Memory & Cognition
1992, 20 (4), 356-373

Memory recall in a process control system: A measure of expertise and display effectiveness

KIM J. VICENTE
Georgia Institute of Technology, Atlanta, Georgia

Previous research has shown that memory-recall performance is correlated with domain expertise. In this study, a process control system was selected as a vehicle for conducting research on memory recall. The primary purposes of the present work were to determine if the classic expertise effects originally obtained in chess generalize to this novel domain and to evaluate the validity of memory recall as a measure of display effectiveness. Experts and novices viewed dynamic event sequences showing the behavior of a thermal-hydraulic system with two different displays, one that only contained information about the *physical* components in the system (P) and another that also contained information about higher order *functional* variables (P+F). There were three types of trials: normal, where the system was operating correctly; fault, where a single fault was introduced; and random, where the system's behavior did not obey physical laws. On each trial, subjects were asked to recall the final state of the system and to diagnose the system state. The P+F display resulted in superior diagnosis performance compared with the P display. With regard to memory, there was some evidence of an interaction between trial type and expertise, with experts outperforming novices but primarily on meaningful trials. In addition, memory for the subset of variables most critical to diagnosis was better with the P+F display than with the P display, thereby indicating that memory recall can be a sensitive measure of display effectiveness. The results also clarify a theoretical problem that has existed for some time in the literature, namely, the conditions under which expertise advantages are to be expected in memory-recall tasks. Collectively, these findings point to the potential benefits of adopting an applied context as a test bed for basic research issues.

Research on the use of memory recall as a measure of domain expertise originated at least 45 years ago with the seminal work of de Groot (1946/1965) on problem solving in chess. In that experiment, four chess players of various levels of expertise were asked to reconstruct meaningful board positions after having been exposed to them for only a few seconds. De Groot discovered that the Master- and Grandmaster-level players were able to perform this task with near-perfect accuracy, whereas the performance of the lesser players was not nearly as impressive. In a subsequent experiment, Chase and Simon (1973a, 1973b) found that when the board positions consisted of randomly placed pieces, the recall performance of Masters plummeted to the level of novices,[1] thereby indicating that the Masters' superior performance on meaningful positions is not merely a result of better overall memory. These findings have since been replicated in many other domains, including bridge (Charness, 1979), figure skating (Deakin & Allard, 1991), and schematic diagrams (Egan & Schwartz, 1979). The general conclusion that emerges from this body of research is that memory-recall performance on meaningful stimuli is correlated with domain expertise (see Vicente, 1988, for a review).

The research presented here extends this paradigm to a novel domain, *process control*. There are four reasons why a process control system, such as the thermal-hydraulic simulation used here, can be a productive choice for research on memory recall. First, this class of systems differs in several respects from the domains to which the memory-recall paradigm has been applied in the past. Process control systems are continuous, dynamic, and governed by well-known physical laws. Furthermore, the state of these systems can continue to evolve even in the absence of control inputs from human operators. Will the classic expertise effects first observed in chess generalize to a system with these characteristics? There is no theoretical basis for predicting whether or not the same pattern of results will be obtained under these conditions.

This work was part of my doctoral dissertation, conducted at the Department of Mechanical and Industrial Engineering at the University of Illinois at Urbana-Champaign. Thank you to Jens Rasmussen for his inspiration and for sharing his wealth of knowledge. Jan Kofoed, Roger Marsh, Jim Burke, Bob Chrystler, Paul Milgram, and Ellen Gibbon contributed to the study at different stages. Also, many people were generous enough to critically discuss this research, including Leo Beltracchi, Ann Bisantz, Sally Cohen, Adriaan de Groot, Dennis Egan, Ken Hammond, Tony Hayter, Alex Kirlik, Gary Klein, John Lee, Jim Reason, Barclay Jones, and Dedre Gentner. My advisors, John Flach and Penny Sanderson, made a substantial intellectual contribution to this work, while also providing me with the freedom to explore my ideas. Finally, thanks to Chris Wickens and two anonymous reviewers for providing thoughtful comments, and in particular to Sue Dumais for being an exceptionally thorough source of constructive criticism and guidance. Correspondence should be addressed to K. J. Vicente, now at Department of Industrial Engineering, University of Toronto, Toronto, ON M5S 1A4, Canada (e-mail: benfica@ie.utoronto.ca).

Thus, the present study allows one to assess the generalizability of the results typically associated with the memory-recall paradigm.

Second, assessing memory recall in a process control system also allows one to address an important applied problem, namely, how to evaluate displays in terms of how well they support problem-solving behavior. The traditional way in which such evaluations have been performed is by testing highly experienced operators in plant simulators with abnormal scenarios. The problem with this methodology is that the scenarios consist of incidents that are overlearned to the point that operators can diagnose the incident through familiar cue-action patterns. What is needed instead is a test that will tap operators' general functional understanding as a function of the display, but in a way that is independent of a particular incident. Thus, one would like to measure the degree of fit between the operators' conceptual understanding of the plant and the knowledge representation of the plant that has been embedded in the display, not how well operators are rotely attuned to the perceptual features of the display.

One way to do this would be to first use theoretical experts who have a veridical understanding of the plant. These subjects are not familiar with the perceptual characteristics of the display and therefore cannot diagnose incidents by relying on rote patterns. It would then be possible to see how well various display representations match the experts' mental model. This degree of fit could be evaluated using diagnosis, but it has also been suggested that the memory-recall methodology can be used for the same purpose (Vicente, 1988). Previous research, cited above, has shown that recall performance is correlated with domain understanding. The idea here is to exploit this finding to evaluate the understanding made possible by different displays for the same system. A good display would allow theoretical experts to deploy their expertise and thereby understand the stimulus, whereas a poor display would impede effective comprehension. Thus, one would predict that the better display (as measured by some independent criterion, such as diagnosis) should result in better memory-recall performance than a weaker display.[2] The plausibility of adopting memory recall as a measure of display effectiveness is reinforced by analogous studies in the domain of computer programming. Several researchers have adopted the memory-recall method to measure programmer comprehension as a function of manipulations in the structure and format of the computer code (e.g., Sheppard, Curtis, Milliman, & Love, 1979; Shneiderman, 1977).

A third reason for selecting a process control system as a vehicle for conducting research on memory recall is that such a system allows one to measure expertise according to two complementary criteria. Traditionally, recall performance has been evaluated by comparing subjects' recall with the actual state of the system (e.g., the configuration on the chessboard), thereby measuring the degree of *correspondence* between subjects' responses and the stimuli presented to them. However, as Hammond, Hamm, and Grassia (1986) have pointed out, competence or expertise is actually a joint function of correspondence and *coherence*. This suggests that it would be beneficial to have a measure of the coherence in subjects' recall to complement the traditional correspondence measure. In contrast to the other domains to which the memory-recall paradigm has been applied, in process control there is an objective reference for evaluating coherence, namely, the time-independent constraints governing the system. These constraints describe the redundant relationships that exist between process variables at a single point in time. Thus, a measure of coherence (or internal consistency) can be derived by calculating the degree of consistency between the values subjects entered for different variables and the relationships that usually exist between those variables, as specified by the process constraints.

Together, the coherence and correspondence measures provide a powerful way to analyze subjects' performance. For example, with the coherence measure it is possible to evaluate how well subjects can reconstruct variables that they may not remember. For any given trial, subjects could receive a poor correspondence score, indicating that they did not accurately remember the state of the system, while also receiving a perfect coherence score, indicating that their recall was entirely consistent with the system's constraints. Thus, going from chess to process control provides an opportunity to evaluate recall performance in a different, yet meaningful, way.

A fourth and final justification for conducting memory-recall research within the applied context of process control is that it allows one to directly address a thorny theoretical issue that has existed for some time in the literature, namely, the conditions under which an expertise advantage is to be expected. As the memory-recall paradigm has been applied to a greater number of domains outside of chess, it has become less clear how one should describe the conditions under which there is an expertise advantage and those in which that advantage is not so great or disappears. The words *familiar* and *meaningful* have been used interchangeably by some to refer to the conditions under which an expertise advantage is observed (e.g., Ericsson & Staszewski, 1989), but, as pointed out elsewhere, these concepts are not interchangeable (Vicente, 1988). For instance, a subject could be presented with an unfamiliar stimulus (i.e., one that had never been seen before), but that stimulus could still be meaningful to the subject. There has also been some uncertainty as to what the functional equivalent of a random chessboard is, particularly in the field of medical diagnosis (cf. Coughlin & Patel, 1987). To confuse matters even more, Myles-Worsley, Johnston, and Simons (1988) conducted a study investigating the effects of expertise on recognition memory for X rays and found that memory *increased* as a function of expertise on abnormal X rays and actually *decreased* with expertise for normal X rays. This result is surprising if one assumes that abnormal and normal X rays are comparable to random and normal chess positions, respectively. It is clear from these observations that the boundary conditions under which expertise advantages are expected have yet to be defined in a sound manner.

This significant theoretical issue can be investigated within the context of process control because it is possible to present subjects with three types of stimuli: random, normal, and fault. The interesting question is whether or not there should be a memory-expertise advantage for fault trials. On the one hand, one could argue that such trials are unfamiliar and therefore should be more similar in nature to random trials than to normal trials. This line of thought would suggest that there would not be an expertise advantage for fault trials. Conversely, one could equally claim that such trials, while unfamiliar, are nonetheless meaningful in that they are physically realizable. In this case, one would predict that fault trials would be more similar in nature to normal trials than to random trials. If this were true, then one should observe an expertise advantage for fault trials. There is no criterion that one can apply from the memory-recall literature to determine which of these hypotheses is correct. Thus, any result obtained from such a manipulation will clarify the boundary conditions under which expertise advantages in memory-recall tasks are to be expected.

In summary, conducting memory-recall research within the context of process control allows one to assess the generalizability of previous results in the literature, to address an important applied problem, and potentially to make both a methodological and theoretical contribution to basic research on memory recall. These are the primary goals to which the present experiment is directed. The next section describes the research vehicle and the displays that were adopted for the experiment.

