Cognitive Engineering Laboratory

# Research on the Characteristics of Long-Term Adaptation (II)

**Xinyao Yu, Gerard Torenvliet, & Kim J. Vicente**

**CEL 98-06, Volume I**

**Final Contract Report**
**September, 1998**

**Prepared for**
**Japan Atomic Energy Research Institute**

Cognitive Engineering Laboratory • University of Toronto • Department of Mechanical & Industrial Engineering
5 King's College Rd. Toronto, Ontario, Canada M5S 3G8
Phone: (416) 978-7399 Fax: (416) 978-3453
Email: benfica@mie.utoronto.ca URL: www.mie.utoronto.ca/labs/cel/

# Cognitive Engineering Laboratory

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

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

**Current CEL Research Topics**

CEL has been funded by Atomic Energy Control Board of Canada, AECL Research, Alias|Wavefront, Asea Brown Boveri Corporate Research - Heidelberg, Defence and Civil Institute for Environmental Medicine, Honeywell Technology Center, Japan Atomic Energy Research Institute, Natural Sciences and Engineering Research Council of Canada, Rotoflex International, and Westinghouse Science & Technology Center. CEL also has collaborations and close contacts with the Mitsubishi Heavy Industries and Toshiba Nuclear Energy Laboratory. Recent CEL projects include:

- Studying the interaction between interface design and adaptation in process control systems.
- Understanding control strategy differences between people of various levels of expertise within the context of process control systems.
- Developing safer and more efficient interfaces for computer-based medical devices.
- Designing novel computer interfaces to display the status of aircraft engineering systems.
- Developing and evaluating advanced user interfaces (in particular, transparent UI tools) for 3-D modelling, animation and painting systems.

**CEL Technical Reports**

For more information about CEL, CEL technical reports, or graduate school at the University of Toronto, please contact Dr. Kim J. Vicente at the address printed on the front of this technical report

## ABSTRACT

This final contract report describes the findings from the second year of a two-year research program investigating the characteristics of long-term operator adaptation in nuclear power plants (NPPs). The report consists of two volumes. This document, volume 1, describes the results of analyses of data from tuning and fault trials in a 6-month longitudinal study of operator adaptation using the novel measures of adaptation that were developed in last year's contract. While last year's research focused on analysing performance for the startup task during normal trials only, this research focused on normal tuning trials and fault management trials. Since the tuning tasks occurred between startup and shutdown tasks, the results were contaminated by the fact that different subjects experienced different system states. Nevertheless, the results showed that the overall performance of all subjects improved over the course of experiment, and that the Ecological Interface Design (EID) group exhibited more consistent behaviour. The analysis of fault management trials provided much more insights about subjects' knowledge of the system functions. It showed that the EID interface supports fault detection, diagnosis, and compensation. The performance of the EID subjects generally was also better than that of the subjects using the P interface. Finally, volume 1 closes by providing a theoretical integration of this year's findings and those obtained in last year's project on long-term operator adaptation. These findings have implications for the design and evaluation of human machine systems in NPPs.

## ACKNOWLEDGEMENTS

This research project was sponsored by a contract from the Japan Atomic Energy Research Institute (Dr. Fumiya Tanabe, Contract Monitor), as well as research and equipment grants from the Natural Sciences and Engineering Research Council of Canada. We would like to thank Dr. Tanabe for his contributions and help.

# TABLE OF CONTENTS

## TABLE OF FIGURES

## TABLE OF TABLES

## OVERVIEW

A predictive model of human cognitive behaviour, which includes the mental strategies used in emergency situations in nuclear power plants (NPPs), is needed in the design and evaluation of human-machine systems. To achieve this objective, a profound understanding of the characteristics of human operators' long-term adaptation to the major behaviour shaping constraints in complex systems is essential. This year's projects builds on the work conducted in last year's project for JAERI by further investigating the usefulness of novel measures of operator adaptation. The results obtained will be useful for the development of a model of human operator cognitive behaviour and of criteria for design and evaluation of human-machine systems.

The work conducted during this project is documented in two volumes. This document, volume 1, describes the results of analyses of data from tuning and fault trials in a 6-month longitudinal study of operator adaptation using the novel measures of adaptation that were developed in last year's contract. In addition, this volume also provides a theoretical integration of this year's findings and those obtained in last year's project on long-term operator adaptation. Volume 2 describes the results of a systematic literature review that was conducted to understand the limitations of well-known behavioural statistical analysis techniques, such as null hypothesis significance testing and analysis of variance. In addition, volume 2 also describes a number of lesser-known statistical analysis techniques that were identified to address the limitations of the more traditional techniques, and shows how these techniques were applied to data from previous experiments conducted for JAERI.

## 1. INTRODUCTION

In the contract report of last year's project, we presented a literature review of approaches to the measurement of operator adaptation. The results of that review indicated that information theory, and especially dynamic system theory, seemed to be promising measurement approaches. We have applied these measures to the startup period of normal trials and reported the results in last year's contract report (Yu et al., 1997a). These measures provided much insight into the long-term adaptation of subjects using different interfaces to control the same system.

This year, we studied the tuning period of normal trials and fault management trials. These control tasks were designed to introduce some sort of perturbation into the system while the subjects were performing normal startup control tasks. In the tuning period, the subjects were

required to satisfy new demand goals after achieving the startup goals. The data from these tasks should help us understand how subjects adapt to goal perturbations. During the fault trials, different faults were introduced into the system which subjects were asked to detect and compensate for to prevent the system from 'blowing up'. The data from these tasks should help us understand how subjects adapt to deal with perturbations of system structure.

As we pointed out in last year's contract report (Yu et al., 1997a), there were a number of important reasons for using the data set collected with DURESS II during earlier research for JAERI (Christoffersen et al., 1994). First, this experiment was conducted on a quasi-daily basis for an unprecedented period of six months. Thus, it provides a rare opportunity to investigate adaptation over a very long period compared to that which is usual for laboratory studies. Second, this experiment manipulated the interface that participants were given. Half of the subjects received an interface designed according to the principles of the EID which contained physical and functional information (P+F), whereas the other half received a more traditional interface containing physical information (P) alone. Thus, this data set allows us to investigate whether the proposed adaptation measures are sensitive, not just to changes with experience, but also to differences across interfaces as well as the interaction between interface and experience.

Third, because of its longitudinal nature, this experiment only included six subjects, three in each interface group. While this resulted in a decrease in statistical power, it has the advantage of making the data analysis with different adaptation measures more manageable. Because there are relatively few subjects, it is easier to examine each individually to see how our understanding of their behaviour varies as a function of the measures we use to look at the data.

Our experience from the research performed under last year's contract has further justified this choice. A number of measures were identified and many of them were shown to be quite illustrative and informative when applied to this data set. For instance, the measures developed based on the abstraction hierarchy and dynamical system analysis demonstrated that the most proficient subjects in the two interface groups used quite different control strategies to control DURESS II. It would be interesting to see if these measures are also informative for the analysis of tuning and fault management trials.

The remainder of this volume is organised as follows. First, we will briefly discuss a number of the beneficial features of the data analysis tool developed last year that facilitated the present research. Second, we will present several analyses of subjects' performance and

adaptation in the tuning period using the measures identified. Third, we will describe the analysis results for fault management trials. Fourth, we will try to integrate all of our empirical findings on long term operator adaptation into a coherent, unified theoretical framework. Finally, we will conclude by describing the contributions of this research, its implications for the design and evaluation of human-machine systems in NPPs, our suggestion for future experiments, the limitations of the current research, and a set of future research topics.

## 2. DATA ANALYSIS TOOL

Due to huge size of the data set, computation for the analyses described in this report was extremely heavy. Thus it was necessary to develop an analysis tool specifically for the data set. In the first project year, the software tool ADAPT (Adaptation Data Analysis & Processing Tool) was developed (Yu et al., 1997b). At that time, only the startup portion of normal trials were analysed. Therefore, the measures included in the ADAPT were defined according to the features of startup tasks. As this was the case, it is not surprising that some modifications were required to fit ADAPT to tuning and fault management trials. Fortunately, the design of ADAPT made these changes relatively simple to implement. First, the architecture of ADAPT is open so that new measures can be added with minimal effort. Second, the code of ADAPT is modular and easy to maintain. As before, the tool served several important roles, including data pre-processing, and computation of various measures. It would have been practically impossible to have manipulated the amount of data used in this study without ADAPT.

## 3. ANALYSIS OF TUNING TASKS

During the experiment, subjects were asked to perform four different types of control tasks: startup, tuning, shutdown and fault management. In the tuning task, the subjects needed to bring the system from an on-line, steady state initial condition (which resulted from a successful task of startup) to a pair of new steady states composed of a different demand goal for each of the reservoirs. These tuning tasks were introduced after subjects had completed 44 startup trials. Therefore, the subjects had considerable experience with controlling DURESS II, and can be assumed to have learned something about the dynamics of the system, the time delay of the components, and even some structural constraints.

The research aim of introducing the tuning tasks in the experiment was to investigate subjects' adaptation to the perturbation of goal variables. Our previous analysis of normal trials (Yu et al., 1997a) revealed how subjects' performance and ability to adapt changed with

experience, and to what extent this was caused by the interface used. Studying control performance under perturbations allows us to derive deeper insights into operators' knowledge structure and cognitive competencies than studying normal trials alone, and allows us to identify operators who had a deeper knowledge of system structure and functions. This rationale is consistent with findings in control theory, where it is well known that stimulating (i.e., perturbing) a system causes hidden behavioural modes of the system to be excited and thereby revealed. This is consistent with the so called "sufficiently rich" characteristic of input control signals in adaptive control technology (see Narenda & Annaswamy, 1989).

