A Longitudinal Study of the Effects of Ecological Interface Design on Skill Acquisition

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

Human Factors: The Journal of the Human Factors and Ergonomics Society 1996 38: 523
DOI: 10.1518/001872096778701917

HUMAN FACTORS, 1996, 38(3), 523–541

A Longitudinal Study of the Effects of Ecological Interface Design on Skill Acquisition

KLAUS CHRISTOFFERSEN, CHRISTOPHER N. HUNTER, *and* KIM J. VICENTE,[1] *University of Toronto, Toronto, Ontario, Canada*

Results from evaluations of interface design concepts conducted over short durations may not generalize to longer time spans. In an attempt to address this issue, this paper presents a longitudinal study of the effect of interface design on skill acquisition in which participants' quasi-daily performance was observed over an unprecedented period of six months. The research was conducted in the context of DURESS II, a real-time, interactive thermal-hydraulic process control simulation that was designed to be representative of industrial systems. The performance of two interfaces was compared, one containing physical and functional (P+F) system representations based on the principles of the ecological interface design framework, and a more traditional interface based solely on a physical (P) representation. Participants were required to perform several control tasks, including start-up, tuning, shutdown, and fault management. The results indicate that the P interface led to significantly less-consistent performance than did the P+F interface; with the former, participants occasionally took up to 2 times longer to complete the required tasks, even after 5.5 months of daily practice. There was very little difference in average performance between the two groups. These results have important implications for designing interfaces that lead to efficient performance under normal operating conditions.

INTRODUCTION

The vast majority of experimental evaluations of interface design concepts reported in the literature have been conducted over a short period, usually ranging from a few hours to, at most, 1 h/day for a few weeks. For several reasons, however, experimental results obtained over the short term may not generalize to the longer term. First, people may be able eventually to compensate for the deficiencies of a poor interface. If so, then performance disadvantages observed in the short term will disappear with practice. Second, differences in effort might also diminish over time. Thus if one interface is found initially to lead to less mental workload, one cannot be certain that this effect will persist over the long term, as the user's processing becomes more automatic.

Third, the full effects of skill acquisition may not be observable over a short period. This is especially true for complex human-machine systems; it may take some time before users can become well adapted to the demands placed on them. In this situation, one might find that differences between interfaces only appear over

[1] Requests for reprints should be sent to Kim J. Vicente, Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada M5S 3G8; benfica@mie.utoronto.ca.

an extended period. Fourth, an interface that initially leads to "better" performance may actually lead to performance deterioration in the long run. For example, people could become overly dependent on the perceptual features of the interface and thereby exhibit a form of "deskilling" over time (see Wickens, 1992).

Therefore, there are many reasons to suggest that short-term evaluations of interface design concepts may generate potentially misleading experimental results. This is important from an applications perspective, given that interfaces for complex work environments will be used on a daily basis for years once they are introduced into the workplace. Thus there is a significant need to investigate the effects of interface design concepts over longer periods.

This paper describes a longitudinal study of ecological interface design (EID), a theoretical framework for designing computer interfaces for complex work environments. In the study, participants controlled DURESS (DUal REservoir System Simulation) II, a real-time, interactive process control simulation, for about 1 h approximately every weekday for a total of six months. To address the impact of interface design on skill acquisition, participants controlled the system with one of two interfaces: an interface based on the principles of EID containing physical and functional (P+F) system representations and a more traditional interface based solely on a physical (P) representation.

BACKGROUND

Theory

EID is a theoretical framework for designing interfaces for complex work environments. It is based on the skills, rules, knowledge taxonomy of levels of cognitive control (Rasmussen, 1983). The framework includes three prescriptive design principles, each directed at providing the appropriate interface support for a specific level of cognitive control. First, to support skill-based behavior, operators should be able to act directly on the interface. Furthermore, the structure of the displayed information should be isomorphic to the part-whole structure of movements. Second, to support rule-based behavior, the interface should maintain a consistent one-to-one mapping between the work domain constraints and the perceptual cues provided in the interface. Third, to support knowledge-based behavior, the interface should represent the work domain in the form of an abstraction hierarchy (Rasmussen, 1985), which can serve as an externalized mental model to support problem solving. This system model contains both physical and functional representations of the system.

Note that, in contrast to other interface design frameworks (e.g., Wickens & Carswell, 1995), EID provides guidance for identifying what information should be included in the interface, not merely for determining the visual form of the interface. See Vicente and Rasmussen (1990, 1992) for a more detailed description and justification of these design principles.

DURESS II

DURESS II is an updated, interactive version of DURESS, a thermal-hydraulic process control microworld that has been used in previous research (Bisantz & Vicente, 1994; Vicente, Christoffersen, & Pereklita, 1995). The original DURESS was a noninteractive simulation that presented participants with real-time, "canned" scenarios. With that simulation, participants were able to view the behavior of the system but not to control it.

DURESS II, in addition to some minor structural changes, differs from the original primarily in that it allows for real-time interaction so that participants may actively control the system. It is important to note that DURESS and DURESS II were designed to be representative of complex work domains, thereby promoting generalizability of research results to operational settings (Vicente, 1991).

The physical structure of DURESS II is illustrated in Figure 1. The system consists of two

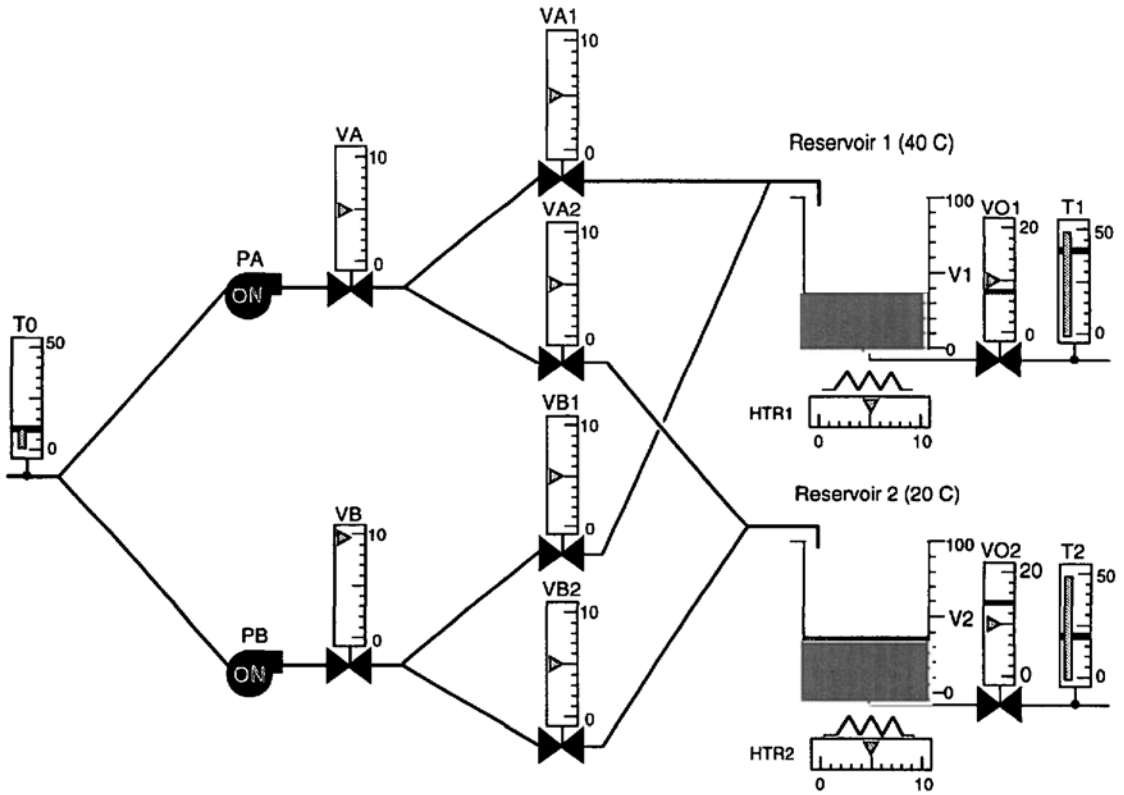

Figure 1. P interface for DURESS II.