DURESS

The present research was conducted within the context of DURESS (DUal REservoir System Simulation), a thermal-hydraulic process simulation (cf. Vicente, 1991). The physical structure of DURESS is illustrated in Figure 1. The system consists of two redundant feedwater streams, each consisting of a pump and three valves, that 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 temperature (T1, T2) and volume (V1, V2) of the two reservoirs are also displayed. The representation in Figure 1 is the physical (P) display that

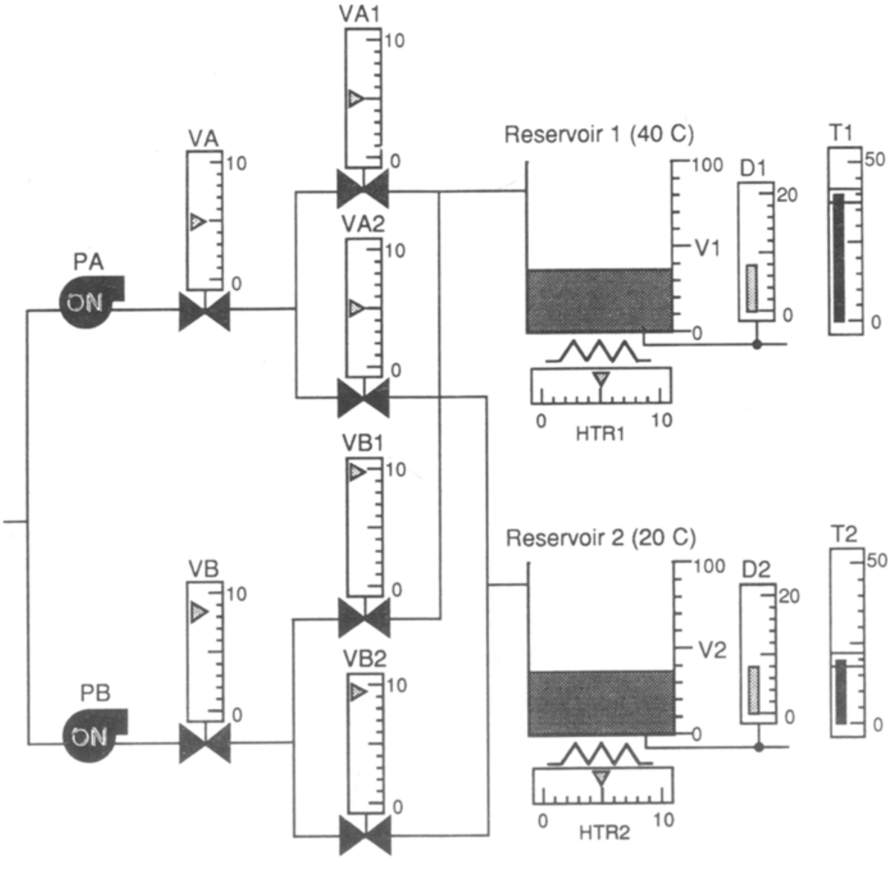


Figure 1. Physical display for DURESS.

was included in the experiment to be described below. It contains information about the goal variables (volume, demand, and temperature) and the state of all of the components.

To evaluate whether the memory-recall methodology can be adopted as a measure of display effectiveness, a second display for DURESS was constructed. This display, illustrated in Figure 2, contains all of the information in the P display as well as some additional information. For reasons to be discussed below, this second display will be referred to as the physical/functional (P+F) display. A brief description of the display follows (see Vicente & Rasmussen, 1990, for a more detailed account).

Beginning on the left side of Figure 2, the valve settings (e.g., VB) and heater settings (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. The relative spatial layout of the components and the connections between them are also represented. The demand (D1, D2) and temperature (T1, T2) setpoints are represented on the right half of the display. 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 flow rates in each feedwater stream (e.g., FVA, FPA, FA1, FA2) and the heating rates (e.g., HTR1) are displayed as bar scales.

The group of graphic representations on the right side of Figure 2 represent DURESS in terms of first principles (i.e., mass and energy conservation laws). The rectangular graphic on the left represents the mass balance (i.e., input flow rate, inventory, and output flow rate) for the reservoir, and the one on the right represents the energy balance. Both operate in a similar manner. Referring to Reservoir 1, the various 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 partialed out according to the two contributors. Thus, the energy added by the feedwater is shown as the lightly shaded bar, and 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 is the case with the mass balance for Reservoir 1 in Figure 2), then it is easy to visualize the consequence, namely, that the volume should

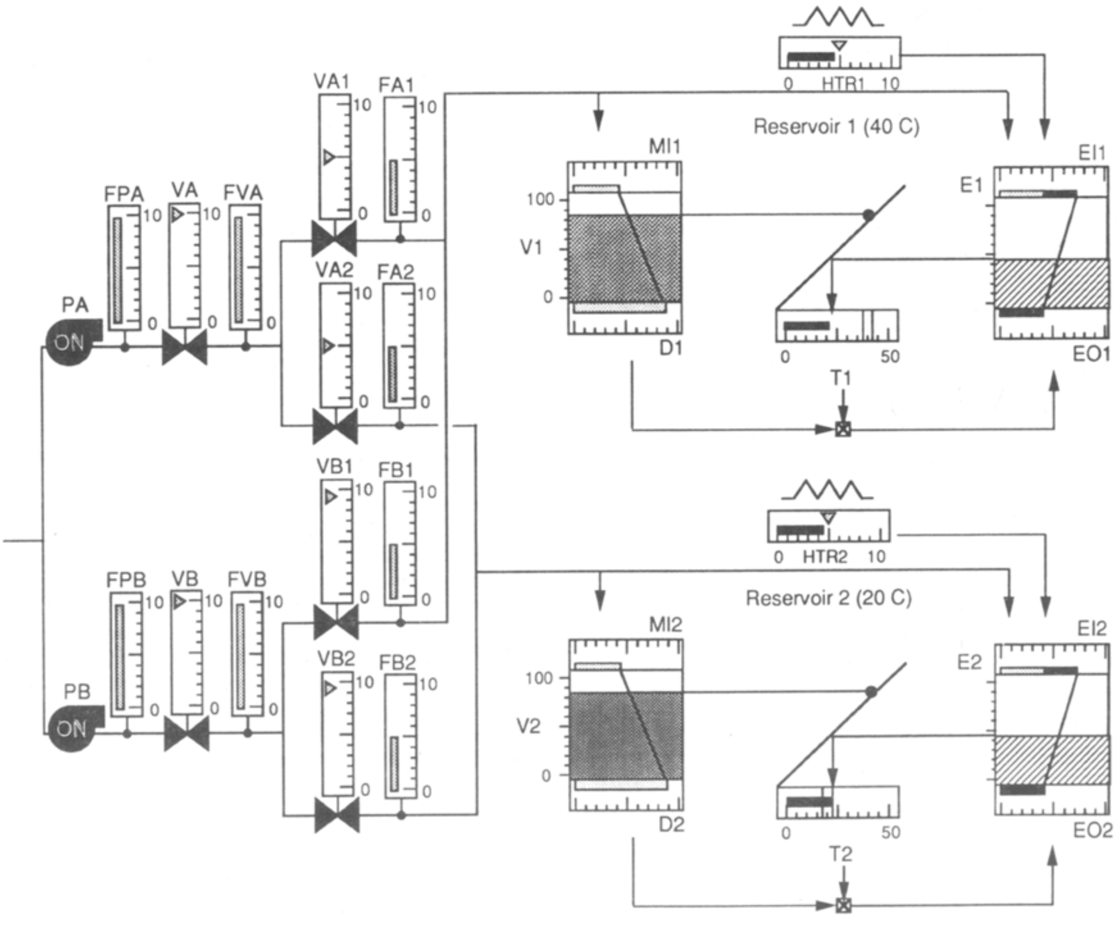


Figure 2. Physical/functional display for DURESS.

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 relationship between volume, energy, and temperature. The horizontal line with a ball on the end that emanates from the current volume level is rigid and of fixed length. Changes in the height of this line always accompany 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 end of the horizontal line. Thus, a change in the vertical position of the horizontal line serves to change the slope of the line in the center display. For example, if volume increases, the horizontal line goes up, causing the diagonal to rotate counterclockwise, 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).

The two displays just described were designed to differentially support problem-solving activities (i.e., knowledge-based behavior; cf. Rasmussen, 1983). More specifically, the P+F display was based on the principles of ecological interface design (EID), a novel theoretical framework for interface design for complex human-machine systems (Vicente, 1991; Vicente & Rasmussen, in press). According to the principles of EID, a display must represent the process at various levels of abstraction if it is to properly support problem solving (see Vicente, 1991, for more detail). The P+F display satisfies this criterion since it was based on a hierarchical representation of the physical and functional relations describing DURESS (cf. Vicente & Rasmussen, 1990), whereas the P display only contains physical information. Therefore, according to the analytical criteria set out by EID, the P+F display provides better support for problem solving than does the P display. This theoretical claim is consistent with previous research that demonstrates that providing people with a device model representing the functional relations between physical components can lead to improved performance (Kieras and Bovair, 1984; Schumacher and Gentner, 1988). The strategy adopted with the P+F display is similar to that adopted in these studies except that the device model was built into the display rather than being communicated to subjects through instruction.

It is important to note, however, that the validity of the EID principles upon which the P+F display has been based has yet to be empirically evaluated. Consequently, a second dependent variable, in addition to memory recall, is required to determine if the P+F display actually results in better performance than does the P display, as predicted. Because the primary problem in process control is identifying the state of the system under abnormal (i.e., fault) conditions, diagnosis accuracy would be a suitable measure of display effectiveness. This would provide an independent empirical criterion that could be used to evaluate the sensitivity of the memory measure to a manipulation of display.

HYPOTHESES

An experiment was conducted to investigate the issues discussed above. The first question to be addressed is whether the classic expertise effects obtained in chess and other domains will generalize to process control. If they do, then one would expect an interaction between trial type (random or meaningful) and expertise such that memory performance will be better for experts than for novices, but only for meaningful trials. This hypothesis can also be evaluated with a measure of coherence memory, in addition to the more traditional measure of correspondence memory. Another important question is whether the P+F display is in fact superior to the P display, as the EID framework claims. If so, then one would predict that the P+F display will result in better diagnosis performance than will the P display. The third issue to be addressed is whether the memory-recall measure is sensitive to differences in display effectiveness. Assuming that the P+F display does result in better diagnosis, one would predict that the P+F display will result in superior memory performance compared with the P display. The final issue to be addressed is the set of conditions under which an expertise advantage is to be expected. As mentioned above, the critical question is whether there is an expertise advantage on fault trials. There is no criterion that one can apply from the memory-recall literature to answer this question. As a result, any result obtained from this manipulation will clarify the boundary conditions under which expertise advantages in memory-recall tasks are to be expected.