3.1 Performance Analysis

In this section, we will present the results of a performance analysis that applied a number of measures to the data from the tuning tasks. First, the analysis will quantitatively describe the performance of each subject and will reveal which subjects were the most proficient at the tuning trials, thereby demonstrating how the interface design impacts individual performance. Second, this analysis will illustrate in detail how, and to what degree, the change of the goal demands affected the subjects' performance. Third, from these measures, we can see how performance changed with experience, that is, which subjects learned more quickly. Fourth, these findings will provide some background information that will be useful in interpreting the various adaptation measures that will be discussed in the sections to follow.

There are two goals for each reservoir in DURESS II. So, with respect to goals, the dynamic behaviour of each reservoir can be seen as a two dimensional state space. For each trial, the behaviour of each reservoir can be characterised by a trajectory in this space (see Howie & Vicente, 1998; Sanderson, Verhage, & Fuld, 1989). Instead of the initial conditions of the startup trials where all components were set to 'off' or 0, a trajectory for a tuning task starts from a non-zero point representing the steady state of the startup task. It is also important to understand that tuning trials involved only changes in output flowrates, and not in the temperature goals. This suggests that the analysis could be conducted within a reduced dimension space, that considers the dynamics of output demands and temperature goals separately, in contrast to what we did with the startup tasks (see Yu et al., 1997a).

**Time to complete the tuning task**

Before discussing fine measures of performance, we will first consider the time taken by subjects to complete the tuning tasks. In last year's project, we applied a similar measure to the

startup tasks. This measure indicated the performance differences between subjects, with AV and TL being the fastest performers in the P+F and P groups, respectively. AS was the slowest in the P+F group, and while ML was the slowest in the P group early on, WL was the slowest at the end of experiment. This measure also indicated the performance difference between interface groups. P+F subjects had a higher level of consistency in their performance in comparison with the P subjects (Yu et al., 1997a). We can expect that similar findings can be made with respect to the tuning trials because there are some similarities between startup and tuning tasks. They both start from a steady state and go toward another steady state. The difference is that the initial conditions for tuning tasks are non-zero and different for each trial of each subject. (This feature of tuning tasks made it difficult to compare results across subjects; we will return to this issue later).

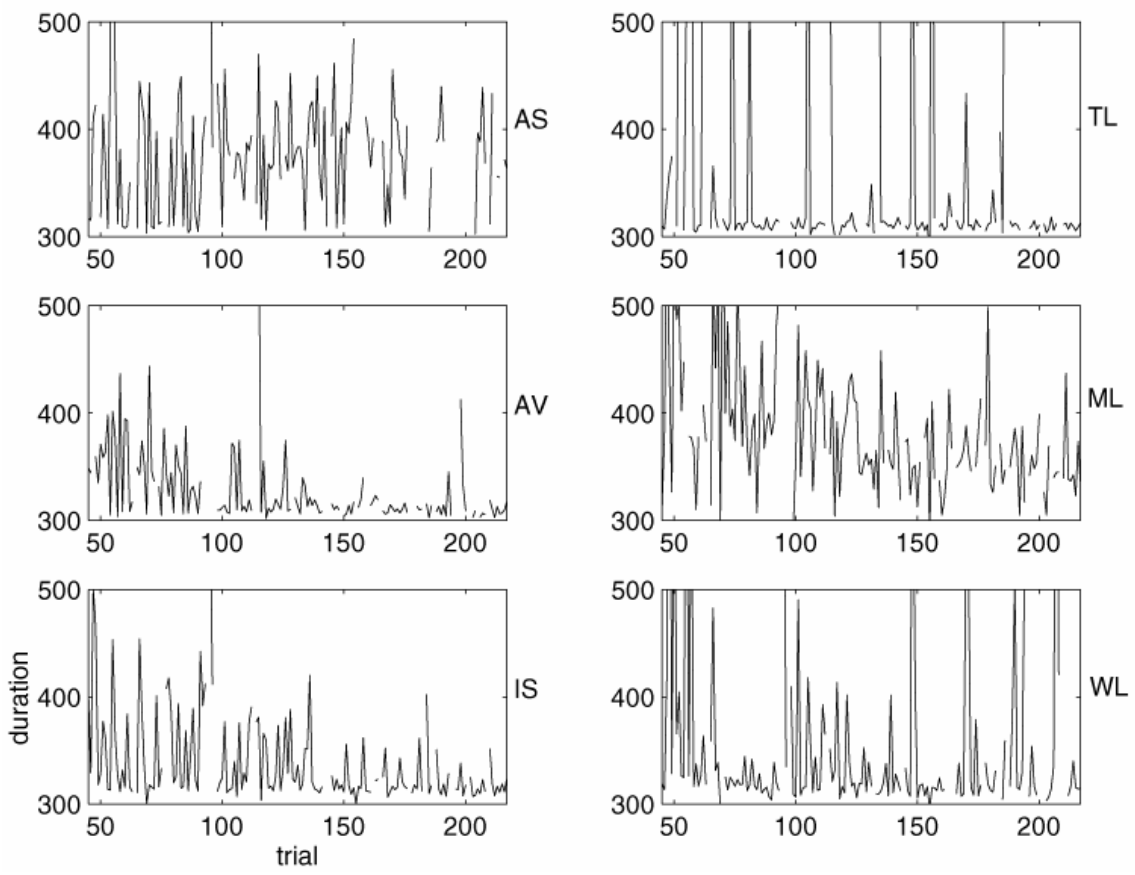


**Figure 1:** Time to complete the tuning tasks for all subjects

Figure 1 shows the results of time to complete the tuning tasks over trials for each subject (note that the tuning tasks were introduced at trial 45). As expected, two findings stand out:

- Beginning with within-group comparisons, AV and IS were clearly better than AS in the P+F group, and TL was the best within the P group. AS and ML were the slowest in their respective groups.
- As for between-group comparisons, the average time to complete the tuning tasks for P+F subjects was shorter than that of P subjects. This indicates that the P+F subjects could, on average, finish their task faster than P subjects.

In terms of performance consistency, it is notable that the P+F subjects were at a higher level of consistency than the P subjects.

Both AV and IS exhibited considerable learning with experience. The time they used to complete the tuning tasks was reduced over the course of experiment. ML also displayed some learning, but this was minimal. TL seemed to improve somewhat in terms of performance consistency. The other subjects failed to have such characteristics. Similar results were found by Christoffersen et al. (1994).

In summary, the subjects who used the P+F interface took less time to finish the tuning tasks, and exhibited more consistent behaviour. These findings are quite similar to the results for the startup task.

**Perturbation on water temperature**

For each reservoir, the water output flow rate and temperature are physically coupled. A change in output rate will change the water temperature of a reservoir, all things considered equal. Larger flow rates cause more energy to be taken away by the water. In the tuning period, when the setpoints of water output flowrates changed, the subjects had to adjust the output valves accordingly. Different strategies used to open the valves could cause different perturbations in the water temperature of each reservoir.

We would expect that more efficient subjects would have less perturbations in water temperature, and if there are any, they should disappear faster than for the less proficient subjects. This can be demonstrated by plotting reservoir temperature over time after a perturbation. An example of one of these plots is shown in Figure 2, where the results of trial 50 completed by AV are depicted. The curves were obtained by subtracting the real water temperatures from the values of temperature goals for each reservoir. The curves quite clearly indicate that the temperatures were perturbed. But without some basis for comparison, we cannot ascertain the relative magnitude of this effect.

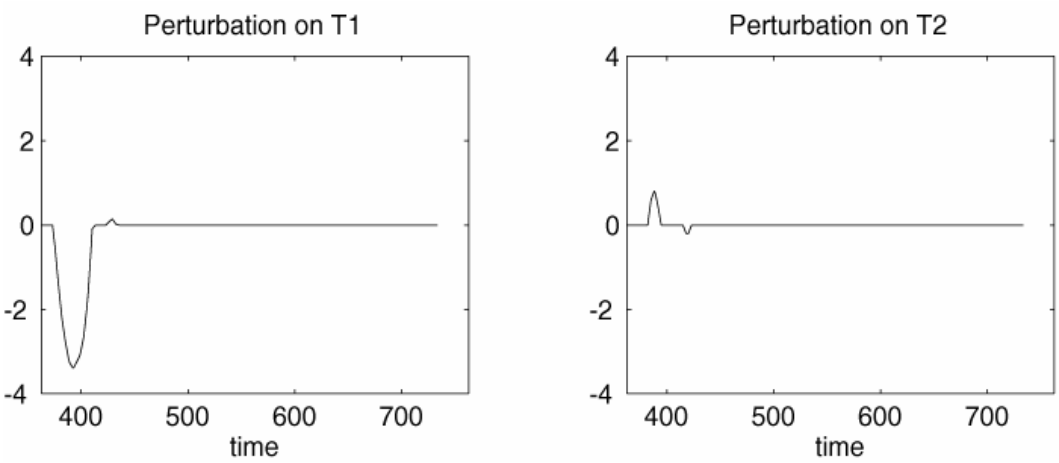


**Figure 2:** Perturbation on temperature in each reservoir for AV at trial 50

**Normalised error of perturbed temperature**

For each individual trial of each subject, plots of temperature perturbation provide important insight into how water temperatures were disturbed by the changes in output water flows. This measure has two disadvantages, however. First, as we can see, it is inefficient because one plot must be constructed for each trial of each subject. Second, from an analysis of each trial alone, it is difficult to determine how severely the temperatures were perturbed. Although we can make between-subjects comparisons for each point of each trial, it is difficult to see how the subjects improved their performance in terms of this measure. Thus, a more efficient measure -- the area under the curves of perturbation of temperature -- was calculated for each trial. To make this measure comparable across trials, the area was normalised with the change of demand goals and the temperature goals. With this measure, a large area indicates a long transient period and/or large deviation from the setpoints. Thus we would expect that more proficient subjects would exhibit smaller areas, and that the areas of all subjects should decrease with experience.