redundant feedwater streams that can be configured to supply water to either, both, or neither of the two reservoirs. Associated with each reservoir is an externally determined demand for water that can change over time. The system purposes are twofold: to keep each of the reservoirs at a prescribed temperature (40° C and 20° C), and to satisfy the current mass (water) output demand rates. To accomplish these goals, the participant has control over eight valves (VA, VA1, VA2, VO1, VB, VB1, VB2, and VO2), two pumps (PA and PB), and two heaters (HTR1 and HTR2). All of these components are governed by first-order lag dynamics, with a time constant of 15 s for the heaters and 5 s for the remaining components. The system input temperature (T0), reservoir output temperatures (T1 and T2), and volumes for both reservoirs (V1 and V2) are also displayed in Figure 1.

P Interface

The interface shown in Figure 1, a physical (P) representation of DURESS II, displays only the state of the physical components and the goal variables. The first meter on the extreme left of the display is a thermometer (T0) measuring the inlet water temperature. The vertical bar increases in height as the water temperature increases. The normal inlet water temperature is 10°C, as indicated by the thin area on the T0 scale. After the thermometer, the input water stream splits and flows to two pumps (PA and PB) that operate as discrete switches (on or off). The participant uses a mouse to click on the pump to change its state. The pumps are displayed in black (with white lettering) if they are off and in light gray (with black lettering) if they are on. The maximum flow rate through each

pump is 10 units/s. If either pump is turned on without any of the downstream valves being opened, the pump will fail after approximately 5 s. This error terminates the trial.

The next set of components are the primary valves (VA and VB), which have a continuous range from 0 to 10. The valve state is set using a mouse; the participant either drags the triangular pointer to the desired setting or simply clicks on the scale at the desired point. From these primary valves, each feedwater stream splits into two secondary valves connecting each stream to both reservoirs. The secondary valves (VA1, VA2, VB1, and VB2) operate in the same manner as the primary valves. The water then flows to each of the two reservoirs, where it is heated and removed through the use of the heaters (HTR1 and HTR2) and the output valves (VO1 and VO2) in order to meet the temperature and demand goals, respectively.

The reservoirs have a maximum capacity of 100 units. Reservoir volume levels are indicated by a scale on the side of each reservoir and by the shaded area depicting water in the reservoir. It is possible to overflow either of the reservoirs if input flow rate is consistently greater than output flow rate. When reservoir volume exceeds the maximum capacity of 100 units, the trial ends automatically.

The heaters (HTR1 and HTR2) also have a continuous range of 0 to 10. The participant can either slide the triangular pointer on the heater scale to the desired setpoint or click on the scale itself at the desired point. Heating an empty reservoir for an extended period will lead to a malfunction. Thus if there is continued heat transfer to a reservoir without any water in that reservoir, then the system will fail and the trial will end. The water temperature in the reservoirs is displayed with thermometers (T1 and T2). The goal temperature is represented as a thin area on the temperature scale. There is a tolerance of ±2°C from the setpoints (40°C for Reservoir 1 and 20°C for Reservoir 2). If the water in the reservoir boils, the system fails and the trial ends.

Finally, participants also have control over the outlet valves (VO1 and VO2) that are used to meet the demand goals. These valves operate in the same manner as the other valves, except that their maximum setting is 20. The demand for each reservoir is indicated by a thin area on the valve setting scale. This goal area, which is ±1 unit around the desired level, moves as a function of changes in the demand.

This interface design format was chosen because it is typical of how existing computer interfaces for process control systems have been designed. To some readers, it may seem like a straw man because it lacks some information (e.g., flow rates for the valves in the feedwater streams). However, this is not atypical of actual control rooms. For example, the Three Mile Island control room did not display the flow rate from the emergency auxiliary feedwater system to the steam generators (Malone et al., 1980). Similarly, the Biblis plant in Germany did not have a pressure sensor for the isolation valve separating the primary system from the low-pressure injection system; as Becker (1991) pointed out, this led to the flow of radioactive water to the outside of the containment.

Other subsystems that are poorly instrumented and that occasionally give rise to complications can also be found in industrial systems (e.g., Vicente & Burns, 1995). Given that (a) such examples exist, (b) it is precisely these subsystems that give operators problems in fault situations, and (c) virtually no research has been done on the effects of physical and functional information on operator performance prior to our studies, we feel that the P interface is not a straw man but, rather, serves as a meaningful baseline condition.

P+F Interface

The P+F interface, based on the principles of EID, is illustrated in Figure 2. The input water thermometer, both pumps, and all the valves operate in the same manner as in the P interface. However, the P+F interface also contains higher-order functional information, identified through an abstraction hierarchy analysis of DURESS (see Vicente & Rasmussen, 1990) that describes the state of the functions that the

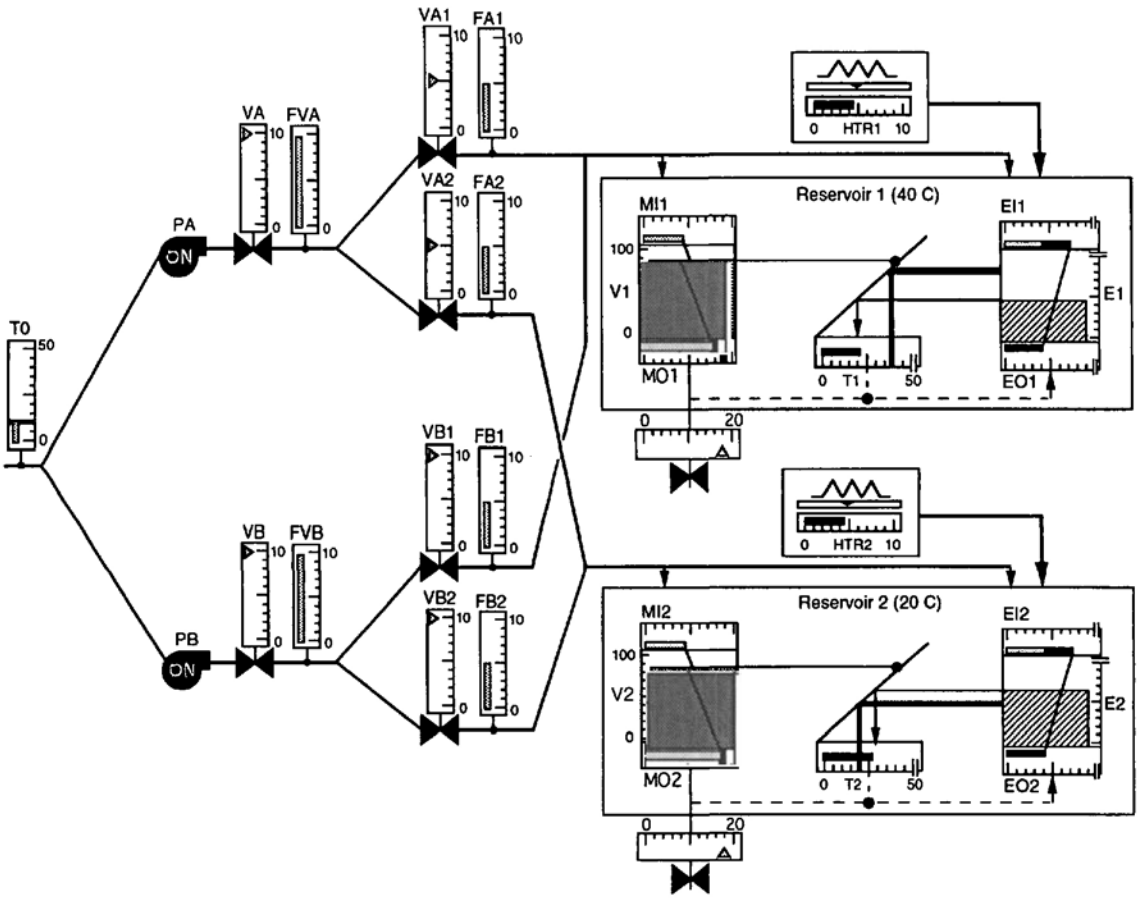

Figure 2. P+F interface for DURESS II.