METHOD

Experimental Design

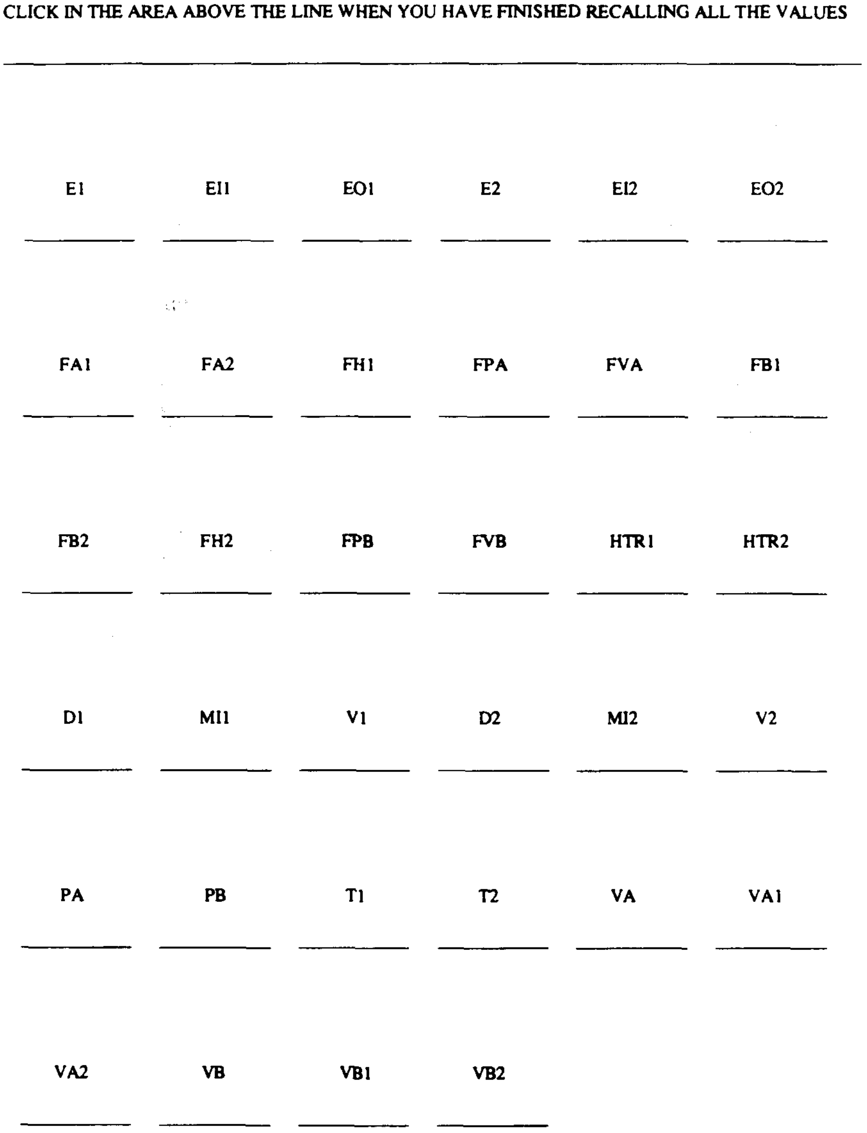
A 2×2×2×2×2 repeated-measures factorial design with two within-subject factors (display and trial type) and three between-subject factors (expertise, order, and sequence) was adopted for this experiment. There were two levels of expertise: experts and novices. The order factor, which refers to the order in which the subjects were exposed to the two displays, had two levels: P display first (P1) and P display second (P2). The sequence factor refers to the order in which the four blocks of 10 scenarios were presented to the subjects. There were two levels: forward (Block 1, Block 2, Block 3, and Block 4) and backward (Block 4, Block 3, Block 2, and Block 1). These three factors resulted in eight different subject groups. The two within-subject factors were factorially crossed and nested within each of the subject groups. As mentioned above, display had two levels: P and P+F. The P display contained 16 variables corresponding to the states of the physical components (see Figure 1), whereas the P+F display contained 34 variables representing both physical and functional variables (see Figure 2). There were also two trial types: semantic, in which the variables were driven by a simulation of DURESS, and random, in which the variables were driven pseudorandomly (see description of trial types below). The subjects used each display for two successive sessions. Each session consisted of 10 trials, with five replications of each trial type.

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 sec. These brief exposure times made the task a challenging one so as to maximize the chances of detecting display and expertise effects and to make sure that the subjects were forced into a problem-solving mode (i.e., analytical reasoning based on knowledge of system structure and functioning, as opposed to pattern recognition based on familiar perceptual cues). This is consistent with the primary concern of the experiment, which was to study the relative level of performance between conditions, not the absolute level of performance of any particular group. The subjects viewed each scenario and were to try to understand and remember as much as they could of what took place. In the instructions, the subjects were told to concentrate on understanding rather than rote recall. While the event was being presented, no response was required. Once the event had finished, the screen went blank, and then the recall screen shown in Figure 3 was automatically displayed.

The recall screen contained the 34 process variables represented in the P+F display (see Table 1). The subjects were required to estimate the value of each of these variables at the end of the preceding scenario. The procedure was the same regardless of which display the subject was using. This means that, for the P display, the subjects were asked to estimate the values of variables that were not displayed. The reason for asking the 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 display.

The format for the recall, shown in Figure 3, was the same for every session. There were 34 bars on the display corresponding to the 34 variables to be recalled. The variables were laid out in a left-to-right and top-to-bottom fashion and grouped according to the variable classes listed in Table 1. Thus, the organization of the variables did not conform to the topographic layout of either dis-



CLICK IN THE AREA ABOVE THE LINE WHEN YOU HAVE FINISHED RECALLING ALL THE VALUES

E1	EI1	EO1	E2	EI2	EO2
FA1	FA2	FH1	FPA	FVA	FB1
FB2	FH2	FPB	FVB	HTR1	HTR2
D1	MI1	V1	D2	MI2	V2
PA	PB	T1	T2	VA	VA1
VA2	VB	VB1	VB2		

Figure 3. Format used for recall of variables.

Table 1
Process Variables in DURESS and Corresponding Labels

	Reservoir 1	Reservoir 2
Temperature variables	T1	T2
	Mass Variables	
Demand (output) flow rate	D1	D2
Mass input flow rate	MI1	MI2
Volume	V1	V2
	Energy Variables	
Total energy stored	E1	E2
Energy input flow rate	EI1	EI2
Energy output flow rate	EO1	EO2
	Heat Transfer Rates	
Flow from HTR1	FH1	
Flow from HTR2		FH2
	Flow Rates	
Flow from VA1	FA1	
Flow from VB1		FB1
Flow from VA2	FA2	
Flow from VB2		FB2
Flow from PA	FPA	
Flow from PB		FPB
Flow from VA	FVA	
Flow from VB		FVB
Heater settings	HTR1	HTR2
	Feedwater Stream A	Feedwater Stream B
Pump settings	PA	PB
	Valve Settings	
Initial valve	VA	VB
Valve 1	VA1	VB1
Valve 2	VA2	VB2

play. The intent was to create a structured (rather than arbitrary) response format, but not to develop an organization that would favor either display.

When the recall screen first appeared, all of the bars were drawn in red. The label for each variable was presented above the respective bar. The left endpoint of the bars represents the zero point, and the right endpoint represents the maximum value for that variable. The subjects were provided with the maximum scale values for each variable in the instructions, which were also made available during recall. The recall estimates were entered using a mouse. The subjects clicked on a point on the bar to indicate where they thought the variable was at the end of the trial. For example, if they thought that the value of the temperature of Reservoir 1 (T1) was 25°, they would click halfway on the bar under the label T1 (the maximum scale value for temperatures is 50°). When a bar was clicked, the color of that bar changed from red to white and blue to indicate the value that had been input. Continuing with the same example, the scale for T1 would be displayed in white from the left endpoint (minimum scale value) to the point that had been clicked on (halfway in this example) and then in blue from that point to the right endpoint (maximum scale value). Thus, the length of the white part of the scale would indicate the value that was entered. Recall estimates could be entered in any order, and the subjects could go back and change the estimate of any variable as many times as they wished. The subjects could consult a list of the 34 variables, their respective labels (see Table 1), their minimum and maximum scale values, and a schematic diagram of DURESS at any point during this procedure. There was no time limit for the recall, but a value had to be entered for all variables.

Once the recall procedure was terminated, the subjects answered a set of structured questions evaluating their diagnosis of the previous event. 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, go to 4.)
3. Describe the fault in as much detail as you can. Where was the fault? What did it consist of? (Stop here.)
4. Given that there was no fault, provide a detailed functional description of what you observed.

In the instructions, the subjects were only told that there would be three types of trials: scenarios that exhibit a normal pattern of behavior following physical principles, scenarios that have a single fault or disturbance, and scenarios 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. Note that the subjects were not told what types of faults could appear nor what the ratios of fault to normal to random trials were.

To summarize, the subjects were tested in two ways, by asking them to remember the values of particular variables and to diagnose the functional state of the system. Knowledge of results was not provided at any point during the experiment.[3]

Subjects

The expert subjects were graduate students in either mechanical or nuclear engineering. In contrast with most previous recall studies, these subjects were theoretical experts, but not experts at controlling the system (see Note 2). The novices were graduate students who had never been enrolled in a science or engineering major. The expert subjects had taken an average of 5.73 graduate or undergraduate physics courses (range of 3-16) and 5.09 graduate or undergraduate thermodynamic or thermal-hydraulic courses (range of 3-9). In contrast, novices averaged 0.75 physics courses (range of 0-2). No novice had ever taken a thermodynamic or thermal-hydraulic course. The two subject groups were from the same university and were roughly equal in terms of age and academic level. There were 12 subjects in each group, 2 females and 10 males. The subjects were paid a total of $24 for participating in the experiment.

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 displays 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 (cf. La Cour Christensen, Kofoed, & Larsen, 1988).