Figures 3 and 4 show the results for the P+F group and P group, respectively. In the P+F group, AV and IS did much better than AS who did not do well consistently. In the P group, TL and WL did better than ML who generally had a larger area. Another interesting point we can see is that all subjects except AS and TL show signs of improvement over the course of the experiment.

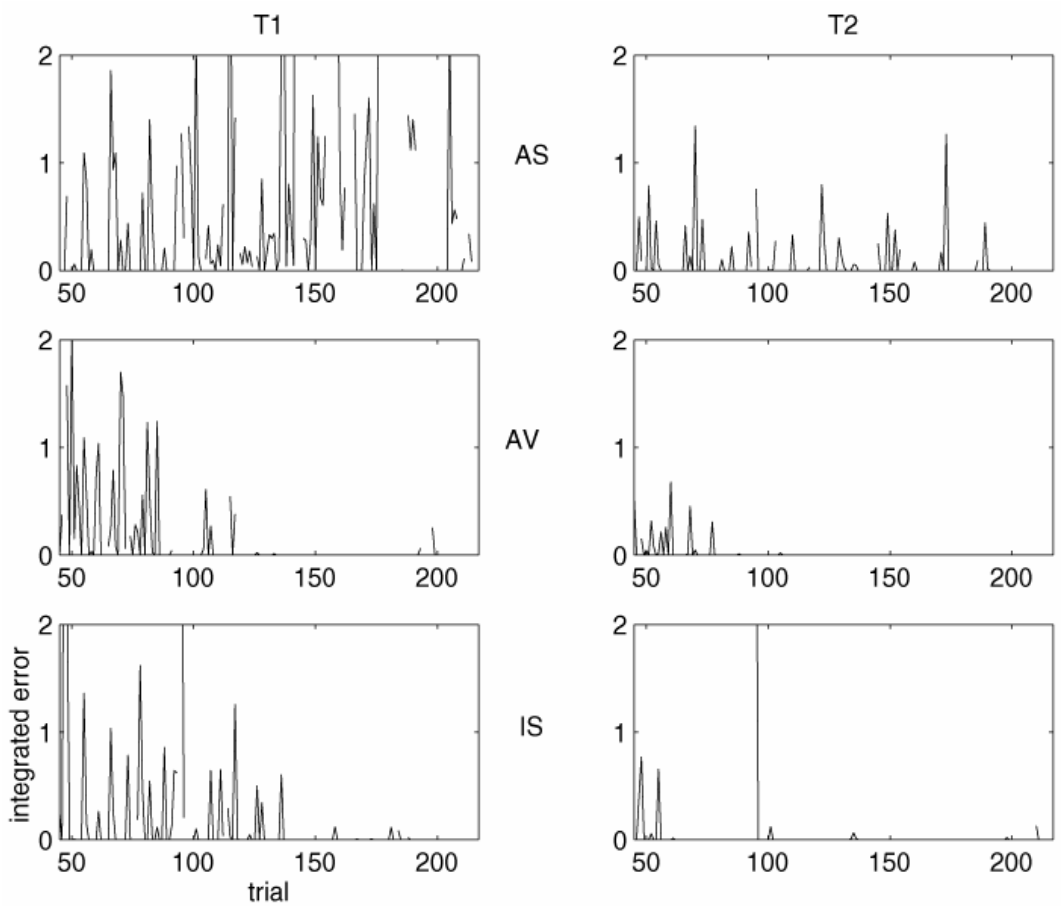


**Figure 3:** Normalised area of perturbation on temperatures for P+F subjects

In the analysis of startup tasks last year, we divided the output responses into three phases. In each of the phases, subjects faced different challenges. For example, during the very early part of a startup trial, the goal is to turn on the required components in order to get the system heading toward the direction of goal state. Following this phase, the subjects should attempt to get the system under control by avoiding large overshoots or undershoots to reach the goal region as quickly as possible. In the last phase, subjects need to maintain the system within the goal region, which is generally accomplished by keeping the system as stable as possible. Unlike the startup tasks, the tuning tasks involved no change in temperature goals (Christoffersen et al., 1994). Instead of having to achieve new temperature goals, the goal in the tuning task is to prevent the reservoir temperature from being perturbed, and to keep it as stable as possible. In what follows, we shall present the results of four more measures. These measures are derived based on the idea of phase division (Yu et al., 1997a). Compared to the measures of perturbation

of temperature and area under the curves of perturbation of temperatures, these four measures are more fine-grained. These measures will show how severely the temperatures were perturbed, how quickly subjects recovered, and how stable the system eventually became with respect to the temperature goal.

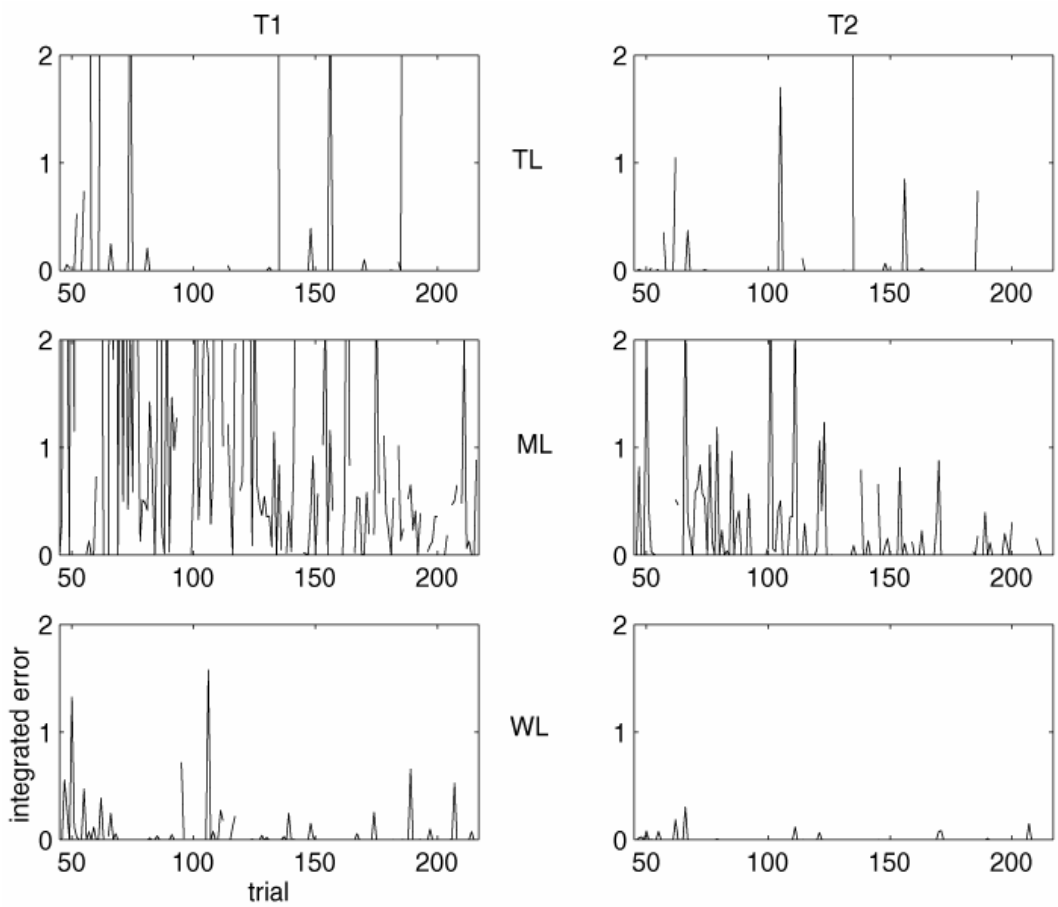


**Figure 4:** Normalised area of perturbation on temperatures for P subjects

**Largest deviation of temperature**

As its name suggests, this measure calculates the largest deviation of temperatures from their setpoints. Since the changes of demand goals and the setpoint of temperature goals must be taken into consideration in order to make comparison across the trials, the temperature deviations were normalised by the demand and temperature goals during the tuning portion of the trial. The larger temperature deviation of the two reservoirs was taken as the largest deviation of temperature of the system.

The results of this analysis are shown in Figure 5. Two findings emerge. First, AV and IS were better than AS in the P+F group, and TL and WL were better than ML in the P group. Second, all the subjects except AS show signs of improvement, while AS failed to have such a tendency. Between the people with best performance in the two groups, it seems that the largest deviations of the P+F subjects (AV and IS) were less than those of P subjects (TL and WL) during the latter half of the experiment, and that subjects using the P+F interface were more consistent.