physical components are intended to achieve. Thus each valve also has a flow meter next to it (FVA, FVB, FA1, FA2, FB1, FB2, and MO1 and MO2 for the mass output flow rates). These flow meters have the same value range as their respective valves.

The boxed group of graphics on the right of Figure 2 provides additional higher-order functional information in the form 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 graphic on the right represents the energy balance. Both representations operate in a similar manner. Referring to Reservoir 1, the various inputs are shown at the top of the graphics (MI1 for mass and EI1 for energy). Inventories for each representation are indicated by scales on the side of each graphic (V1 for volume/mass and E1 for energy). The outputs, MO1 for mass and EO1 for energy, are shown at the bottom of each graphic. The energy inputs to each reservoir (EI1 and EI2) are partialled out according to the two contributors. Thus the energy added by the feedwater stream is shown as the lightly shaded bar, and the energy added by the heater is shown as a dark bar.

Intuitively, the mass and energy graphics rely on a funnel metaphor. For example, if the bottom is wider than the top (i.e., output > input), as is the case with the mass balance for Reservoir 2 in Figure 2, then it is easy to visualize the consequence—namely, that volume should

decrease. Thus the slope of the line represents the rate at which the mass (or energy) inventory should be changing. If input equals output, then the line is perpendicular, indicating that the level should not be changing.

The graphic in the middle, between the mass and energy balances, illustrates the relationship between mass, energy, and temperature. A horizontal line with a ball on the end emanates from the current mass inventory level (V1 and V2). Changes in the height of this line always accompany any change in mass inventory (i.e., the bar will always be at the same height as the water level, V1 or V2).

The diagonal line in the center display rotates around its leftmost endpoint (connected to the top left of the T1 box) and 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 diagonal line in the center display. For example, if volume increases, the horizontal line goes up, causing the diagonal to rotate counterclockwise, increasing the slope of the diagonal line. The slope of the diagonal line represents the function that maps the relationship between mass and energy onto temperature.

This mapping is indicated by the line emanating from the current energy inventory level (E1 and E2) that comes across and reflects off the diagonal line at a right angle down onto the temperature scale (T1 and T2). The goal temperature is indicated by the thin shaded area on the temperature scale. This goal area reflects back from the temperature scale, off of the diagonal line, and onto the energy inventory scale. In addition, off-scale markers are added to the output temperature scales and the energy input, inventory, and output scales as well. These were added to the interface by creating a gap in the scale at the off-scale point, thereby allowing participants to discriminate the maximum value from off scale (Mumaw, Woods, & Eastman, 1992).

For a detailed description of the rationale behind the design of the P+F interface, see Vicente and Rasmussen (1990). For a more detailed illustration of the interface, see Pawlak and Vicente (1996).

Purpose of This Study

Previous empirical investigations of EID have compared the P and P+F interfaces for DURESS (Vicente et al., 1995) and DURESS II (Pawlak & Vicente, 1996). Although these studies have led to a promising set of findings, many issues have not yet been addressed. Perhaps one of the most salient is the effect that an EID interface can have on operator skill acquisition over a prolonged period. As we mentioned in the introduction, one cannot assume that the short-term results of the effects of interface design on performance can necessarily generalize to the longer term. This paper addresses this issue by describing a study that was carried out over six months to ensure that the long-term influences of EID on skill acquisition could be meaningfully investigated. The primary questions addressed here are (a) Is there an effect of interface on skill acquisition for normal (i.e., nonfault) trials? and (b) Is there an interface effect for normal trials after extensive practice?

As Crossman (1958) observed several decades ago, acquisition of skill can lead not only to changes in mean performance but also to changes in performance variability. Thus we investigated both types of changes in this study.

METHOD

Experimental Design

A repeated-measures, between-subjects design with the type of interface (P or P+F) as the primary manipulation was adopted for this experiment. Participants were assigned to one of the two interfaces and participated for a total of six months. At the end of the experiment, a transfer manipulation was conducted with participants controlling the system for six trials with the alternate interface.

Participants

The criteria adopted for selection of participants were constrained by considerations of

representativeness and experimental control. With regard to representativeness, in some countries nuclear power plant operators have a high school background, whereas in other countries they have engineering degrees. Also, in Canada at least, all licensed nuclear power plant operators are men. With regard to control, it was important to select participants who were reliable, who would be willing to endure the full duration of the experiment, and who did not have many extended periods during which they would not be able to come in on a daily basis.

Another relevant consideration was that previous research on tasks involving manual control of dynamic systems has shown significant gender effects (e.g., Jagacinski, Greenberg, Liao, & Wang, 1993). As a result of all these considerations, the decision was made to recruit male participants (to enhance representativeness and experimental control) who either worked at the university (making it easier for them to participate reliably) or were graduate students at the university (and who therefore did not have a two-week final examination period, as did undergraduates). An attempt was also made to recruit participants with a science and engineering background to enhance representativeness.

As a result of this selection process, six men ranging in age from 23 to 32 years participated in the study. The participants, with one exception, had either science or engineering backgrounds. Participants were assigned to interface groups so as to roughly match for background. A summary of the participant groups is presented in Table 1.

Each participant was paid $5 per session. These regular wages were paid approximately every six weeks. A bonus of $2 per session was offered for completing the entire experiment. Extra bonuses were also offered for "good" performance, although participants were not given any details about how performance would be measured. These additional bonuses were paid only on completion of the experiment and were designed to maintain participants' motivation over the course of the six months of the experiment.

Apparatus

The DURESS II simulation runs on a Silicon Graphics IRIS Indigo R4000 computer workstation. The simulation code was written in C, whereas the two interfaces were constructed using a graphical construction set called FORMS. Verbal protocols were collected using a Sony CCD-TR81 Hi-8 Handycam with an external microphone.

Experimental Tasks

During the experiment, participants performed four different types of control tasks:

Start-up. For this task the participant was presented with a shut-down system and was asked to bring the system to steady state, meeting predefined setpoints consisting of temperature and demand goals for each reservoir.

Tuning to new setpoints. In this task the participant needed to bring the system from an online, steady-state initial condition to a pair of new steady-state demand setpoints.

Shutdown. During this task the participant was required to bring the system from an

TABLE 1

Summary of Participants' Backgrounds, by Interface Group

Interface	Participant	Educational Background
P + F	IS	Medical genetics (Ph.D. level)
	AS	Mechanical engineering (master's level)
	AV	Industrial engineering (master's level)
P	TL	Electrical engineering (master's level)
	WL	Geophysics (Ph.D. level)
	ML	Commerce/political science (bachelor level)

on-line, steady-state condition to a shut-down condition.