Trial Types

Each trial consisted of a dynamic event sequence illustrating DURESS' behavior. However, the settings of the components did not change during the trial. Thus, the trajectory followed by the process variables (see Table 1) 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 sec 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.

As mentioned above, there were two general classes of trial types: semantic and random. There were five different types of trials in the semantic condition, each occurring once within a session:

1. *Steady state*. For this trial type, there were no changes in any of the system variables.

2. *Change in reservoir volume*. In this condition, there was a change in the volume of one of the reservoirs (either an increase or a decrease) caused by a difference between the mass input flow rate and the current demand. The change in volume was not a result of a fault, but was rather the result of a mismatch between the current water supply rate and the current output demand.

3. *Reservoir leak*. With this trial type, a leak was introduced in one of the two reservoirs. This meant that the volume gradient was less than it should have been, given the current input and output flow rates for the reservoir in question.

4. *Blocked valve*. This fault class 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) that the failed valve is connected to.

5. *Change in inlet water temperature*. This fault trial type consisted of a disturbance in the temperature of the inlet water (either an increase or a decrease). This caused a corresponding change in the temperature of both reservoirs.

The random trials were constructed by randomly sampling from the semantic trials. For every trial, the time history of each variable is defined by a trajectory, for a total of 34 trajectories per trial. The random scenarios contain the same set of trajectories as do the semantic scenarios except that trajectories that once belonged together are now placed in different scenarios. In this way, the average temporal distribution properties of the variables displayed in the random scenarios are identical to those of the semantic scenarios. The primary difference between the two trial types is that the random scenarios do not obey the laws of physics, whereas the semantic scenarios do. In spite of this, there is still some constraint between variables in the random condition because of the sampling procedure adopted (e.g., the flows through the pumps and the first valves in each feedwater stream [e.g., FPA and FVA] tended to be greater than the flows through the latter two valves [e.g., FA1 and FA2]). Thus, the random scenarios are actually pseudorandom, not fully random.

Procedure

Sessions. The entire experiment consisted of one introductory session followed by four data-collection sessions. Each session was conducted on a different day. The subjects performed the task with one display for two sessions, and then with the other display for another two sessions. The order in which the two displays were presented was counterbalanced. The entire experiment lasted from 4 to 6 h.

Introductory session. During the first session, the subjects were presented with a general introduction to the experiment consisting of an introductory statement outlining the purpose of the experiment, reading and signing an informed consent form, and filling out a demographic questionnaire. The subjects then read a brief description of the physical properties of the DURESS simulation. Afterwards, they wrote a pretest of thermal-hydraulic knowledge consisting of 20 multiple-choice questions couched within the context of DURESS. This pretest was intended to evaluate the subjects' theoretical thermal-hydraulic knowledge. There were two general types of questions: quantitative (e.g., given the pump and valve settings for one feedwater stream, derive the flow rates) and qualitative (e.g., given certain assumptions, what effect will increasing the heater have?). A maximum of 30 min was allotted for taking the test. Finally, the subjects were given descriptions of the 34 variables in DURESS, the labels that were used to identify each variable throughout the experiment (see Table 1), and the procedure for recalling the state of these variables. The latter consisted of five blocks of simulated recall, each consisting of 34 trials. This practice was provided to allow the subjects to become familiar with the labels and locations of the variables they were required to recall. For each trial, the experimenter called out a variable (e.g., the setting of the heater in Reservoir 1), and the subjects were required to identify the corresponding variable label (e.g., HTR1) and then find that variable on the recall screen and click on it. Each block of trials consisted of going through the 34 variables in a random order.

Data-collection sessions. At the beginning of the first session with each display, the subjects were introduced to the display they would be using for the next two sessions. For the first data-collection session only, the experimental task was explained to the subjects. This explanation consisted of a description of what would happen on each trial, a brief description of the three different types of trials (random, fault, normal), a review of the recall procedure, and a description of the diagnosis questions that would be posed after each trial.

Each session consisted of 10 trials, 5 semantic and 5 random. The order of the trials was randomized within a session with the constraint that no more than 3 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 (see description of experimental design above).

Performance measures. The primary performance measures can be divided into three classes: recall correspondence, recall coherence, and diagnosis accuracy.

Recall correspondence provides a measure of the difference between the subjects' estimates and the actual state of the variables at the end of each trial. An error score from 0 to 1 was calculated by subtracting the subjects' recall from the actual state of the variable, normalizing this difference with respect to the maximum scale value for each variable, and then taking the absolute value. Thus, 0 indicates perfect recall and 1 represents the worst possible recall.

Another way to evaluate the subjects' recall is to determine how internally consistent it is. There are nine pairs of time-independent constraints between variables when DURESS is functioning normally (see Table 2). Thus, it is possible to see how consistent the subjects' responses are with this set of constraints. The coherence measures were calculated solely from the subjects' recall estimates, *not* from the actual values. For each equation of constraint, the left side is subtracted from the right side, the resulting difference is normalized by dividing by the maximum possible difference score, and the absolute value is taken, resulting in a normalized error score. An error score is obtained for each pair of isomorphic constraints (e.g., the conservation of mass constraint for Reservoir 1 and for Reservoir 2), and the two error scores are averaged. The result is a set of nine measures that indicate how consistent the subjects' recall was with each of the nine equations of constraint. A score of 0 indicates perfect consistency, and a score of 1 indicates the worst possible consistency.

A few words about the method of evaluating memory performance are in order. First, recall was scored as a continuous variable because the process variables themselves are continuous and can therefore take on an infinite number of values. Second, the relationship between coherence and correspondence should also be elaborated. On random trials, these measures are not necessarily related. However, on semantic trials, coherence and correspondence memory are related in an asymmetrical fashion; perfect correspondence guarantees perfect coherence, but perfect coherence says nothing about correspondence.

The final performance measure, diagnosis accuracy, was evaluated in several ways. Three levels of analysis were adopted that evaluated how well the subjects could discriminate: (1) random from semantic trials, (2) random from normal from fault trials, and (3) the exact trial type (see description of trial types above). These levels of discrimination are shown in Figure 4.

RESULTS

The analyses presented here are based on data collected from the second session with each display only. The first session with each display served as practice to allow the subjects to become accustomed to finding and reading the variables in each display. Unless mentioned otherwise, the analyses of variance (ANOVAs) followed the five-factor experimental design described above, and a level of $\alpha = .05$ was adopted to test for significance of effects. Planned comparisons were conducted with simple-effect *F* tests using experimentwise error terms, and post hoc pairwise comparisons were evaluated with a Neuman-Keuls test. This section is divided into five subsections according to the following analyses: pretest scores, diagnosis accuracy, correlation between diagnosis and memory, correspondence memory accuracy, and coherence memory accuracy. For additional analyses of the data generated from this study, see Vicente (1991).

Thermal-Hydraulic Pretest

The results from the pretest of thermal-hydraulic knowledge will be described first. As mentioned, the test consisted of 20 multiple-choice questions. The experts' test scores ranged from 10 to 18, with a mean of 14.67.

Table 2
Algebraic Constraints Governing DURESS Variables Under Normal Operations

Algebraic Equations	Constants
1. $E1(t) = T1(t)\ V1(t)\ c_p\ rho$ • relationship between energy, volume, and temperature	rho: density of water c_p: specific heat capacity
2. $EI1(t) = FH1(t) + c_p\ T_I MI1(t)$ • conservation of energy from heater and inflow	T_I: inlet water temperature
3. $MI1(t) = FA1(t) + FB1(t)$ • conservation of mass from two feedwater streams	
4. $EO1(t) = D1(t)\ c_p\ T1(t)$ • energy leaving reservoir	
5. $FA(t) = FA1(t) + FA2(t)$ • conservation of mass in feedwater stream	
6. $FH1(t) = HTR1(t)$ • conservation of energy from heater	
7. $FA1(t) = \frac{FA(t)\ VA1(t)}{VA1(t) + VA2(t)}$ • flow split relation	
8. If pump is *off*, then $FPA(t) = 0$, otherwise: If $[VA1(t) + VA2(t)] > VA(t)$, then $FPA(t) = VA(t)$, Else $FPA(t) = VA1(t) + VA2(t)$. • flow through pump	
9. $FA(t) = FPA(t)$ • conservation of mass in pipe	

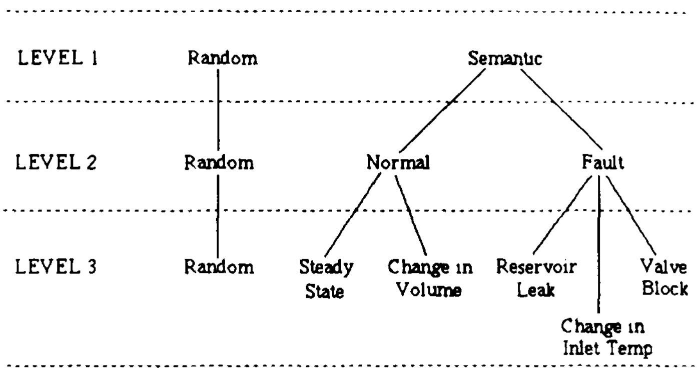


Figure 4. Three levels of analysis adopted for scoring diagnosis accuracy.

The novices' test scores ranged from 4 to 12, with a mean of 9.58. A Mann-Whitney *U* test (Siegel, 1956) indicated that the difference in means between the two groups is statistically significant ($U = 10$, $p < .002$).

Diagnosis Accuracy

The next dependent variable to be analyzed is diagnosis accuracy. As illustrated in Figure 4, three levels of analysis were adopted. Level 1 measured how well the subjects could discriminate random from semantic trials. Level 2 measured how well the subjects could discriminate random from normal from fault trials. Level 3 measured how well the subjects could discriminate between random trials and the five specific semantic trials that were adopted for the experiment.

Two different types of statistical tests were used to analyze these data. First, standard parametric ANOVAs were conducted. Second, following the methodological example set by Hammond, Hamm, Grassia, and Pearson (1987), each individual subject's data were also analyzed using nonparametric tests. The results from the parametric tests are described first.

Parametric tests. A four-way ANOVA, with display, expertise, sequence, and order as factors, was conducted. The dependent variable was the number of diagnosis questions (out of 10) that were correctly answered for each session. Since the data are based on frequency counts, the scores were first converted to percent correct and then a square root transformation was performed (Myers, 1972). Thus, the resulting measure of diagnosis ranged from 0 to 10, with 10 being a perfect score. A separate ANOVA was performed for each level of analysis.