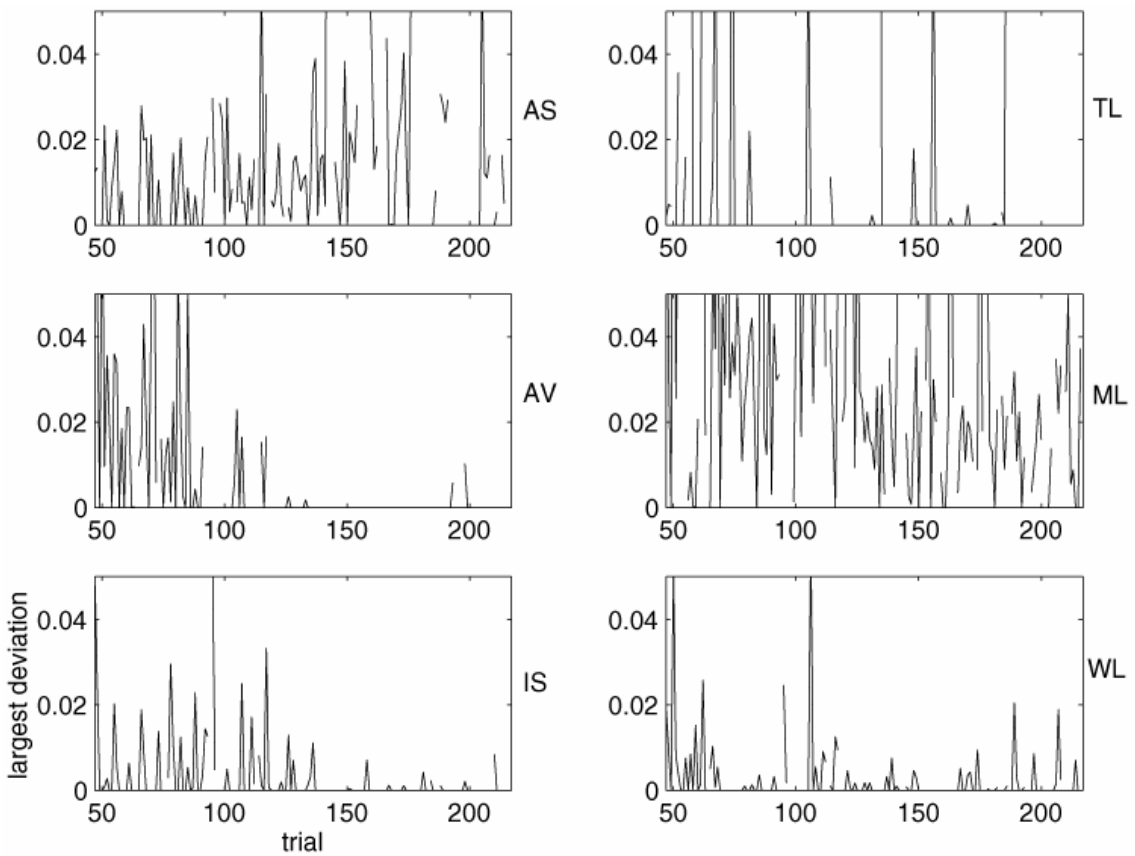


**Figure 5:** Largest deviation of temperature for all subjects

**Recover time of temperature**

The measure of largest deviation of temperature indicates how severely the temperatures were disturbed due to the changes in water flowrates. If there were such deviations away from the range of the temperature goals, then the subject would have to take actions to counteract them. The effort exerted by subjects to recover may depend on two factors. The first is the system state, especially the largest deviations of temperature. The second is the strategies used

by the subjects. The more severely temperatures were disturbed, the more effort would be needed to re-stabilise the system. Of course, a good strategy can reduce effort as well. We expect that people who had better performance would take less time to recover temperature goals.

Figure 6 shows the results of an analysis of the recover time of temperatures over the course of experiment. In order to make a comparison between trials, we again normalised the data. The actual time that the subjects used to stabilise the temperature was normalised by the values of demand changes and temperature goals. It can be seen that the findings are quite consistent with those we achieved using the measures discussed above.

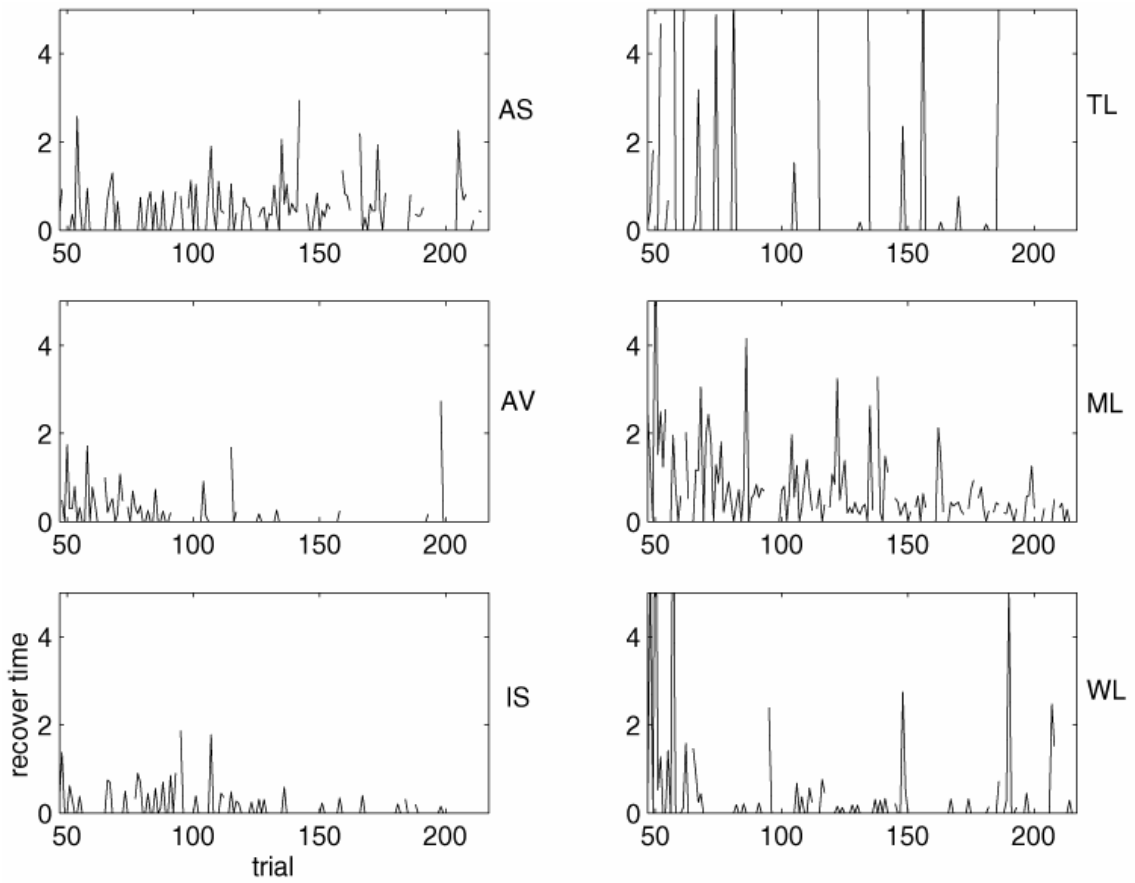


**Figure 6:** Recover time of temperature for all subjects

**Time to goal boundaries**

Just as in last year's project, the time to goal boundaries (TTGB) of temperature was calculated to measure the stability of the last five minutes' period in a task, this time of the tuning tasks. Since the system is already in the goal boundaries during this period, the goal boundaries can be viewed as repellors that must be avoided. Thus, we can quantify performance

by measuring the time to reach the goal boundaries, which can be considered as a measure of boundary avoidance. Assuming that subjects aim to avoid the goal boundaries, the higher a subjects' TTGB is at any point in time, the better. The time to goal boundaries can be plotted over time for each trial for each subject. An example is shown in Figure 7 for AV at trial 50.

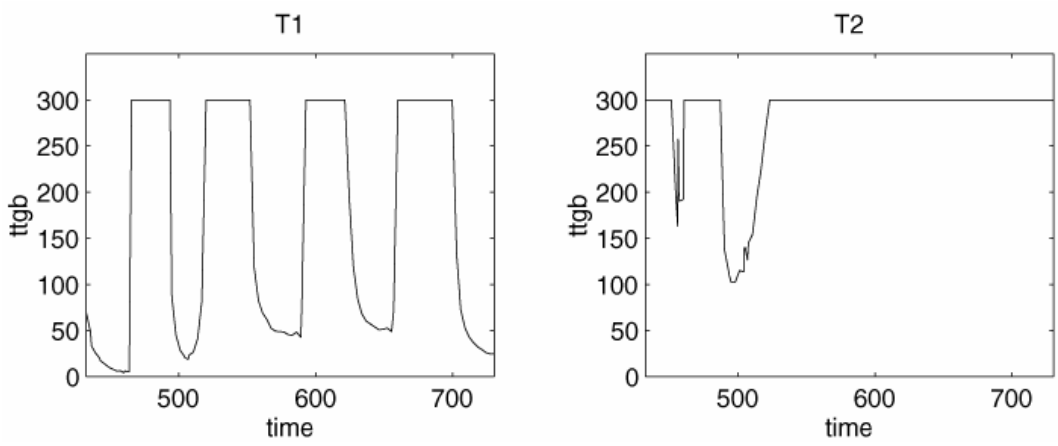


**Figure 7:** Time to goal boundaries for AV at trial 50

**Area under curves of time to goal boundaries of temperature**

While TTGB graphs show in detail the stability of temperature during the later period of the tuning task, they suffers from two disadvantages. First they are inefficient, since one graph displays the information for only one trial. Second, they make it difficult to compare the results across trials as the demand changes were different from trial to trial. To remedy these shortcomings, we can summarise subjects' performance for the entire experiment by taking the area under the curves of time to goal boundaries for each trial (ATTGB). We again used the normalisation technique for this measure. The area was normalised with the demand changes and temperature goals. Since the best case is to keep the goal boundaries at maximum (300 seconds) for the last 300 seconds of a trial, the area was also normalised by the maximum possible area, $300^2$. Thus, a normalised area of 1 will indicates perfectly stable temperature, while any value less than 1 indicates decreasing stability. We expect that proficient subjects will have higher normalised areas, increasing with experience for each subject.