Fault management. After the introductory phase of the experiment, when participants had a reasonable amount of practice at controlling the system, they were occasionally presented with trials during which a fault would occur. Participants were told that their task was to detect, diagnose, and compensate for any such faults. Fault management results will not be presented in this paper, so this aspect of the method will be described only briefly.

For all control tasks, *steady state* was defined as maintaining both reservoirs in the goal regions (both temperature and output demand) for five consecutive minutes.

Trial Types

Trials normally consisted of the first three control tasks described earlier, performed either in isolation or in sequence within the same trial. Flow rate demand setpoints for start-up and tuning tasks were varied to keep participants from consistently adopting simplistic control strategies. New tuning setpoints were chosen to promote the use of multiple or complex control strategies within a trial. During the introductory phase, the demand pairs were selected to gradually increase in complexity (as determined by a cognitive work analysis of DURESS II, Vicente & Pawlak, 1994).

Procedure

There were four distinct phases in the experiment: an introductory session, in which the experimental protocol was explained to participants; an introductory practice phase, during which the complexity of tasks was gradually increased; an extended practice phase, characterized by repeated exposure to normal trials and occasional exposure to faults; and a final phase examining the long-term effects of each interface on operators' knowledge. This final phase included three fault trials and six transfer trials. The entire experiment lasted six months, during which participants controlled the system for about 1 h/day approximately every weekday.

There were 224 trials per participant. Six days, referred to as test days, were also set aside during the experiment to have the participants perform a set of knowledge elicitation tests. The experiment concluded with a debriefing session. A detailed schedule of experimental sessions is provided in Christoffersen, Hunter, and Vicente (1994).

Introductory session. The introductory session was the same for both groups. Participants were (a) presented with a description of the experiment, (b) asked to complete an informed consent form and a demographic questionnaire, and (c) provided with a technical description of DURESS II. Participants then received a complete list of the system variable names so that they could become familiar with those labels. A pretest questionnaire consisting of 20 multiple-choice questions was then administered to assess participants' prior knowledge of thermal hydraulics. Thirty minutes were allotted for completing the questionnaire.

After the pretest questionnaire, participants were given an explanation of the interface to which they had been assigned, as well as a description of the different types of trials that they would encounter throughout the experiment, as outlined earlier. The experimenter answered questions about the interface the participants would be using, though questions about system functioning (i.e., the constraints governing the system) were not answered, as this was left to the participant to discover through experience with the system. A copy of all of the forms and instructions that were used in this introductory session can be found in Christoffersen et al. (1994).

Introductory practice phase. In this phase, lasting approximately one month, participants were gradually introduced to the various types of control tasks that they would be performing. They began by performing trials consisting of a start-up task alone. After 22 trials, they began performing trials consisting of a start-up task

immediately followed by a shutdown task. Finally, after 44 trials, participants were asked to perform trials consisting of a start-up task, followed by a tuning task, followed by a shutdown task. This last type of trial will hereafter be referred to as a *standard trial.*

Throughout the experiment, knowledge of results was never provided to the participants. However, in both interfaces a timer measuring the elapsed time in each trial was continually displayed in the upper left corner of the screen. As a result, participants could monitor the time they took to complete the various control tasks, if they so desired. In addition, when the participant damaged a system component, causing the simulation to stop prematurely, a message describing the component failure was displayed (e.g., "Reservoir 1 heated empty"). These messages provided participants with feedback that they could use to modify their control strategies.

Extended practice phase. This was the longest phase of the experiment, lasting approximately four months. During this phase participants repeatedly performed standard trials, as described earlier. Nine fault management trials were distributed over the course of this phase. Collection of verbal protocols also began during this phase. The purpose of the protocols was primarily to examine participants' performance on fault trials. During this phase participants received a two-week rest for Christmas and New Year holidays and a one-week rest during spring break.

Examination of long-term effects. In the final phase, constituting the final month of the experiment, we examined the effects of the large amount of experience with the system that participants had accumulated. This phase included three fault management trials. The experiment concluded with six transfer trials, for which the participant groups switched to using the other group's interface. The goal of the transfer manipulation was to determine whether the results obtained up to that point were interface dependent or participant dependent. At this point, the new interface was described to the participants and any questions they had about the interface only were answered. Participants were also told that the tasks they would be required to do would be the same as before.

Performance Measures

Many measures of performance were adopted in this experiment, but the primary measure pertinent to the analyses presented in this paper was trial completion time. This coarse measure was subdivided to obtain times for each of the requisite tasks in a trial (i.e., start-up, tuning, and shutdown).

RESULTS

Only the results pertinent to performance in normal trials will be described here (see Christoffersen et al., 1994, for additional analyses). The section is organized into five subsections corresponding to analyses of the effect of interface on skill acquisition, asymptotic performance, number of trials not completed, transfer performance, and participants' comments.

Skill Acquisition

The first set of analyses examine the effect of interface design on skill acquisition. Fault and transfer trials were not included, and incomplete trials (i.e., blow ups) were treated as missing data. The most interesting results were those obtained from analyses of performance variability, so these data are presented first.

Performance variability. Using Crossman's (1958) graphing procedure, we examined the distributions of times for total trials and for the individual subtasks as well. Frequency distributions were taken for successive blocks of trials over the course of the experiment. In order to more closely examine the introductory phases of the experiment wherein a large proportion of participants' learning was expected to occur, the analysis employed a block size of 22 trials for the first two blocks and block sizes of approximately 40 trials for the remaining data. Complete graphs for each participant for each trial type are included in Christoffersen et al. (1994).

The two interface groups seemed to progress

in a similar manner, with the notable exception that the P+F group seemed to display consistently less variability than did the P group. This pattern can be clearly seen when one compares Figures 3 and 4, which show start-up time data for a P+F and P participant, respectively. Even for the last block of trials, representing performance after about five months of practice, the P participant exhibits extreme outliers in his distribution. Similarly, the other P participants occasionally had very slow times, even after a great deal of experience. In contrast, the P+F participants were much more consistent, with few outliers. This pattern was observed in all