The results of the three ANOVAs are illustrated in Figure 5. The effect of display was highly significant at each level of analysis, with the P+F display consistently outperforming the P display [$F(1,16) = 19.47, p < .001$, for Level 1; $F(1,16) = 18.95$, $p < .001$, for Level 2; $F(1,16) = 24.04$, $p < .001$, for Level 3]. Thus, there is very strong evidence indicating that the P+F display resulted in better diagnosis than did the P display.

Although the data in Figure 5 seem to suggest an expertise × display interaction, this result did not reach significance at any of the three levels [$F(1,16) < 1$ for Level 1; $F(1,16) = 1.12$, $p < .3048$, for Level 2; $F(1,16) < 1$ for Level 3]. These results suggest that experts and novices benefited no differently from the P+F display.

The data in Figure 5 also seem to indicate that experts outperformed novices, particularly at finer levels of analysis, but this result also did not achieve statistical significance. The novices did not perform significantly differently from the experts. This is an unexpected finding, but perhaps it can be attributed to the method of scoring the analysis. To investigate this possibility, the diagnosis data were analyzed according to nonparametric tests.

Nonparametric tests. Following the example of Hammond et al. (1987), the predicted superiority of the P+F

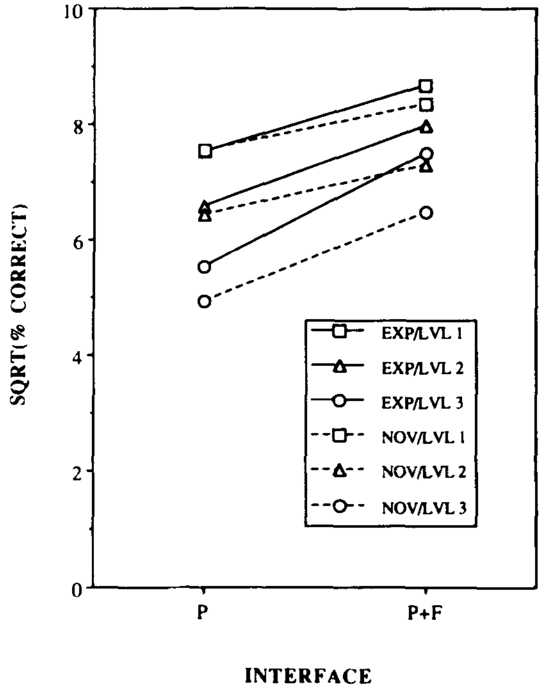


Figure 5. Primary results from analysis of variance of diagnosis accuracy.

Table 3
Results from Sign-Test Analysis of Diagnosis Accuracy for Experts and Novices

Level	P + F > P	P + F = P	p
	Experts (n = 12)		
1	9	1	<.066
2	9	3	<.004
3	11	1	<.001
	Novices (n = 12)		
1	7	3	n.s.
2	6	1	n.s.
3	7	1	n.s.

Table 4
Results From Contingency-Table Analysis of Diagnosis Accuracy (One Tailed)

	P + F > P	P > P + F	P + F = P
	Level 1 (p > .10)		
Experts	9	2	1
Novices	7	2	3
	Level 2 (p = .0211)		
Experts	9	0	3
Novices	6	5	1
	Level 3 (p = .0466)		
Experts	11	0	1
Novices	7	4	1

display over the P display 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 (Siegel, 1956). In this way, there were actually 24 individual experiments testing the theoretical prediction, one for each subject.

The results from this nonparametric analysis are presented in Table 3. For the experts, the results indicate that the P+F display was clearly better than the P display at Levels 2 and 3 (ps < .004 and .001, respectively) and approached significance at Level 1 (p < .066). For the novices, on the other hand, the effect of display was not significant at any of the three levels of analysis. In contrast with the ANOVAs, these results clearly indicate that the P+F display resulted in superior diagnosis when compared with the P display for experts but that there was no statistically significant difference between displays for novices.

The preceding test only provides a test of the effect of display for each of the two expertise groups. A more direct test of the display × expertise interaction can be performed using an 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 displays (see Table 4). 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 of the one-tailed test are illustrated in Table 4. At Level 1, the result failed to reach significance (p > .10). However, at Level 2, there was a statistically significant display × expertise interaction (p = .0211). The same result was obtained for Level 3 (p = .0466). Again in contrast with the ANOVAs, the results obtained from Levels 2 and 3 suggest that the experts benefited more from the P+F display than did the novices.

Correlation Between Diagnosis and Memory

Another way to evaluate the importance of presenting higher order functional variables is to examine the correlation between diagnosis and memory for physical and functional variables. If functional information 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. To test this claim, the 34 process variables were divided into two sets, 16 physical variables (those that are represented in both displays) and 18 functional variables (those that are displayed only in the P+F display). Two average memory measures were obtained for each trial, one for each variable class. Then, separate Pearson product-moment correlations were calculated between diagnosis scores at each of the three levels of resolution and memory for either physical or functional variables. Only the data from the semantic trials were analyzed. Data were averaged across trials within a session and then across interfaces within each subject, resulting in 24 data points, one per subject. For Level 1 diagnosis, the results indicate that diagnosis is significantly correlated with memory for functional variables [r(22) = −0.41072, p < .0462] but not with memory for physical variables [r(22) = −0.2017, n.s.]. For Level 2 diagnosis, the correlation between diagnosis and memory for functional variables is marginally significant [r(22) = −0.39996, p < .0528], whereas the correlation with memory for physical variables is again not significant [r(22) = −0.25769, n.s.]. A similar pattern of results was obtained for Level 3 diagnosis [r(22) = −0.35601, p < .0877, for memory for functional variables; r(22) = −0.23444, n.s., for memory for physical variables]. Thus, the pattern of results indicates that the better the diagnosis, the lower the memory error for functional variables. This finding shows that functional variables are important in diagnosing system state.

Correspondence Memory

The predictions regarding correspondence memory were evaluated in two ways. First, correspondence memory based on separate measures for physical variables (those presented in the P and P+F display) and functional variables (those presented only in the P+F display) was analyzed. This analysis allows one to evaluate the predictions made regarding display and expertise effects. Second, the global memory results averaged across all 34 variables were analyzed as a function of normal, fault, and random trial types (rather than the semantic vs. random classification). The latter analysis allows one to determine whether there is an expertise effect for fault trials.

Physical versus functional variables. As in the correlation analysis, two independent performance measures were obtained for each trial, memory for physical variables and memory for functional variables. A six-way ANOVA was then conducted with the five factors mentioned in the experimental design and a variables factor with two levels, physical and functional.

Before describing the results from this analysis, several points need to be mentioned. First, because they are not displayed, the functional variables must be derived from the physical variables in the P condition. This is not necessarily true in the P+F condition since all variables are displayed. Second, it is also important to realize that, under normal circumstances, it is possible to derive functional variables from physical variables (following the constraints listed in Table 2), whereas the inverse is not true. For example, knowledge of the flow rates may not allow one to derive the valve settings since various combinations of valve settings could account for a given set of flows. This information provides the context required to interpret the following results.

The effects pertinent to expertise, illustrated in Figure 6, will be described first. Several important results were obtained. First, experts outperformed novices [$F(1,16) = 18.66, p < .001$], thereby validating the expertise criterion adopted for the experiment. Second, the results indicated that type was a highly significant factor, with semantic trials clearly outperforming random trials, as expected [$F(1,16) = 211.83, p < .001$]. Third, the predicted interaction between expertise and trial type was also significant [$F(1,16) = 8.00, p < .0121$]. In contrast to the chess studies, however, simple-effect F tests revealed that there was a significant expertise advantage for both random and semantic trials [$F(1,16) = 8.34, p < .05$, and $F(1,16) = 52.23, p < .001$, respectively], although it was much greater for the latter than for the former. The advantage of semantic over random trials was statistically significant for both experts and novices [$F(1,16) = 184.3, p < .001$, and $F(1,16) = 85.3, p < .001$, respectively]. These effects are similar to the classic memory-expertise effects first observed in chess, with the exception that there was also a smaller but nonetheless significant expertise effect on random trials.

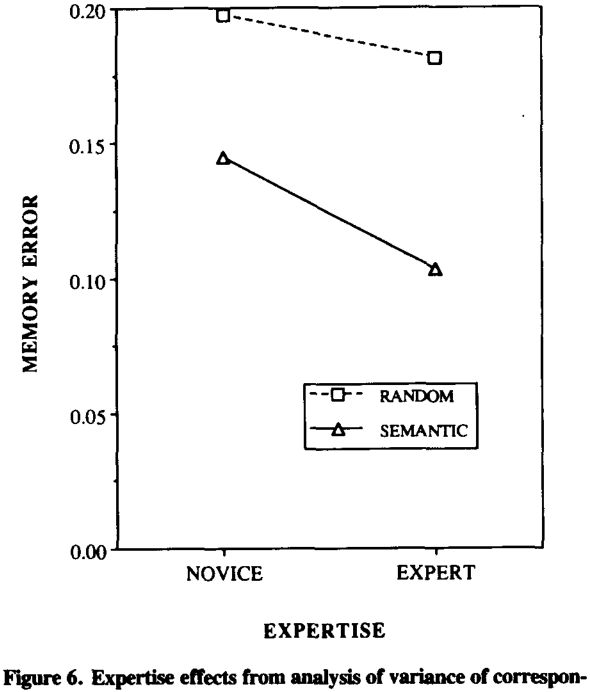


Figure 6. Expertise effects from analysis of variance of correspondence memory for physical and functional variables.

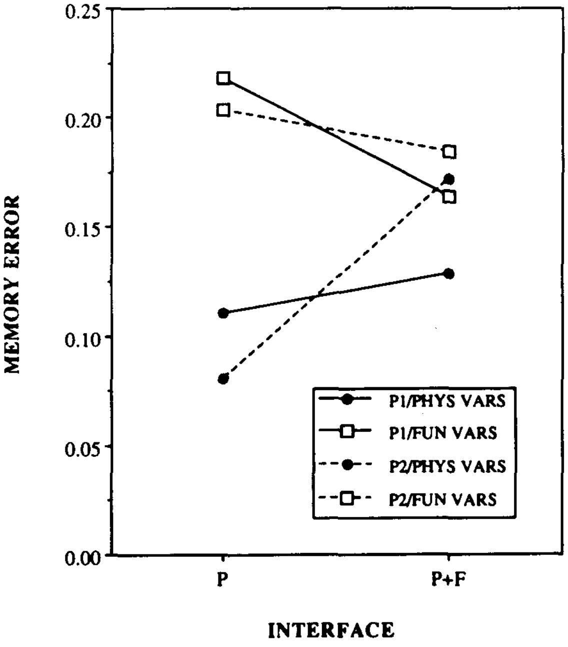


Figure 7. Order effects from analysis of variance of correspondence memory for physical and functional variables.