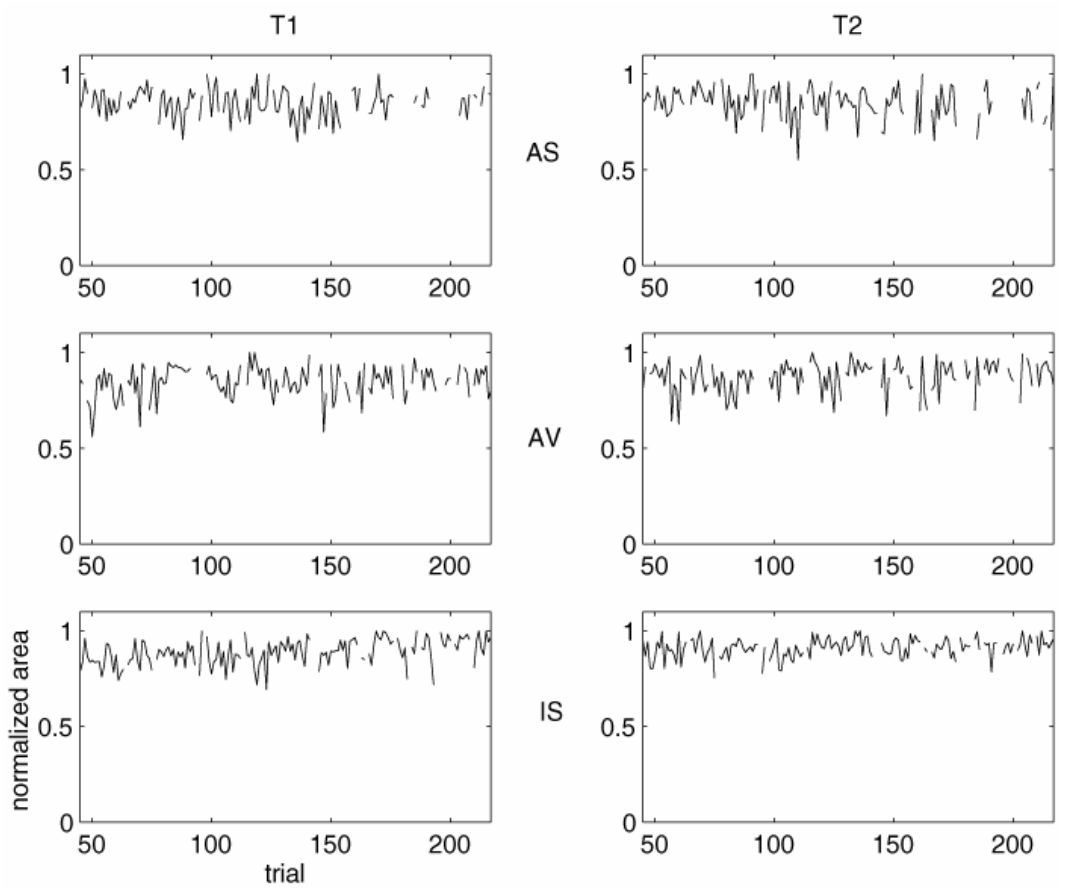


**Figure 8:** Normalised area under the curves of time to goal boundaries of temperatures for P+F subjects

The results are shown in Figure 8 for the P+F group and Figure 9 for the P group. In the P+F group, AV and IS both exhibit more stability than AS. In the P group, TL was more stable than WL. In addition, there were some signs that some of the subjects (such as AV and ML) improved with experience. In general, these findings are not very strong, and the results from this measure are not very informative. Similarly uninformative results were achieved in the analysis of startup tasks (Yu et al., 1997a)

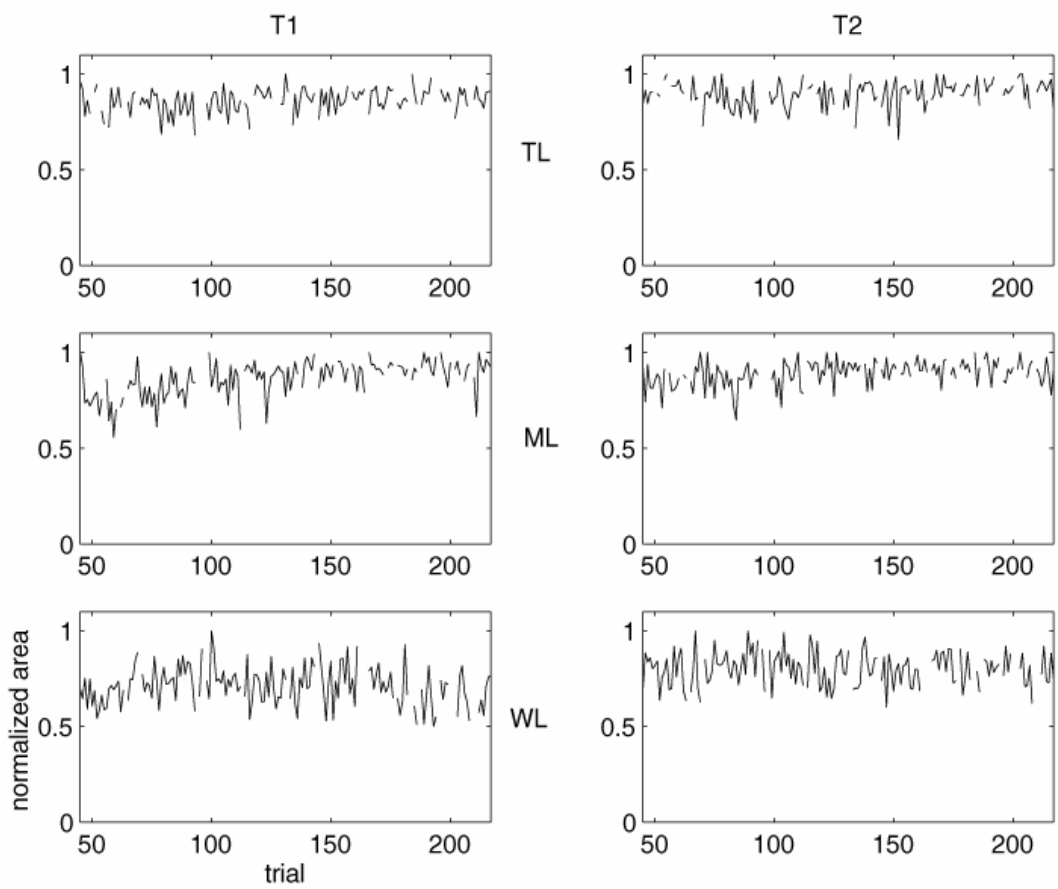


**Figure 9:** Normalised area under the curves of time to goal boundaries of temperatures for P subject

**Summary**

We have presented the results of overall performance in terms of time to complete the tuning tasks, as well as dynamic and steady state performance in terms of temperature goal for all subjects. Within groups, AV and IS were the best in the P+F group, and TL was the best in the P group. The proficient subjects took less time to finish the task, experienced less perturbation of temperature, and were able to regain system stability more quickly. There were also some signs of improvement for some of the subjects over the course of experiment, especially AV and IS in the P+F group.

In the rest of this section, we will present the results of analysing the system performance in terms of output flowrates. By carefully examining the data set, we found that for most trials, the subjects seldom had any overshoot in the demand outputs in the tuning tasks. As a result, there is no oscillation period as was defined previously for the demand output (Yu et al., 1997a).

We used seven measures to characterise system performance. For overall performance, we used the error of output water flowrates in response of the changes in setpoints, and the area under the curves of error responses. For the first part of the dynamic period, we used rise time (RT), time to contact (TTC), and area under the curves of time to contact (ATTC). For the rest of the dynamic period, we used TTGB and ATTGB. For the sake of briefness, we shall only describe here the results of the integrated measures that summarise the results for each subject across trials.

**Area under the curves of error responses of output flowrates**

Figure 10 shows the results of this analysis. Before making any comparisons, it is worthwhile to note that for the experiment from which we derived the dataset for this study, DURESS II's output valves were configured to a time constant of five seconds. If subjects were to respond to changes in demand goals immediately as they happened, the area under the curves of error response should be 5 units. If there was some delay in response or if the subject had some difficulty in regulating the output flow rates, the area would be greater than that 5 units.

