelligent man-machine system for BWR nuclear power plants," in *Proceedings of AI 91: Frontiers in Innovative Computing for the Nuclear Industry*, pp. 383–392, LaGrange Park, IL: American Nuclear Society, 1991.

[14] S. E. Olsen and J. Rasmussen, "The reflective expert and the pre-novice: Notes on skill-, rule-, and knowledge-based performance in the setting of instruction and training," in L. Bainbridge and S. A. R. Quintanilla (Eds.), *Developing Skills with Information Technology*, pp. 9–33, Chichester, UK: Wiley, 1989.

[15] W. S. Pawlak, "Inducing effective control strategies through ecological interface design," (Unpublished M.A.Sc. thesis). Toronto, Canada: University of Toronto, 1994.

[16] C. Perrow, *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books, 1984.

[17] J. Rasmussen, "Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models," *IEEE Trans. Syst. Man Cyber.*, vol. SMC-13, pp. 257–266, 1983.

[18] ——, "The role of hierarchical knowledge representation in decision-making and system management," *IEEE Trans. Syst. Man Cyber.*, vol. SMC-15, pp. 234–243, 1985.

[19] ——, *Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering*. New York: North-Holland, 1986.

[20] S. Siegel, *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill, 1956.

[21] K. J. Vicente, "Adapting the memory recall paradigm to evaluate interfaces," *Acta Psychologica*, vol. 69, pp. 249–278, 1988.

[22] ——, "Supporting knowledge-based behavior through ecological interface design," (Unpublished dissertation). Urbana, IL: University of Illinois at Urbana-Champaign, 1991.

[23] ——, "Memory recall in a process control system: A measure of expertise and display effectiveness," *Memory and Cognition*, vol. 20, pp. 356–373, 1992.

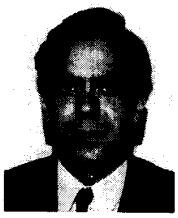
[24] ——, "Multilevel interfaces for power plant control rooms I: An integrative review," *Nuclear Safety*, vol. 33, pp. 381–397, 1992.

[25] K. J. Vicente, and J. Rasmussen, "The ecology of human-machine systems II: Mediating direct perception in complex work domains," *Ecological Psychology*, vol 2, pp. 207–250, 1990.

[26] ——, "Ecological interface design: Theoretical foundations," *IEEE Trans. Syst. Man Cyber.* , vol. SMC-22, pp. 589–606, 1992.

[27] D. D. Woods, "Process tracing methods for the study of cognition outside of the experimental psychology laboratory," in G. A. Klein, J. Orasanu, R. Calderwood, and C. E. Zsambok (Eds.), *Decision Making in Action: Models and Methods*, pp. 228–251. Norwood, NJ: Ablex, 1993

[28] D. D. Woods, J. F. O'Brien, and L. F. Hanes, "Human factors challenges in process control: The case of nuclear power plants," in G. Salvendy (Ed.), *Handbook of Human Factors*, pp. 1724–1770. New York: Wiley, 1987.



Kim J. Vicente (S'88–M'91) received a B.A.Sc. in industrial engineering from the University of Toronto in 1985, an M.S. in industrial engineering and operations research from the Virginia Polytechnic Institute and State University, Blacksburg, VA, in 1987, and a Ph.D. in mechanical engineering from the University of Illinois at Urbana-Champaign in 1991.

During 1987–88, he spent one year as a visiting scientist in the Section for Informatics and Cognitive Science of the Risø National Laboratory in Roskilde, Denmark. During 1991–92, he was on the faculty of the School of Industrial and Systems Engineering at the Georgia Institute of Technology, Atlanta, GA.. Currently, he is an assistant professor in the Department of Industrial Engineering at the University of Toronto, and director of the Cognitive Engineering Laboratory there. He is interested in the design of interfaces for complex human-machine systems, the study of expertise, and more broadly, in the design and analysis of complex work environments. In 1991, he received the award for the best paper published in the Human Factors Society Bulletin.

He is currently an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS and a member of the IEEE SMC Administrative Committee.

Klaus Christoffersen was born in Copenhagen, Denmark, on October 30, 1969. He received the B.A.Sc. degree in industrial engineering from the University of Toronto in June 1993. He is currently pursuing a M.A.Sc. degree, also in industrial engineering at the University of Toronto, specializing in human factors.

Since May 1993, he has been working as a graduate research assistant at the University of Toronto's Cognitive Engineering Laboratory.

His interests include cognitive aspects of human-machine interaction in complex work environments, interface design, expertise, and learning.

Alex Pereklita was born in Toronto, Ontarion, Canada on August 1, 1969. He received the B.A.Sc. degree in industrial engineering from the University of Toronto in June 1993.

He is currently working and studying East Asian culture in Kyoto, Japan.