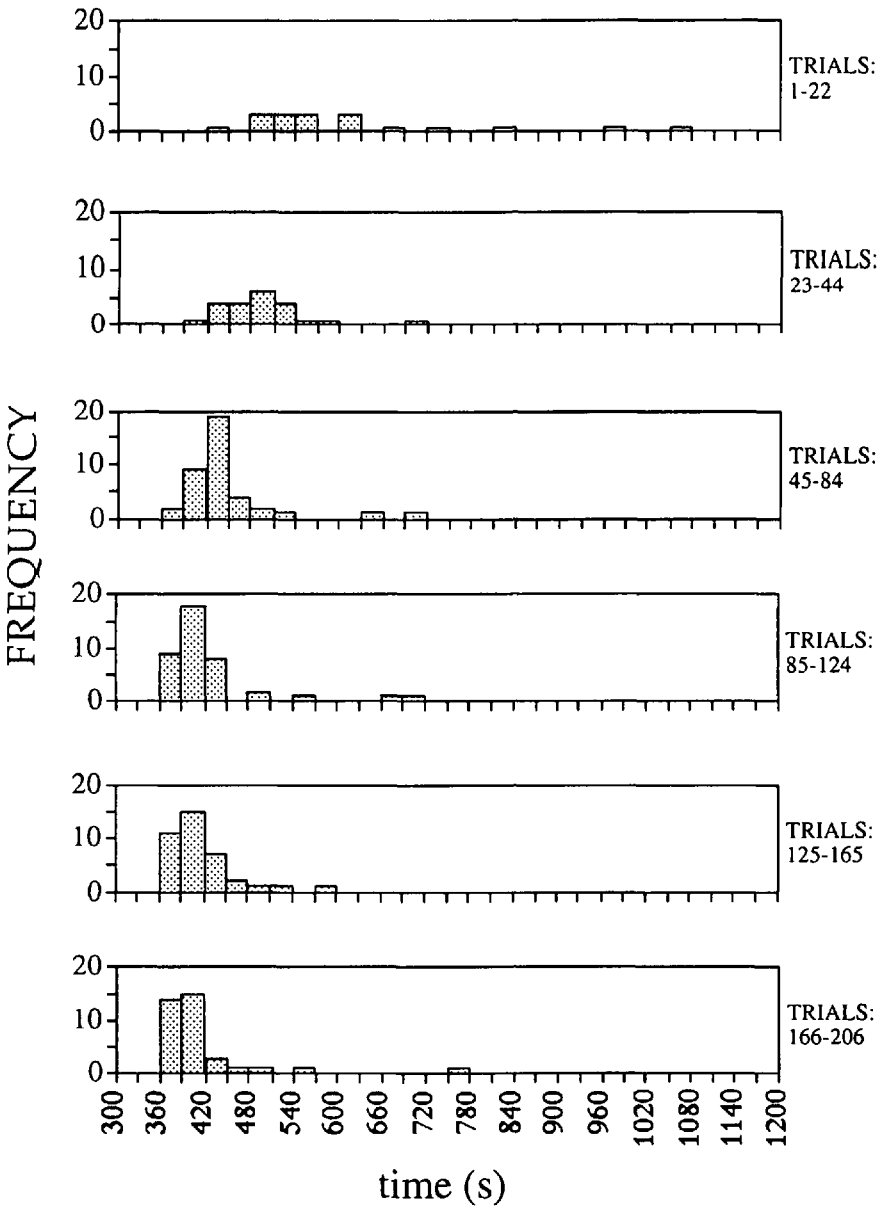

Figure 3. Distributions of start-up times at successive points in the experiment for a typical P participant.

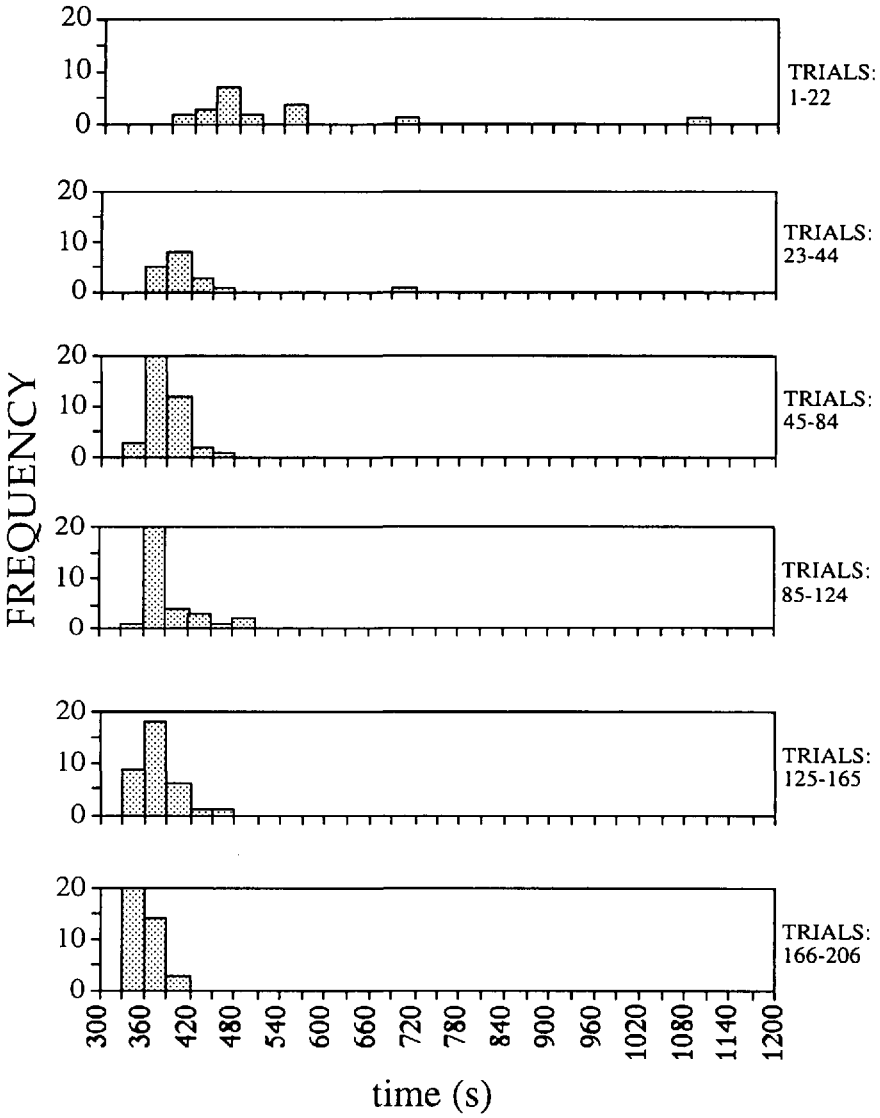

Figure 4. Distributions of start-up times at successive points in the experiment for a typical P+F participant.

control tasks but was most notable for tuning tasks.

Four Cochran tests (Winer, 1971) were conducted to determine whether these differences in variability were statistically significant. The results are summarized in Table 2. For total trial times, a Cochran test revealed that the P group displayed a significantly higher variance than did the P+F group. This effect can be seen in Figure 5, which shows the greater variability of the P group, particularly toward the end of the experiment. For the start-up task, a Cochran test revealed that, in this case, the P+F group exhibited a significantly higher variance than did the P group. This was caused in large part by the early performance of Participant AS in the P+F group. During the first several trials, he had a particularly difficult time in reaching steady

TABLE 2

Results of Cochran Tests (C) for Skill Acquisition Performance

Task	Interface	Variance	$C(2\ 456) =$	$p <$
Total time	P	32 970	0.6099	.01
	P + F	21 090		
Startup	P	11 880	0.6022	.05
	P +F	17 990		
Tuning	P	11 109	0.8025	.01
	P + F	2 735		
Shutdown	P	1 954	0.6098	.01
	P + F	1 253		

state, resulting in times that were as much as three to four times longer than those of most of the other participants.

As for the tuning task, a Cochran test revealed a highly significant difference between the variances of the two interface groups; the P+F group was much more consistent than the P group. Figure 6 shows the high variability the P group exhibited relative to the P+F group, particularly in the second half of the experiment. For the shutdown task, a Cochran test revealed that the P+F group again had a significantly smaller variance in completion times than did the P group.

Mean performance. Four two-way ANOVAs (with interface and trial number as the independent variables) were conducted to examine

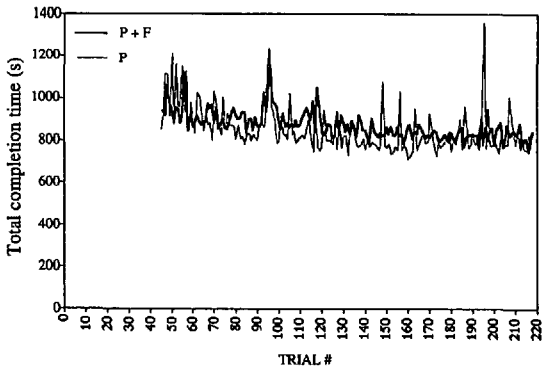

Figure 5. Interface × Trial interaction for total trial times.