A first impression from Figure 10 is that ML consistently had the largest area under the curve of error response of all subjects. While meeting the output demand goals is not difficult by itself, changing the output flow rates has an effect on the energy balance of the reservoir. Thus, these results are consistent with the results in Figure 5, where ML was shown to have consistently large deviations from the temperature goal.

AV and TL were the best in the P+F and P group, respectively, while AV was more consistent than TL. It is interesting to note that AS, the subject with the worst performance in P+F group, had low values on this measure except during the middle portion of the experiment (about trials 120-145). From the point of view of temperature, a better way to meet the demand goals is to move the output flowrates step by step toward the setpoints, as an immediate change of outflow rates would cause more severe temperature disturbances. Therefore, lower values of this measure (as exhibited by AS) are not necessarily better. Rather, a better strategy is to strike a balance between high and low values of the area under the error responses of output flowrates. In conclusion, AV and AS had better performance in the P+F group, while TL and WL had better performance in the P group. In general, subjects in the P+F group had more consistent performance than those in the P group.

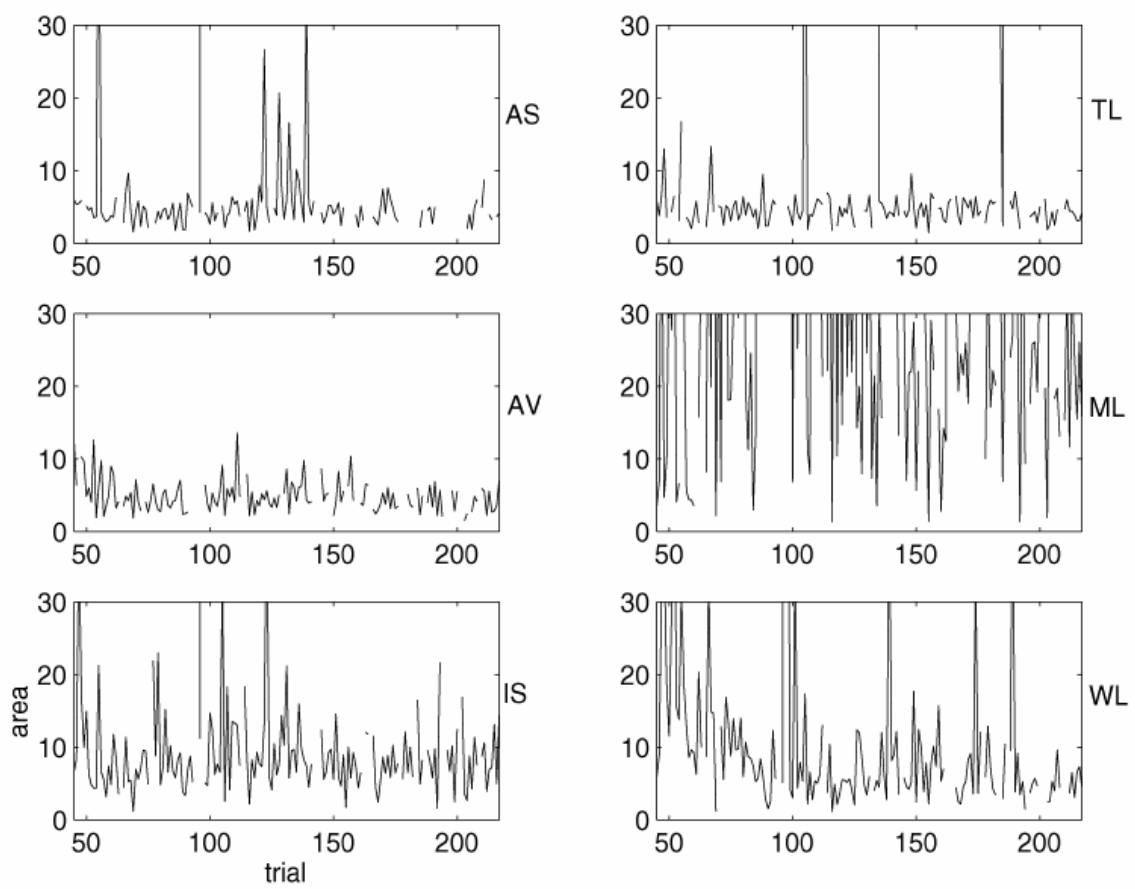


**Figure 10:** Area under the curves of error responses of output flow rates for all subjects

**Rise time of output flowrates**

The area under the curves of error responses of output flowrates depends on both the rise time and the value of error over time. Longer rise times and larger errors will result in a larger area. Since the output valves have first order dynamics, the area should be mainly determined by the time to reach the goal boundaries. Thus, we can expect that the measure of rise time of output flowrates should be consistent with the results in Figure 10. The results of this analysis are shown in Figures 11 and 12 for the P+F and P groups, respectively. Not surprisingly, it does confirm what we saw in Figure 10, indicating that ML used the output valves to regulate temperature, and so had a relatively long rise time. Again it should be noted that shorter rise times are not necessarily indicative of good performance. Rather, it is better to have a moderate rise time.