The results pertinent to the physical/functional variable distinction are illustrated in Figures 7 and 8. There was a highly significant main effect of variables [$F(1,16) = 493.72, p < .001$], with memory for physical variables being more accurate than that for functional variables. However, the variable effect interacted with trial type [$F(1,16) = 67.00, p < .001$], indicating that the difference between the two variable classes was significantly greater on random trials.

The display × variables interaction was also statistically significant [$F(1,16) = 549.21, p < .001$], as shown in Figure 7. For physical variables, the P display outperformed the P+F display [$F(1,16) = 295.49, p < .001$], whereas for functional variables, the P+F display outperformed the P display [$F(1,16) = 132.99, p < .001$]. This result makes intuitive sense if one considers the information represented in each display. With the P display, physical variables are remembered better than are functional variables because the former are displayed, whereas

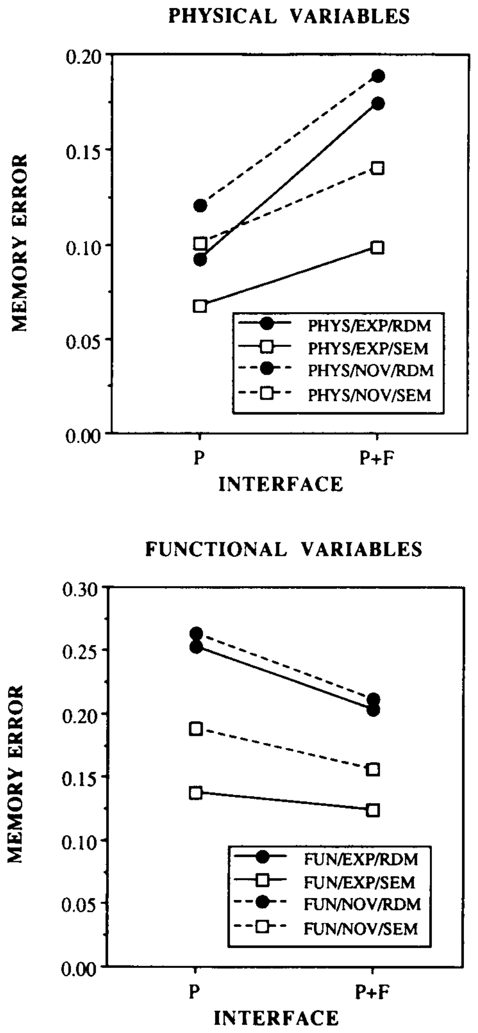


Figure 8. Primary results from analysis of variance of correspondence memory for physical and functional variables.

the latter must be derived. In contrast, with the P+F display, the functional variables are displayed, and so it is natural that they were "remembered" more accurately than with the P display. The fact that memory for physical variables is better with the P display can be attributed to the fact that there is less information in this display. While the subjects have to allocate their attention across both functional and physical variables with the P+F display, with the P display the subjects are only presented with the physical variables, thereby allowing them to focus their attention on a smaller number of variables (i.e., 16 variables as opposed to 34 variables for the P+F display).

The ANOVAs also revealed significant display × order and display × order × variables interactions [$F(1,16) = 32.71$, $p < .001$, and $F(1,16) = 23.85$, $p < .001$, respectively]. This result, illustrated in Figure 7, suggests that experiencing the information-laden P+F display first (order P2) results in a large improvement in memory for physical variables when transferring to the less complex P display, as compared with having the P display first (order P1). This may result from the fact that the subjects become accustomed to viewing a denser display, thereby making it easier to extract information from the less dense P display.

Referring now to Figure 8, the display × type × variables interaction was also highly significant [$F(1,16) = 55.83, p < .001$]. The important observation here is that the P+F display results in better memory than does the P display, but *only* for functional variables [$F(1,16) = 132.99, p < .001$]. This advantage was observed for both semantic and random trials [$F(1,16) = 25.71, p < .001$, and $F(1,16) = 126.29, p < .001$, respectively]. This result indicates that the predicted superiority of the P+F display is limited to the functional variables, which, as the correlation analysis indicates, are the variables most relevant to diagnosing system state.

A significant display × expertise × type × variables interaction was also obtained [$F(1,16) = 4.51$, $p < .0496$]. Given the nonparametric-analysis finding of a significant interaction between expertise and display, one might expect that this result would be caused by a significant display × expertise × type interaction for functional variables, indicating that the advantage of the P+F display on semantic trials is greater for experts than it is for novices. This would indicate that a certain amount of expertise is required to fully exploit the benefits of the P+F display. However, a simple-effect F test revealed that the interaction was not due to this type of effect [$F(1,16) = 1.49$, n.s.].

Normal versus fault versus random. The purpose of this second analysis of correspondence memory was to determine whether there was an expertise effect for fault trials. A five-way ANOVA identical to that described in the Method section was conducted, with the exception that the trial type factor consisted of three levels (normal, fault, and random) instead of two (semantic and random). Global correspondence memory (averaged over all 34 variables) was the dependent variable. The expertise effect and the display × order interaction were both significant [$F(1,16) = 18.50$, $p < .001$, and $F(1,16) = 30.72$, $p < .001$, respectively]. Experts outperformed novices, and memory was better for whichever display was experienced second.

Other significant effects are illustrated in Figures 9 and 10. Figure 9 shows that there was a highly significant main effect of trial type [$F(2,32) = 135.46, p < .001$]. Memory for normal trials was best, for fault trials second best, and for random trials the worst of all. All three means were significantly different from each other, as evidenced by a Neuman-Keuls pairwise comparison. The expertise × type interaction also attained significance [$F(2,32) = 6.24, p < .0052$]. Simple-effect F tests reveal that there was a significant expertise effect for each trial type. However, the advantage was greatest on normal trials [novices − experts = .05257, $F(1,32) = 53.4$, $p < .001$], second largest for fault trials [novices − experts = .03041, $F(1,32) = 17.9$, $p < .001$], and smallest for random trials [novices − experts = .01697,

$F(1,32) = 5.26, p < .05$]. The magnitude of the expertise advantage on normal trials is significantly greater than that on fault trials [$t(32) = 3.3731, p < .01$]. Similarly, the magnitude of the expertise advantage on faults trials is, in turn, significantly greater than that on random trials [$t(32) = 2.5572, p < .05$]. The latter result has important theoretical implications that will be discussed in the following section.

Finally, the display × expertise × type interaction was also significant [$F(2,32) = 5.37, p < .0097$], as illustrated in Figure 10. The interaction seems to be caused by the point representing novices' performance with the P display on normal trials, which is higher than the corresponding point for experts. A possible interpretation of this result is that experts can compensate for the P display during normal trials by deriving variables and thereby attaining a good correspondence score, whereas novices cannot.

Coherence Memory

The coherence memory data were aggregated for each trial by averaging over the nine pairs of algebraic constraints, thereby resulting in a global measure for each trial of how well the subjects' recall conformed to the relationships between variables that exist when the system is operating normally (see Table 2). A score of 0 represents perfect coherence, whereas a score of 1 indicates the worst possible coherence. This global measure of coherence memory was analyzed with a five-way ANOVA following the design described in the Method section.

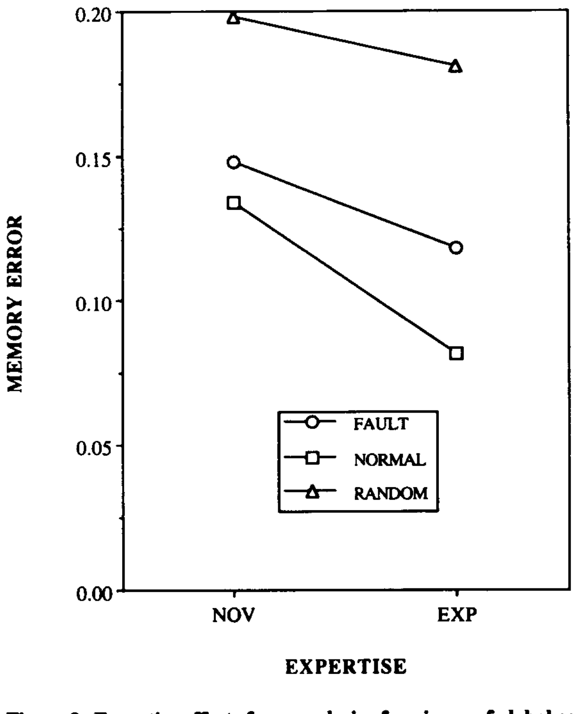


Figure 9. Expertise effects from analysis of variance of global correspondence memory for fault, normal, and random trials.

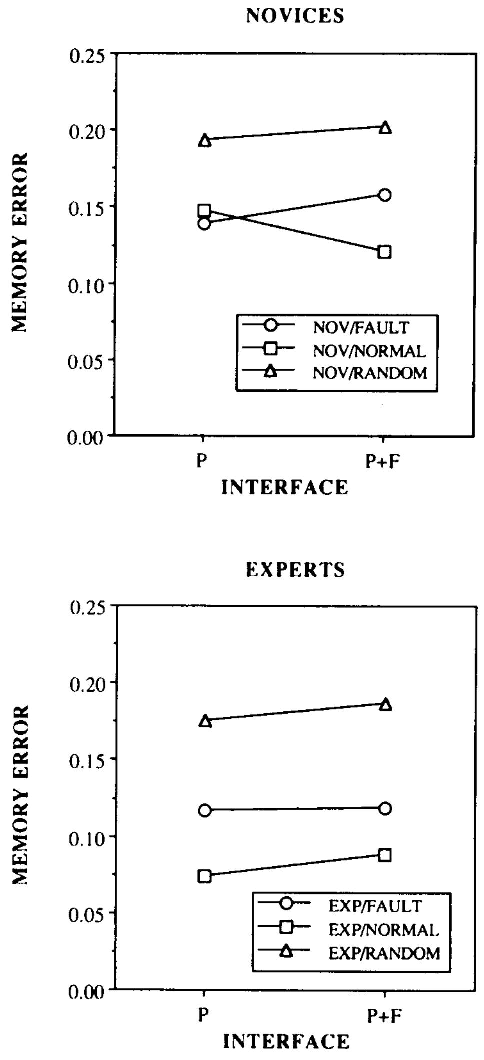


Figure 10. Primary results from analysis of variance of global correspondence memory for fault, normal, and random trials.