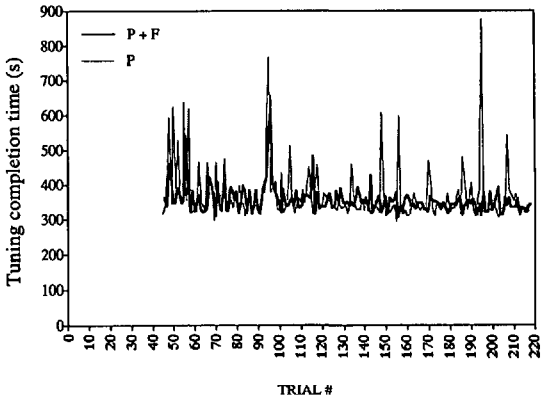

Figure 6. Interface × Trial interaction for tuning times.

mean performance: one for total trial completion time and one for the time on each individual control task (start-up, tuning, and shutdown). For total trial times, the analyses start with data from Trial 45 (the first standard trial). No significant differences were found between the interface groups for total time, $F(1, 4) = 0.08$, n.s. Not surprisingly, trial was statistically significant, $F(161, 579) = 1.48$, $p < .0006$.

There was also a significant interaction between interface and trial, $F(161, 579) = 1.27$, $p < .0256$. The Interface × Trial interaction can be seen in the plot of average completion times illustrated in Figure 5. There appears to be a crossover effect, with the P+F group being slightly faster initially. The P group seems to equal the performance of the P+F group around Trial 60 to 70 and then to surpass it for most of the remainder of the experiment. In the last group of 20 to 30 trials, however, both groups seem to be performing at an approximately equal level once again.

The ANOVA of the start-up trials began with data from Trial 1 (the first start-up trial). Again, there was no significant difference between interface groups, $F(1, 4) = 0.20$, n.s. Trial was again highly significant, $F(205, 723) = 6.64$, $p < .0001$. However, in this case, there was no interaction between trial number and interface, $F(203, 723) = 0.93$, n.s.

For tuning times, the ANOVA included data from Trial 45 (the first trial including a tuning

task) onward. Once again, no significant differences were observed between the interface groups, $F(1, 4) = 1.07$, n.s. Trial was significant, $F(161, 575) = 2.79$, $p < .0001$, as was the interaction between trial and interface, $F(161, 575) = 1.58$, $p < .0001$. Figure 6 contains the plots of the average tuning times by interface group against trial number. The Interface × Trial interaction seems to consist of a convergence of the two groups' times. The P+F group began as slightly faster, but the two groups appear to grow more comparable in their times as the experiment continued.

Finally, the ANOVA of the shutdown times included data from Trial 23 (the first trial including a shutdown task) onward. For this task there was a significant interface effect, $F(1, 4) = 9.51$, $p < .0368$; the P group had a mean of 78.1 s and the P+F group had a mean of 97.9 s. Trial was significant, $F(183, 651) = 3.18$, $p < .0001$, but there was no significant interaction between trial and interface, $F(183, 651) = 0.93$, n.s.

Asymptotic Performance

The analyses in this section investigate the differences between groups toward the end of the experiment, once they had reached roughly asymptotic performance.

Performance variability. Cochran tests were conducted to investigate differences in variability between the two interface groups over the last 20 nonfault, nontransfer trials. Again, four tests were conducted, one for the total trial time and one for each of the individual control tasks. The results are summarized in Table 3. For total trial, start-up, and tuning times, the P group's variance was significantly greater than that of the P+F group, sometimes by 300% or 400%. For the shutdown times, no significant differences were observed, but the trend was in the same direction.

TABLE 3

Results of Cochran Tests (C) for Asymptotic Performance

Task	Interface	Variance	$C(2\ 55) =$	$p <$
Total time	P	18 225	0.6981	.01
	P + F	7 885		
Startup	P	6 131	0.8010	.01
	P + F	1 521		
Tuning	P	8 046	0.8510	.01
	P + F	1 414		
Shutdown	P	1 289	0.6140	n.s.
	P + F	812		

Mean performance. Another four ANOVAs were performed on participants' performance over the last 20 nonfault, nontransfer trials. Again, interface and trial number were the independent variables, with completion time for the total trial and for each of the three individual subtasks as the dependent measures.

Detailed results of these analyses are presented in Christoffersen et al. (1994), but considering that the findings were relatively uniform, only a brief summary will be provided here. First, there were no significant differences between interfaces in average times for the final phase of the experiment on normal trials. One interpretation of this result is that the two groups had reached an indistinguishable level of performance after 5½ months of practice. Alternatively, this null result may also have been attributable to the small sample size. Second, the results show no effect of trial or interactions between interface and trial, indicating that performance on normal trials had roughly stabilized by the end of the experiment.

Trials Not Completed

As mentioned earlier, several equipment failure modes were modeled in the DURESS II simulation. These failure modes were intended to add an element of risk to the simulation, serving as a representative set of behavior-shaping constraints. A number of χ^2 analyses were conducted to investigate whether the frequency of blowups varied according to participant, trial, and interface (Christoffersen et al., 1994). However, none of these analyses revealed any significant effects.

Transfer Trials

For the final six trials of the experiment, a transfer manipulation was conducted so that participants who had been using the P+F interface used the P interface, and vice versa. Of these six trials, the first four were normal standard trials (the last two were fault trials). The goal of this transfer manipulation was to explore which aspects, if any, of participants' performance were interface dependent and which were based on participants' knowledge. If the effects documented in the preceding section are attributable to differences in the interfaces, then one would expect that the relative differences between the two groups' performance would be reversed when they used a different interface. In contrast, if the documented effects were attributable merely to differences in participants' abilities or background, then using a different interface should have no effect on the relative performance of the two groups.

Performance variability. Cochran tests for homogeneity of variance were performed for the transfer trials. The results are presented in Table 4. (The interface categorizations refer to the interfaces used during the transfer trials, not the original interface groups.) These results show once again that the group using the P interface (the original P+F group) displayed a consistently higher degree of variability in their completion times than did the group that used the P+F interface during the transfer trials. This is an important result because it clearly shows that the P+F interface had a dramatic effect in helping participants to perform more consistently under normal conditions. The relative differences in variability are tied to the interfaces themselves, rather than resulting from the individual abilities or background of the participants in the two groups.

TABLE 4

Results of Cochran Tests (*C*) for Transfer Trial Performance

Task	Interface	Variance	$C(2\ 9) =$	$p <$
Total time	P	227 529	0.8188	.05
	P + F	50 176		
Startup	P	104 976	0.8613	.05
	P + F	16 900		
Tuning	P	63 001	0.8902	.01
	P + F	7 744		
Shutdown	P	4 761	0.9123	.01
	P + F	441		

Note. The interface categorizations refer to the interfaces used during the transfer trials, not the original interface groups.

Mean performance. Again, completion times for total trials and for individual tasks were examined. In order to provide a baseline against which to measure participants' performance during these trials, the average completion times from the previous 20 normal trials for each participant were calculated and incorporated as data points in the analyses. (Recall that the analyses of asymptotic performance showed that there was no significant learning effect for these 20 trials, thereby justifying the decision to take the average of these data.) Unsuccessfully completed trials (in which a system component was damaged) were treated as missing data.