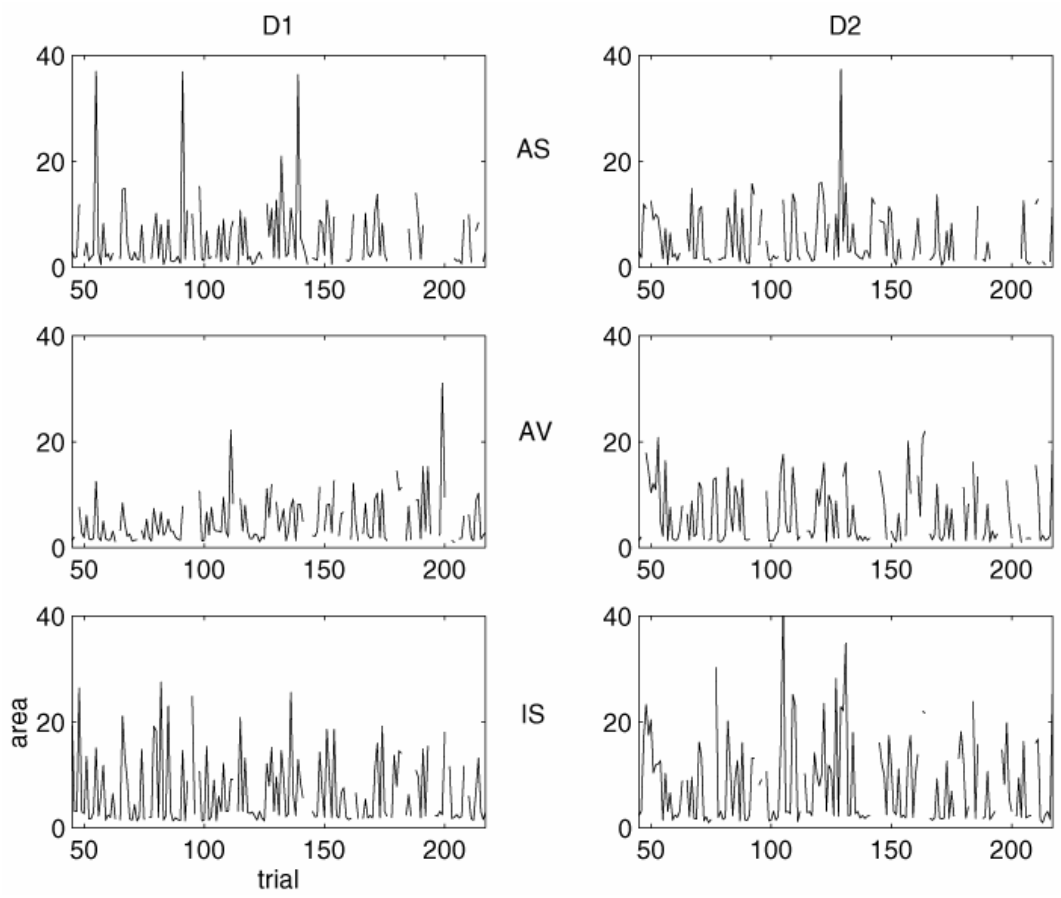


**Figure 11:** Rise time at tuning period for P+F subject

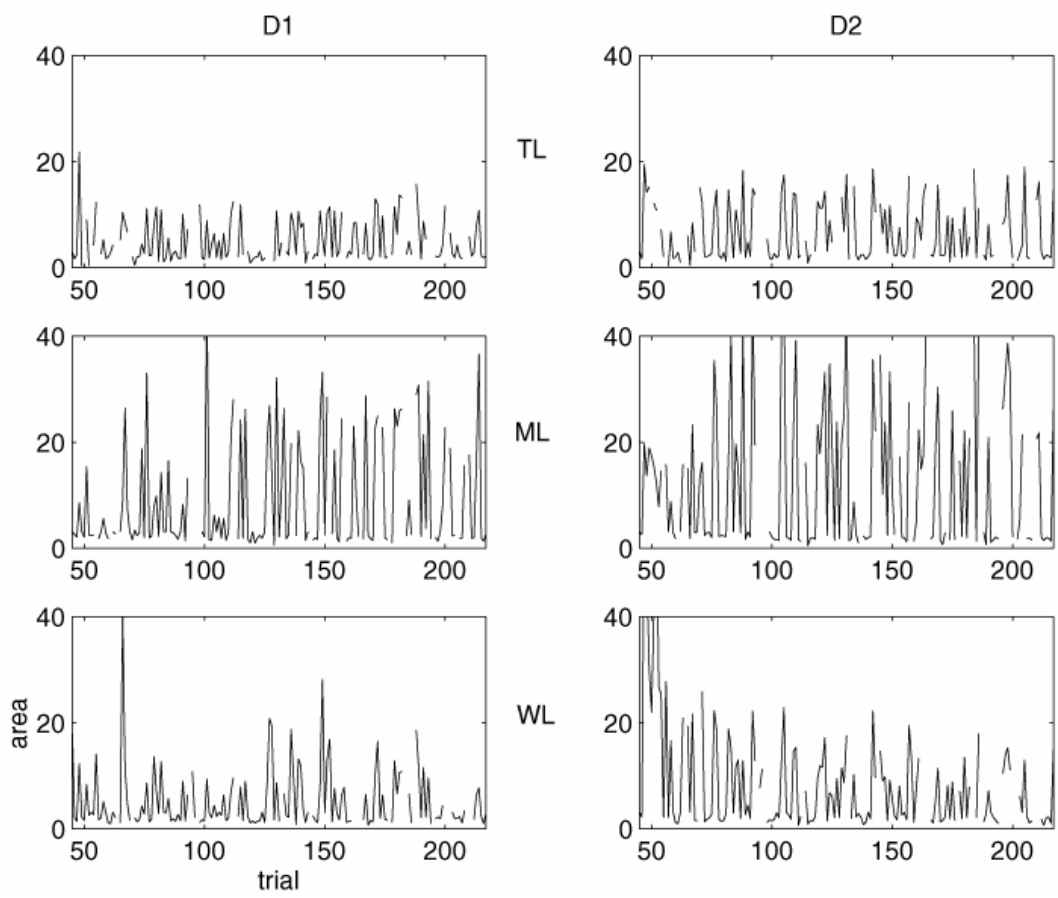


**Figure 12:** Rise time at tuning period for P subjects

**Area under the curves of time to goal boundaries for output flowrates**

Figures 15 and 16 show the results of this analysis. The method to perform this analysis can be viewed as an analysis on ATTGB for temperature (see Yu et al., 1997a). We would expect that subjects whose system control was more stable in the last 5 minutes would have values on this measure approaching 1. Unfortunately, this measure was not as informative as we expected. Since the goal of this period was to move the output only to within *coarse* boundaries, as long as the output flowrates were within these boundaries, the subject did not need to initiate any control actions, even if the TTGB was quite small. We expect that this measure would be more informative if the tolerances around the output goal boundaries were decreased.