The results obtained from this analysis are illustrated in Figure 11. Several important findings were obtained. First, the expertise effect was significant in the expected direction [$F(1,16) = 11.57, p < .0036$], thereby validating the expertise criterion. Second, the effect of trial type was highly significant, with performance on the random trials being significantly worse than on the semantic trials [$F(1,16) = 103.59, p < .001$]. Third, the expertise × trial type interaction was also significant [$F(1,16) = 5.57, p < .0313$]. Experts outperformed novices on both semantic and random trials [$F(1,16) = 100.43, p < .001$, and $F(1,16) = 44.68, p < .001$, respectively], although the difference was greater on semantic trials. The advantage of semantic over random trials was statistically significant for both experts and novices [$F(1,16) = 78.6, p < .001$, and $F(1,16) = 30.56, p < .001$, respectively]. These results parallel those observed for correspondence memory.

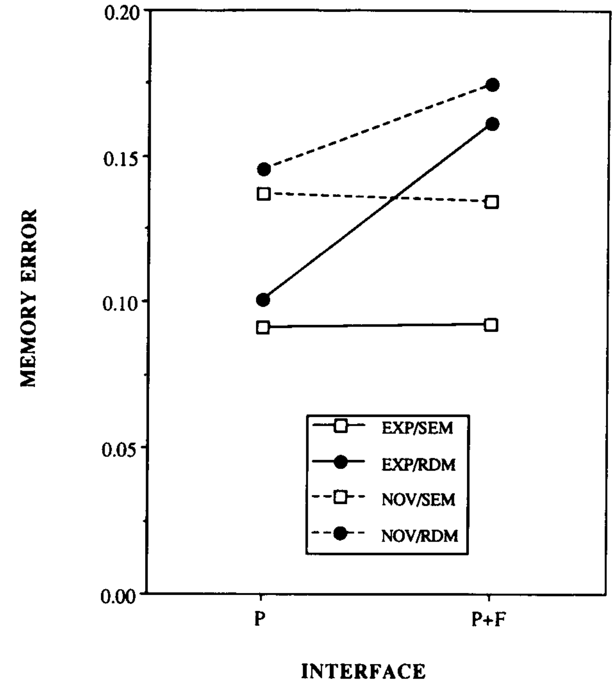


Figure 11. Primary results from analysis of variance of global coherence memory.

As shown in Figure 11, the effect of display was also significant, with the P display resulting in better coherence than the P+F display [$F(1,16) = 31.80, p < .001$]. However, this result can only be meaningfully interpreted within the context of the significant display × type interaction [$F(1,16) = 35.27, p < .001$]. Figure 11 indicates that the effect of display was specific to the random trial type. For semantic trials, there was no significant difference between the two displays [$F(1,16) < 1$]. In contrast, with random trials, the P+F display resulted in more incoherent memory than did the P display [$F(1,16) = 69.15, p < .001$]. Paradoxical as it may seem, worse coherence on random trials is actually a more appropriate response pattern. The reason for this is that the random trials themselves are not coherent because most of the constraints that hold between variables when the system is operating normally have been violated. Thus, for the P+F display, the subjects' recall is more coherent when the stimulus is coherent (semantic trials) and more incoherent when the stimulus itself is incoherent (random trials). This difference between trial types on the P+F display is statistically significant [$F(1,16) = 100.6, p < .001$]. In contrast, with the P display one does not observe this discrimination between random and semantic trials. The coherence memory for both types of trials is virtually identical [$F(1,16) = 2.65$, n.s.].

DISCUSSION

Expertise Effects

The diagnosis results will be discussed first. The picture here is a fuzzy one since the results from the parametric and nonparametric tests are inconsistent. With the traditional ANOVA, the expertise and display × expertise effects were not statistically significant. The nonparametric tests, on the other hand, indicated a difference between novices and experts in their ability to diagnose system state as a function of display. This conflict is probably due to the fact that there is more variability in the novice group than in the expert group. When the data for the two groups are analyzed together, as in the ANOVA, it is reasonable that the high variance of the novice group will overwhelm any effects of which that group is a part. However, when the two groups are treated separately, as in the nonparametric test, the high variance of the novices leads to a nonsignificant result, but it no longer suppresses the significant effect for experts. Although these results do not allow one to derive firm conclusions, it should nevertheless be noted that there are good reasons for putting more weight on the results obtained from the nonparametric tests (cf. Dar, 1987; Hammond et al., 1987; Meehl, 1967, 1978).

With regard to the pretest results, two clear findings emerged. First, experts clearly outperformed 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.

As for memory, the results indicated that experts strongly outperformed novices on meaningful trials. This finding is similar to the memory-expertise effects obtained in various other domains. However, this seems to be the first time that this result has been obtained with continuous, dynamic stimuli of the sort presented to the subjects in this experiment. Thus, one contribution of this study was to generalize a finding from basic psychological research to a new domain, process control. It is also worthwhile pointing out that the same result was obtained for both correspondence and coherence memory measures. This also seems to be the first time that expertise effects in memory recall have been evaluated according to degree of coherence. The typical measure has traditionally been one of correspondence. Thus, a second contribution of this research is that the memory-expertise effect was replicated with a measure of coherence memory.

There were also some unexpected results. In contrast with the results from chess, in the present study there was a small but significant expertise advantage on random trials. One possible explanation for this finding is that novices and experts differed on cognitive abilities that were not assessed in this experiment (e.g., memory capacity). Another finding that was not predicted was the strong expertise advantage obtained for fault trials. These two findings bear on an important question: Under what conditions is an expertise advantage to be expected in memory tasks?

A likely answer is that there will be a memory-expertise advantage in cases where there are goal-relevant constraints that experts can exploit to structure the material with which they are presented. The more constraint available, the greater the expertise advantage should be. In fully random events, there are no constraints and thus an expertise advantage would not be expected. This conjecture will subsequently be referred to as the *constraint-attunement hypothesis.*[4] Although it is a new theoretical explanation for expertise effects in memory recall, this hypothesis has existed for years as the cornerstone of ecological theories of skill acquisition (see Flach, Lintern, and Larish, 1990; Fowler & Turvey, 1978; Gibson, 1969; Owen, 1990).

The constraint-attunement hypothesis can explain why the significant expertise advantage for fault trials is significantly larger than that for random trials. In these scenarios, only one or two constraints that usually govern the system under normal operating conditions are violated. Thus, the vast majority of the constraints listed in Table 2 still hold. The operating constraints allow experts to structure the situation and thereby enable them to remember more than novices, despite the fact that the event is abnormal and has never been observed before.

The hypothesis can also explain why there was a weaker but still significant expertise advantage on random trials. Because of the sampling procedure that was adopted, these trials were actually pseudorandom in that a few of the constraints governing the system under normal circumstances still held (see Method). Since there are still some (albeit fewer) constraints to pick up on, the expertise advantage is much smaller but still significant under these conditions.

The constraint-attunement hypothesis also accounts for the results of Myles-Worsley et al. (1988) described earlier. The critical finding from this experiment was that recognition memory increased as a function of expertise on abnormal X rays, but decreased with expertise for normal X rays. This result would seem paradoxical if one equated abnormal X rays with random chess positions and normal X rays with normal chess positions. However, the goal in reading an X ray is to detect any existing abnormality. Thus, the goal-relevant constraints are those relations that signify an abnormal X ray, not those that characterize a normal X ray. From this perspective, the results of Myles-Worsley et al. are easily interpreted. Experts were more attuned to the goal-relevant constraints in the X rays than were novices, and novices were more attuned to irrelevant information in the X rays than were experts.

To summarize, a third contribution of this research is a clarification of the boundary conditions under which expertise advantages are to be expected in memory-recall tasks. According to the constraint-attunement hypothesis, there will be a memory-expertise advantage in cases where there are goal-relevant constraints that experts can exploit to structure the stimulus, and the more constraint available, the greater the expertise advantage will be (see Vicente, 1991, for an application of this principle in the context of the various layers of constraint in chess).

A final issue pertinent to expertise that is usually investigated in memory-recall studies is the amount of clustering in subjects' recall. Various studies have found that experts' categorization classes tend to be organized according to functional properties, whereas novices' classes tend to be organized according to surface features (Glaser & Chi, 1988). To see if this pattern of results was observed here, the degree of clustering in the subjects' recall was analyzed according to two criteria for categorizing the variables being recalled. One classification scheme was based on the two-way physical/functional distinction. A second classification scheme grouped variables according to their surface features. This resulted in the eight mutually exclusive categories listed in Table 1. The results of these clustering analyses, reported in Vicente (1991), indicate that the effects that have typically been found in recall studies were not observed in this experiment. Neither analysis led to any significant effects of expertise. Instead, the organization in recall tended to be driven by the structure of the response format (Figure 3) and the task demands.

Display Effects

The first step in evaluating the validity of memory recall as a measure of display effectiveness is to independently establish which of the two displays provides greater support for problem solving. There was strong evidence, both from parametric and nonparametric analyses, indicating that the P+F display was indeed superior to the P display in terms of diagnosis. This result is consistent with the a priori prediction based on the principles of EID. Explicitly representing higher order functional information in the display enhances problem-solving performance. There was also some evidence from nonparametric tests to indicate that this advantage was greater for experts than for novices.