Four ANOVAs were conducted with interface and trial as the dependent variables. The only significant interface effect occurred for shutdown completion time, $F(1, 4) = 9.89, p < .0347$. This effect shows that members of the original P group retained the performance advantage they exhibited before the transfer trials, despite switching interfaces. Their average time was 78.9 s, whereas the original P+F group had an average time of 119.7 s with the transfer interface. It seems, therefore, that the differences that led to the original P group's superior performance on shutdown were not interface dependent but, rather, the result of individual participants' strategies, which transferred to the P+F interface.

The effect of trial was significant for total time, $F(4, 12) = 4.31, p < .0217$, and for shutdown time, $F(4, 12) = 3.61, p < .0373$. We examined these two main effects in more detail by performing pairwise comparisons using Student-Newman-Keuls (SNK) tests ($\alpha = .05$). For total trial time, the analysis showed that the first transfer trial was significantly slower than

all other trials, with the exception of the second transfer trial, which was not significantly different from the first. Apart from the first transfer trial, all trials—including the baseline trial—did not differ significantly. In other words, there was a general decrement in performance on the first transfer trial, but participants generally approached their original performance levels again rapidly.

The SNK test for the trial effect on shutdown performance did not reveal any differences between individual trials. Nevertheless, the order of the times was identical to those obtained for total trial completion times, with the first transfer trial being the slowest, followed by the second, third, baseline, and fourth transfer trials.

Participants' Comments

After each of the six transfer trials, participants were asked to comment on the new interface. All of the participants found at least some features of the P+F interface to be beneficial. This is noteworthy because the participants who transferred to the P+F interface had only six trials in which to become accustomed to it. After the first transfer trial, all these participants stated that the P+F interface appeared much more complicated and more difficult to use than the P interface. However, after only a few more trials, all of them felt, to varying degrees, that the P+F interface could help them to control the process more effectively. The two fault trials presented at the end of the transfer sessions had a particularly important impact in forming this opinion.

Conversely, participants who transferred to the P interface complained that the functional information to which they were accustomed was gone. In general, they said that they had to simplify their strategies and make inferences about system state because there was not enough information present in the interface to directly perceive the state of the system and therefore control it reliably. Thus participants' preference for the P+F interface is consistent with the results obtained from the objective performance measures.

DISCUSSION

The results bearing on the two questions posed at the beginning of this paper will now be discussed in turn.

1. Is there an effect of interface on learning/adaptation for normal trials? The graphing technique developed by Crossman (1958) showed that the P+F group's performance was generally more consistent than that of the P group. The P participants occasionally had trials that were much slower than usual, sometimes by a factor of almost two. These differences were confirmed by a series of Cochran tests, which showed that the P+F group's times were significantly more consistent than those of the P group for total trial times, tuning times, and shutdown times.

In contrast, the P+F group exhibited a significantly higher variance for start-up times over the course of the experiment. This effect seemed to be caused by the performance of Participant AS, who, for the first several trials, had a particularly difficult time in reaching steady state. This resulted in times that were as much as three to four times longer than those of most of the other participants. This suggests that the P+F interface may be more difficult to use initially for some people.

The ANOVAs revealed only two reliable differences in learning performance for the entire task or its constituent subtasks, as measured by the completion times for each. First, the P group showed a significant speed advantage for shutdown. Subsequent analyses of participant strategies revealed that this difference was caused by a strategy difference (Christoffersen et al., 1994). Participant TL in the P group and Participant AV in the P+F group discovered efficient strategies for shutdown, but TL adopted this strategy much earlier in the experiment than did AV. It is not known why the P interface should lead one to adopt this strategy earlier than does the P+F interface.

Second, the interaction between trial number

and interface group was significant for total trial times and for tuning times. With respect to the former, there appeared to be a crossover effect, with the P+F group being slightly faster initially. The P group seemed to equal the performance of the P+F group around Trial 60 to 70 and then to surpass it for most of the remainder of the experiment. In the last group of 20 to 30 trials, however, both groups seemed to be performing at an approximately equal level once again. As for the significant interaction on tuning times, the P+F participants appeared to be faster initially, but the difference between groups was gradually reduced over time. By the end of the experiment the two groups' times were comparable.

2. Is there an interface effect for normal trials after extensive practice? The Crossman analyses and the Cochran tests showed significant differences in variability between interface groups after extensive practice. The P group exhibited greater variability for total trial time, start-up time, and tuning time. This difference is particularly striking when one considers that participants had had almost 200 trials of almost daily practice over 5½ months by this point.

Why did the P+F interface lead to more consistent performance than the P interface? Recall that participants were required to maintain each reservoir within its goal conditions for five *consecutive* minutes before steady state was defined to be reached. Thus if either the temperature or output goal condition was violated, participants would be forced to reattain the goal conditions and maintain them for an additional five minutes in order to move on to the next phase of the trial. Falling out of the goal conditions could obviously inflate completion times significantly, particularly if the miscue occurred when the five required minutes within the goal region were nearly completed.

It was hypothesized that this difference in the number of outliers was attributable to the fact that participants using the P+F interface were able to control the reservoir temperatures more precisely as a result of the extra information available to them. If this were the case, one would expect to see evidence of it in the relative proficiency of the groups at maintaining the temperatures within the goal regions.

To keep the analysis to a manageable scale and to minimize the influence of learning effects, we decided that only the asymptotic phase would be examined in detail. Plots of reservoir temperatures versus time were created for each trial during the asymptotic period for each individual participant. For each trial, the number of times a participant overshot or fell below the goal regions (of either reservoir) after initially entering them were counted. The inevitable drop in temperatures during the shutdown phase was ignored. The results indicated that the P group violated the temperature goal boundaries a total of 76 times, whereas the P+F group went outside the temperature goal region a total of 45 times. A χ^2 test revealed that this difference was highly significant, $\chi^2(1) = 7.942$, $p < .01$.

The results of this analysis lend clear support to the hypothesis that the P+F interface allows participants to control temperature more precisely, even after extensive practice. Although there is no direct evidence bearing on this issue, it seems plausible that the added information in the P+F interface (in particular, the energy balance graphics; the display showing the relationships among energy, mass, and temperature; and the heat transfer rate) allowed these participants to exert more consistent control over the system.

The results from the normal transfer trials show how powerful this effect is. When the P+F participants moved to the P interface, their performance became significantly more variable than that of the P participants, who had moved to the P+F interface. This result clearly shows that the relative differences in variability observed for nontransfer trials were tied to the interfaces themselves, not to the participants. The fact that this effect was significant within a mere four trials after switching interfaces further emphasizes this result.

The ANOVAs of the last 20 normal trials failed to show any significant differences between

groups or any learning effect. Thus the mean performance of the two groups was statistically indistinguishable after extensive practice. Also, performance seemed to have stabilized by this point.

CONCLUSIONS

The primary result to emerge from this longitudinal study of the effect of EID on skill acquisition was that the P+F interface leads to more consistent performance than the P interface. This effect is specific to the interface, not to the participants, and still holds after extensive practice. For the most part, the two interface groups were comparable in terms of average performance on normal trials. This in itself may be a surprising result to some, considering that the P+F interface appears to be visually more complex than the P interface. However, the data from the experiment show that there is no substantial performance cost caused by this added information, except perhaps initially less consistent performance for some people.

The fact that these results were obtained over a period of six months makes generalization to operational settings more likely. As we will discuss, these results have several implications for the design of interfaces for complex human-machine systems.

Limitations

There are several important limitations to this study. First, the experimental design does not allow us to determine the extent to which the observed effects were caused by differences in content versus differences in visual form between the two interfaces. However, previous research (Vicente, 1991) indicates that the advantage of the P+F over the P interface cannot be explained solely by differences in visual form. Second, the participants received no training and had to engage in discovery learning. There may be an interaction between training and interface design. Such effects should be addressed in future research.