Having established the superiority of the P+F display, the next question is: Does that superiority reflect itself in the memory-recall measure? The results revealed that the P+F display resulted in better memory than did the P display, but *only* for functional variables. This is not very surprising, since the functional variables were presented in the P+F display, but not in the P display. Thus, a simple interpretation of this result is that memory is worse for variables that have to be derived than it is for those that have to be recalled. The problem with this interpretation is that it fails to account for why there are marginally significant correlations between diagnosis accuracy and memory for functional variables, but *not* memory for physical variables. This result shows that there is a close association between functional variables and the ability to accurately diagnosis system state, thereby providing empirical justification for the theoretical motivation for including higher order functional information in the P+F display. This correlational result also suggests an alternate interpretation of the display effects: memory for those variables most relevant to diagnosis is better with the P+F display than with the P display.

Is this finding really surprising? One might argue that all that has been demonstrated is that a display with more information is better than a display with less information. This may seem to be a reasonable criticism to some, so it is important to clearly point out why the advantage of the P+F display cannot be solely attributed to more information. Although 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 a display that had the information that was in the P display as well as some extra information. 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 display containing more information than the P display. The key is that the added levels of information are *goal relevant*, as the significant correlation between diagnosis and memory for functional variables shows. The EID framework used to design the P+F display provides a *principled* approach to identifying the goal-relevant information that needs to be included in the display.

In summary, the findings indicate that the P+F display resulted in better diagnosis performance than did the P display because the former represented the state of functional variables, whereas the latter did not. In addition, the correspondence memory measure was sensitive to this display advantage, as indicated by the superior memory for functional variables for the P+F display as compared with the P display. The significant correlation between memory for functional variables and diagnosis performance suggests that the advantage of the P+F display on the memory task was localized to those variables that were most relevant for diagnosis.

CONCLUSIONS

This study has demonstrated that memory-recall performance on meaningful trials in a process control system varies as a function of expertise, thereby generalizing a classic finding in the literature to a novel domain. This result was obtained with measures of both correspondence and coherence memory. It was also found that explicitly representing higher order functional information in a display can result in enhanced performance, as evidenced by the fact that the P+F display resulted in much better diagnosis performance than did the P display. Furthermore, memory for the subset of variables that was most relevant for diagnosis was observed to be better with the P+F display than with the P display, thereby indicating that memory recall can be a sensitive measure of display effectiveness. Finally, the constraint-attunement hypothesis was proposed as a novel theoretical explanation for the conditions under which expertise effects in memory-recall tasks are to be expected. This hypothesis accounts for the results of this experiment, those of Myles-Worsley et al. (1988) in X-ray diagnosis, and various other memory-recall experiments (cf. Vicente, 1991).

On a more general note, the present work also bears on the relationship between basic research and applied concerns. Traditionally, these two areas of interest have been relatively segregated. Those concerned with basic research sometimes shun applications as being of little scientific value, whereas those faced with applied problems have often pointed to the irrelevance of basic research. In the human-factors community, these tendencies have unfortunately led to a conflicting dichotomy between basic research and applied problems (for discussions, see Flach, 1990; Rouse, 1985). However, some have argued that there need not be a conflict between basic research and applied concerns, and even more strongly, that an interaction between basic research and pragmatic challenges can be of great benefit to both interests (Gomez & Dumais, 1986; Landauer, 1987).

The research presented here supports the latter view. A basic finding from psychological research, the relationship between memory recall and expertise, was adapted to address an applied problem, evaluation of degree of display support for problem-solving activities. Furthermore, adopting an applied context as a test bed led to methodological and theoretical insights into basic research in the form of the coherence measure of memory recall and the constraint-attunement hypothesis, respectively. Therefore, although addressing a relatively narrow and concrete set of issues, the present work speaks to a much broader issue, namely, the fruitful interplay that can be achieved between basic research and applied problems.

REFERENCES

Charness, N. (1979). Components of skill in bridge. *Canadian Journal of Psychology*, **33**, 1-16.

Chase, W. G., & Simon, H. A. (1973a). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215-281). New York: Academic Press.

Chase, W. G., & Simon, H. A. (1973b). Perception in chess. *Cognitive Psychology*, **4**, 55-81.

Coughlin, L. D., & Patel, V. L. (1987). Processing of critical information by physicians and medical students. *Journal of Medical Education*, **62**, 818-828.

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist*, **42**, 145-151.

Deakin, J. M., & Allard, F. (1991). Skilled memory in expert figure skaters. *Memory & Cognition*, **19**, 79-86.

de Groot, A. D. (1965). *Thought and choice in chess*. The Hague: Mouton. (Original work published 1946)

Egan, D. E., & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory & Cognition*, **7**, 149-158.

Ericsson, K. A., & Staszewski, J. J. (1989). Skilled memory & expertise: Mechanisms of exceptional performance. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 235-267). Hillsdale, NJ: LEA.

Flach, J. M. (1990). The ecology of human-machine systems I: Introduction. *Ecological Psychology*, **2**, 191-205.

Flach, J. M., Lintern, G., & Larish, J. F. (1990). Perceptual motor skill: A theoretical framework. In R. Warren & A. Wertheimer (Eds.), *The perception and control of self motion* (pp. 327-356). Hillsdale, NJ: LEA.

Fowler, C. A., & Turvey, M. T. (1978). Skill acquisition: An event approach with special reference to searching for the optimum of a function of several variables. In G. E. Stelmach (Ed.), *Information processing in motor control and learning* (pp. 1-40). New York: Academic Press.

Gibson, E. J. (1969). *Principles of perceptual learning and development.* New York: Appleton-Century-Crofts.

Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv-xxviii). Hillsdale, NJ: LEA.

Gomez, L. M., & Dumais, S. T. (1986). Putting cognitive psychology to work: Examples from computer system design. In T. J. Knapp & L. C. Robertson (Eds.), *Approaches to cognition: Contrasts and controversies* (pp. 267-290). Hillsdale, NJ: Erlbaum.

Hammond, K. R., Hamm, R. M., & Grassia, J. (1986). Generalizing over conditions by combining the multitrait-multimethod matrix and the representative design of experiments. *Psychological Bulletin*, **100**, 257-269.

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of intuitive and analytical cognition in expert judgement. *IEEE Transactions on Systems, Man, & Cybernetics*, **SMC-17**, 753-770.

Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science*, **8**, 255-273.

La Cour Christensen, P., Kofoed, J., & Larsen, N. (1988). *PC-DYSIM: A program package for simulation of continuous dynamic processes* [User's manual]. Roskilde, Denmark: Risø National Laboratory, Department of Information Technology. (Risø-I-342)

Landauer, T. K. (1987). Relations between cognitive psychology and computer system design. In J. M. Carroll (Ed.), *Interfacing thought: Cognitive aspects of human-computer interaction* (pp. 1-25). Cambridge, MA: MIT Press.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, **34**, 103-115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting & Clinical Psychology*, **46**, 806-834.

Myers, J. L. (1972). *Fundamentals of experimental design* (2nd ed.). Boston: Allyn & Bacon.

Myles-Worsley, M., Johnston, W. A., & Simons, M. A. (1988). The influence of expertise on X-ray image processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 553-557.

Owen, D. H. (1990). Perception and control of changes in self-motion: A functional approach to the study of information and skill. In R. Warren & A. Wertheimer (Eds.), *The perception and control of self motion* (pp. 289-326). Hillsdale, NJ: LEA.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, & Cybernetics*, **SMC-13**, 257-266.

Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology*, **80**, 16-20.

Rouse, W. B. (1985). On better mousetraps and basic research: Getting the applied world to the laboratory door. *IEEE Transactions on Systems, Man, & Cybernetics*, **SMC-15**, 2-8.

Schneider, W., Körkel, J., & Weinert, F. E. (1989). Domain-specific knowledge and memory performance: A comparison of high- and low-aptitude children. *Journal of Educational Psychology*, **81**, 306-312.

Schumacher, R. M., & Gentner, D. (1988). Transfer of training as analogical mapping. *IEEE Transactions on Systems, Man, & Cybernetics*, **SMC-18**, 592-600.

Sheppard, S. B., Curtis, B., Milliman, P., & Love, T. (1979, December). Modern coding practices and programmer performance. *IEEE Computer*, **12**, 41-49.

Shneiderman, B. (1977). Measuring computer program quality and comprehension. *International Journal of Man-Machine Studies*, **9**, 465-478.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Vicente, K. J. (1988). Adapting the memory recall paradigm to evaluate interfaces. *Acta Psychologica*, **69**, 249-278.

Vicente, K. J. (1991). *Supporting knowledge-based behavior through ecological interface design.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Vicente, K. J., & de Groot, A. D. (1990). The memory recall paradigm: Straightening out the historical record. *American Psychologist*, **45**, 285-287.

Vicente, K. J., & Rasmussen, J. (1990). The ecology of human-machine systems II: Mediating "direct perception" in complex work domains. *Ecological Psychology*, **2**, 207-250.

Vicente, K. J., & Rasmussen, J. (in press). Ecological interface design: Theoretical foundations. *IEEE Transactions on Systems, Man, & Cybernetics.*

NOTES

1. This random control condition has often been incorrectly attributed to de Groot (cf. Vicente & de Groot, 1990).

2. Two additional considerations should be mentioned. First, the study here differs from most previous recall studies in that the subjects are familiar with the content of the stimulus but not with its perceptual form. However, several studies have shown that the classic interaction between expertise and meaningfulness of stimulus is also obtained under these conditions (cf. Recht & Leslie, 1988; Schneider, Körkel, & Weinert, 1989; and the review in Vicente, 1988), thereby lending support for applying the recall method in this manner. Second, if such a result is obtained with theoretical experts as subjects, then one would need to see if the same pattern of results is obtained with experienced operators who are typically not theoretical experts. The latter issue has not been addressed in the present research. If the memory measure is validated and the boundary conditions of its validity established, then future studies could evaluate interfaces using the memory-recall measure alone.

3. It might be useful at this point to provide some justification for the conditions under which the experiment was conducted. First, the subjects were not given any feedback so that the novices would remain such. Second, the subjects were not very familiar with the perceptual properties of the display, they were presented with unfamiliar events, and they were not given very much time to view each scenario. This was to ensure that the task was challenging enough so that the subjects would be forced into a problem-solving mode (i.e., Rasmussen's, 1983, knowledge-based behavior), which was the focus of this study.

4. The constraint-attunement hypothesis is a theoretical claim regarding product, not process. That is, it tries to predict under what conditions there will be an expertise effect in recall tasks, not the psychological processes that cause such effects.

(Manuscript received May 8, 1991;
revision accepted for publication March 20, 1992.)