Third, there were only three participants in each group, so some null results (e.g., the ANOVAs for asymptotic performance) may have resulted from a small sample size. However, in a costly longitudinal study of this type, it is difficult to overcome this limitation. Fourth, the generalizability of these results to industrial-scale systems and to systems outside the process control domain also needs to be established.

Design Implications

This study provides additional empirical support for the principles of EID. These results indicate that it is possible to design an interface that supports operators effectively during both normal and abnormal conditions. The EID framework was geared primarily (though not exclusively) to support operators during unanticipated fault events (Vicente & Rasmussen, 1992). The results from the fault management portion of this experiment (see Christoffersen et al., 1994) show that the P+F interface exhibits a clear advantage over the P interface under such circumstances. However, the results presented in this paper also show that the P+F interface can lead to better performance under normal situations as well, in the form of greater consistency.

If operators are able to control the system more consistently with an EID interface, as the results of this study suggest, then a great deal of money could potentially be saved. For example, performance variability may lead to a greater chance of gradually entering unsafe plant conditions. By allowing operators to be more consistent, an EID interface could minimize the frequency of these costly excursions, especially in industrial systems that have many more failure modes than DURESS II.

Reduced performance variability may also allow operators to achieve plant goals in a more cost-effective manner because, in many cases, deviations from predefined operating regions are a source of inefficiency, even if they never lead to unsafe conditions. In addition, because most complex systems operate under normal conditions the vast majority of the time, small economic gains realized during nonfault

conditions can add up to significant savings in the long run.

The practical relevance and potential benefits of EID have not gone unnoticed by industry. For example, both Honeywell and AECL Research have incorporated portions of the P+F interface for DURESS into demonstration prototypes that are intended to represent the state of the art in advanced interface design. More important, Toshiba in Japan has adopted EID as the basis for designing its advanced control room for a next-generation boiling water reactor plant (Monta et al., 1991). It has also incorporated and adapted specific features of the P+F interface for DURESS (e.g., the mass balance graphics) into some of its displays. This application is notable in that it has been conducted at the scale of a full-scope nuclear power plant simulator.

More recently, Mitsubishi Heavy Industries in Japan has also demonstrated a strong interest in EID (Watanabe, Takaura, Fujita, & Hayashi, 1995) and has contracted Battelle to initiate a multiyear research program focusing solely on EID (Lee & Sanquist, 1995). This technology transfer to industry must be followed by large-scale evaluations if EID is to be defensibly used in designing interfaces for complex systems.

ACKNOWLEDGMENTS

This research project was sponsored by a contract from the Japan Atomic Energy Research Institute (Fumiya Tanabe, contract monitor) and by a graduate scholarship, a research grant, and an equipment grant from the Natural Sciences and Engineering Research Council of Canada. We thank Dimitris Nastos, Andy Sun, and Dr. Tanabe for their contributions to this research, and Bill Howell and several anonymous reviewers for their comments on earlier drafts. Finally, we are extremely grateful to the six participants who consistently devoted so much of their time to this experiment. Without their continued cooperation, this study would not have been possible.

REFERENCES

Becker, G. (1991). Analysis of human behaviour during NPP incidents: A case study. In *Balancing automation and human action in nuclear power plants* (pp. 517–526). Vienna, Austria: International Atomic Energy Agency.

Bisantz, A. M., & Vicente, K. J. (1994). Making the abstraction hierarchy concrete. *International Journal of Human-Computer Studies, 40,* 83–117.

Christoffersen, K., Hunter, C. N., & Vicente, K. J. (1994). *Research on factors influencing human cognitive behaviour I* (CEL 94–05). Toronto: University of Toronto, Cognitive Engineering Laboratory.

Crossman, E. R. F. W. (1958). A theory of the acquisition of speed-skill. *Ergonomics, 2,* 153–166.

Jagacinski, R. J., Greenberg, N., Liao, M., & Wang, J. (1993). Manual performance of a repeated pattern by older and younger adults with supplementary auditory cues. *Psychology and Aging, 8,* 429–439.

Lee, J. D., & Sanquist, T. F. (1995). *Review of ecological interface design research: Applications of the design philosophy and results of empirical evaluations* (Battelle Tech. Report). Seattle, WA: Battelle.

Malone, T. B., Kirkpatrick, M., Mallory, K., Eike, D., Johnson, J. H., & Walker, R. W. (1980). *Human factors evaluation of control room design and operator performance at Three Mile Island-2* (NUREG/CR-1270). Washington, DC: U.S. Nuclear Regulatory Commission.

Monta, K., Takizawa, Y., Hattori, Y., Hayashi, T., Sato, N., Itoh, J., Sakuma, A., & Yoshikawa, E. (1991). 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). La Grange Park, IL: American Nuclear Society.

Mumaw, R. J., Woods, D. D., & Eastman, M. C. (1992). *Interim report on techniques and principles for computer-based display of data* (STC Report 92-1SJ4-CHICR-R1). Pittsburgh: Westinghouse Science and Technology Center.

Pawlak, W. S., & Vicente, K. J. (1996). Inducing effective operator control through ecological interface design. *International Journal of Human-Computer Studies, 44,* 653–688.

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

Rasmussen, J. (1985). The role of hierarchical knowledge representation in decision making and system management. *IEEE Transactions on Systems, Man and Cybernetics, SMC-15,* 234–243.

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

Vicente, K. J., & Burns, C. M. (1995). *A field study of operator cognitive monitoring at Pickering nuclear generating station-B* (Tech. Report CEL 95-04). Toronto: University of Toronto, Cognitive Engineering Laboratory.

Vicente, K. J., Christoffersen, K., & Pereklita, A. (1995). Supporting operator problem solving through ecological interface design. *IEEE Transactions on Systems, Man and Cybernetics, SMC-25,* 529–545.

Vicente, K. J., & Pawlak, W. S. (1994). *Cognitive work analysis of the DURESS II system* (Tech. Report CEL 94-03). Toronto: University of Toronto, Cognitive Engineering Laboratory.

Vicente, K. J., & Rasmussen, J. (1990). The ecology of human-machine systems: II. Mediating "direct perception" in complex work domains. *Ecological Psychology, 2,* 207–249.

Vicente, K. J., & Rasmussen, J. (1992). Ecological interface design: Theoretical foundations. *IEEE Transactions on Systems, Man and Cybernetics, SMC-22,* 589–606.

Watanabe, O., Takaura, K., Fujita, Y., & Hayashi, Y. (1995). Evaluation of ecological interface design. In Y. Anzai, K. Ogawa, & H. Mori (Eds.), *Symbiosis of human and artifact* (pp. 977–982). Amsterdam: Elsevier.

Wickens, C. D. (1992). Virtual reality and education. In *Proceedings of the 1992 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 842–847). Piscataway, NJ: IEEE.

Wickens, C. D., & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors, 37,* 473–494.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Klaus Christoffersen received his M.A.Sc. in industrial engineering from the University of Toronto in 1996 and is a Ph.D. student in the Ohio State University Department of Industrial, Welding, and Systems Engineering.

Christopher N. Hunter received his M.A.Sc. in industrial engineering from the University of Toronto in 1995. He is currently pursuing a law degree at the University of Victoria.

Kim J. Vicente received his Ph.D. in mechanical engineering from the University of Illinois at Urbana-Champaign in 1991. He is an assistant professor of mechanical and industrial engineering and of biomedical engineering at the University of Toronto and director of the university's Cognitive Engineering Laboratory.

Date received: January 17, 1995
Date accepted: March 11, 1996