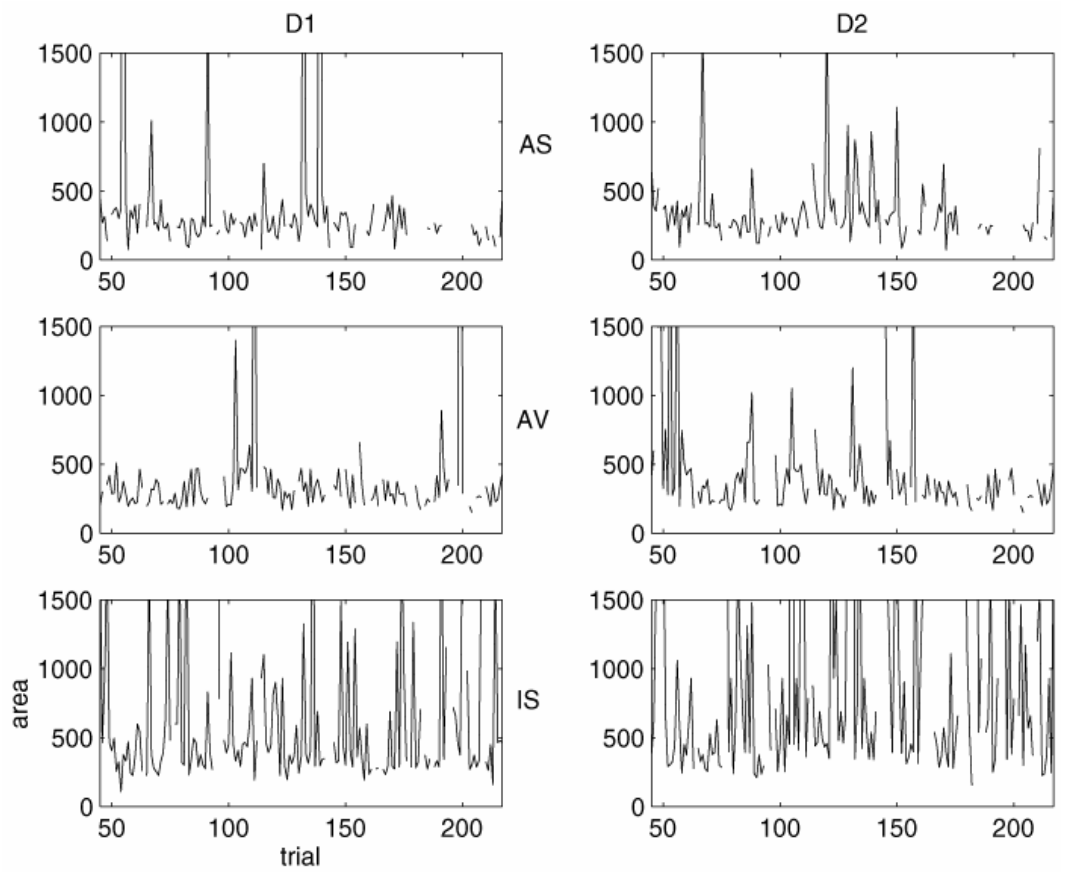


**Figure 13:** Area under the curves of time to contact the goals for output flowrates for P+F subjects

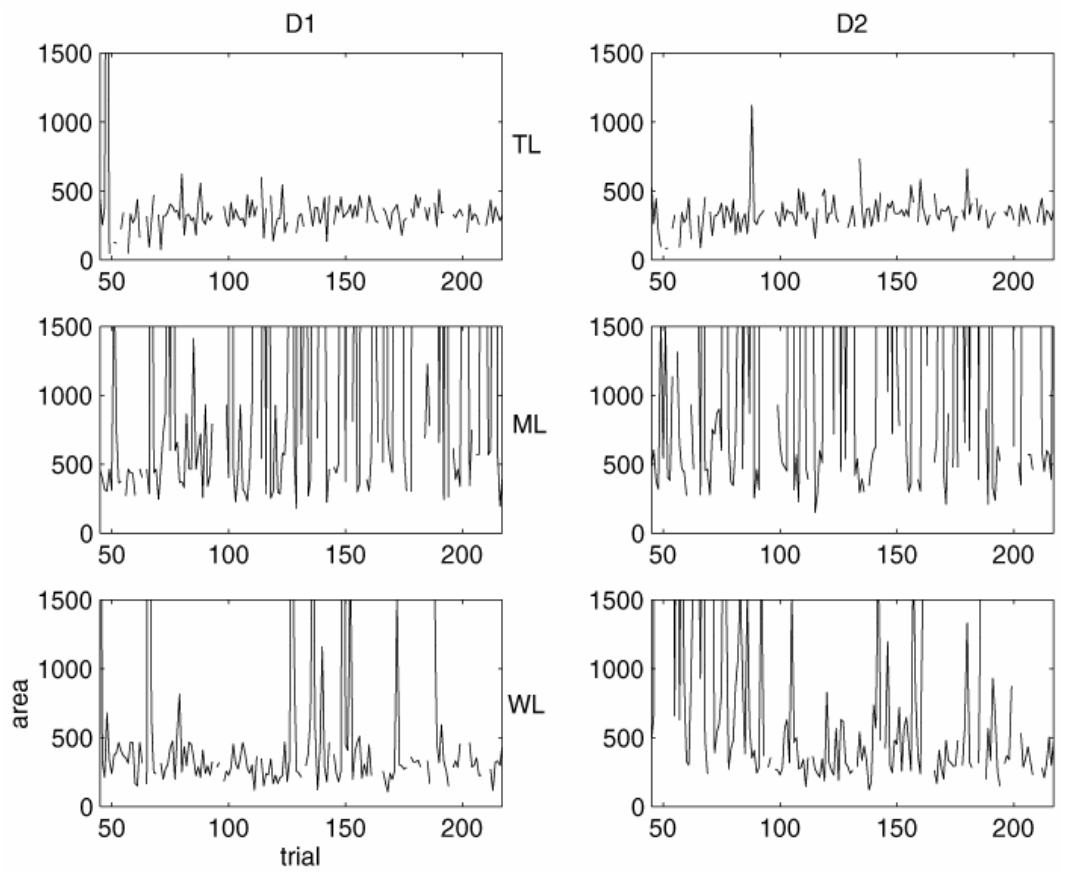


**Figure 14:** Area under the curves of time to contact the goals for output flowrates for P subjects

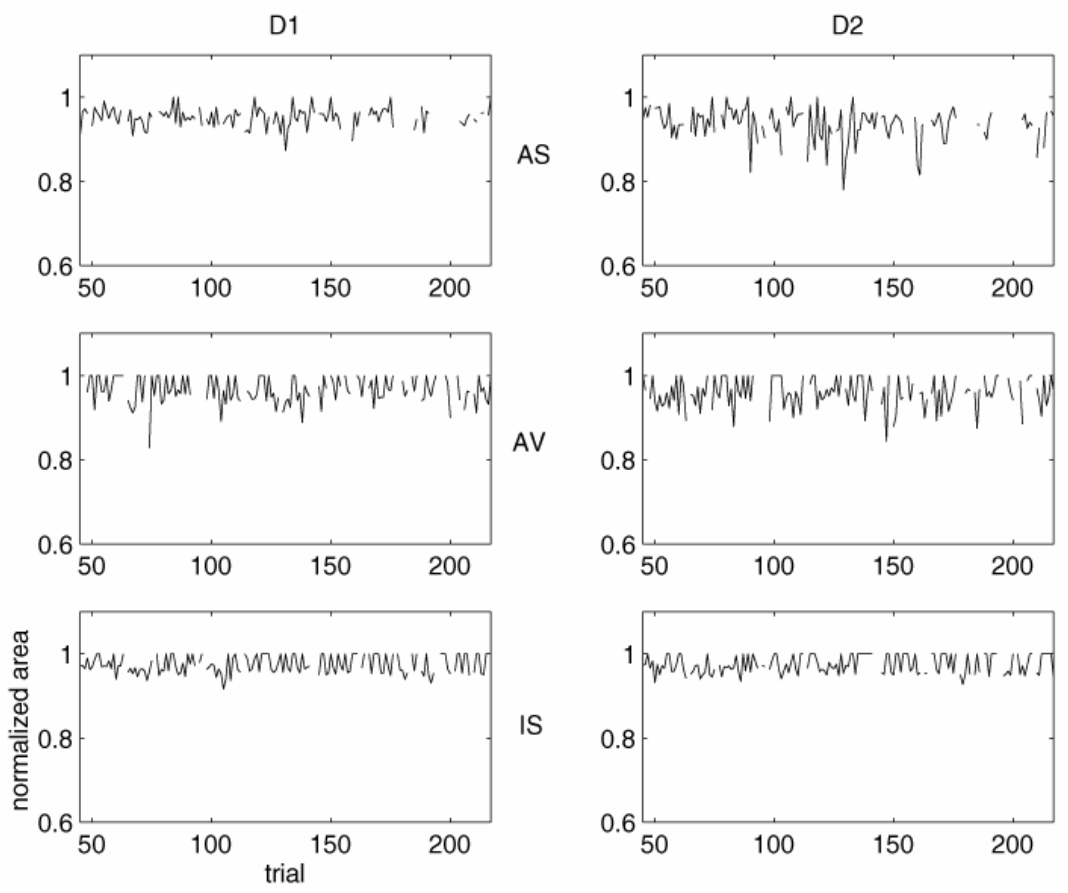


**Figure 15:** Area under the curves of time to goal boundaries for output flowrates for P+F subjects

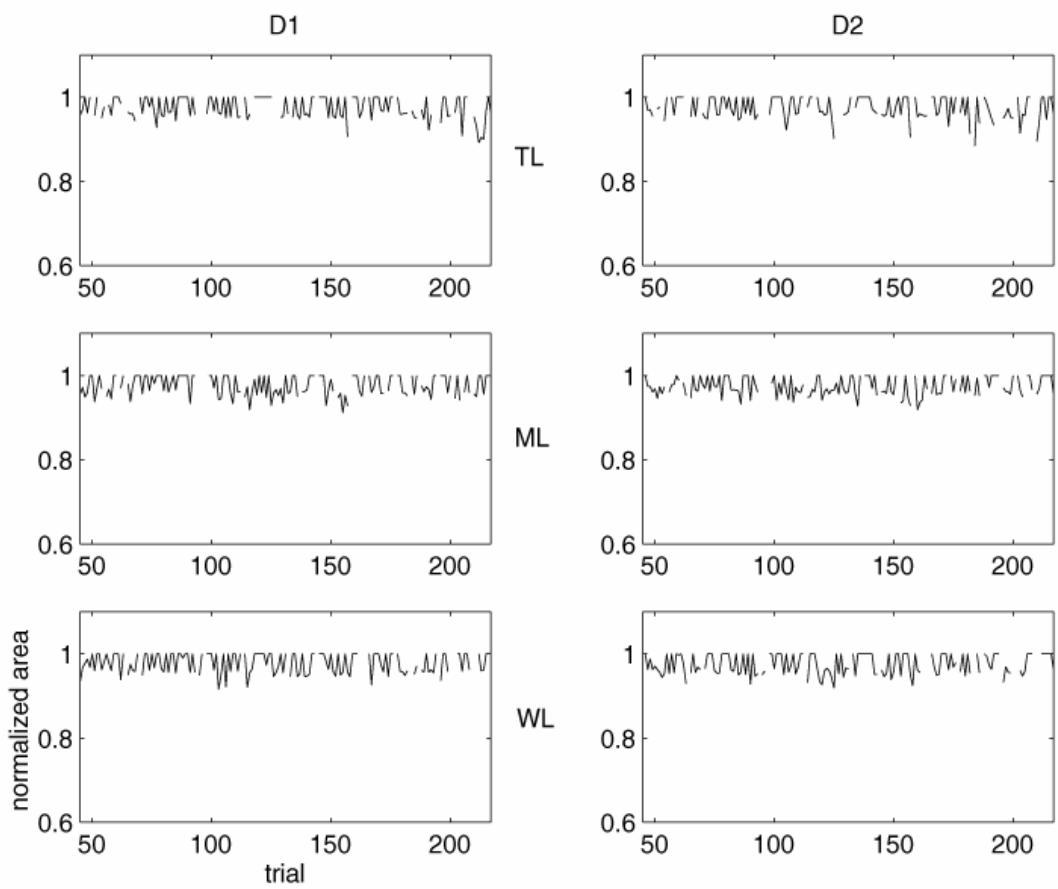


**Figure 16:** Area under the curves of time to goal boundaries for output flowrates for P subjects

**Summary**

While normal startup trials lend themselves to an analysis in a two-dimensional goal space, tuning tasks necessitate two one-dimensional goal spaces representing the temperature and output flowrate. Analyses on these goal spaces led to the following observations:

- The performance of most subjects improved over the course of experiment.
- AV and IS were the best performers in the P+F group, and they seemed to exhibit comparable performance. TL was the best performer in the P group.
- In general, the P+F group exhibited more consistent performance than the P group.

Interestingly enough, these findings are quite similar to what was found for the normal startup trials (see Yu et al., 1997a). This may indicate commonalties between the startup and tuning tasks, which both require subjects to lead the system from one stable state to another. Considering subjects' adaptation to DURESS II, the performance of the P+F group seems to improve with experience to a greater extent than that of the P group.

**Discussion**

In this section, we have presented the results of performance analyses for all subjects, using various measures of dynamic and steady state system responses. The analysis of temperature and demand goals was separated in consideration of the special features that were designed into tuning tasks in the longitudinal experiment. Several interesting findings are:

- Overall, the P+F subjects completed the tuning tasks more quickly and consistently than the P subjects.
- AV and IS exhibited comparable dynamic performance in the P group, while AS was the worst performer in this group. In the P group, TL exhibited the best dynamic performance.
- AV and IS showed significant signs of improvement in overall dynamic performance, while the rest of the subjects did not show a strong indication of learning.
- There was no clear difference in steady-state performance between subjects or interface groups.

These findings show the quantitative performance difference between subjects and interface groups, and how the subjects adapted to changing system goals. It seems that the P+F interface promoted better performance and adaptation to system goals. These findings are also serve as background information to interpret the various measures of adaptation that will be discussed in the rest of this section.

3.2 Analysis of Action Variability

Our previous analysis of action variability for normal startup trials confirmed the content of the control recipes completed by subjects describing their methods for controlling DURESS II (Yu et al., 1997a). While the analyses already described in this report focussed on the plant being controlled by the subjects, the emphasis will now shift to an analysis of action variability that emphasises the contributions of the human operator. Human control actions can be driven in various ways, including by environmental stimuli or by some internal knowledge (i.e., experience). It would be interesting to investigate how the human operators adapted in terms of their control actions, and what factors influence such adaptation.

Instead of reprinting the various control recipes here (see Christoffersen et al., 1994) , it suffices to recall the contrast between the control recipes of TL (P) and AV (P+F):

- TL, the best subject in the P group, reported setting certain components to specific values. In other words, his recipe was proceduralised.

- AV, the best subject in the P+F group, reported setting the components to suit the goal to be achieved. In other words, his recipe was goal directed.

Unfortunately, subjects were not asked to write control recipes for tuning trials. Analyses of action variability were nevertheless conducted for these trials, for two reasons. First, because there are similarities between the startup and the tuning tasks, one can reasonably assume that similar control strategies would have been used by the subjects in both tasks. Second, because these higher level measures gave results that were consistent with what we would expect from the control recipes, they have been tentatively validated and show promise for application to the tuning tasks as well.

**Action frequency distribution graphs**

One method of measuring to what extent subjects put components at certain settings is to plot the numbers of control actions versus settings for a block of trials for each subject. We can expect that any preferences that the subject had for one setting or a group of settings would show up as peaks on these action frequency distribution graphs. Although in this case we have no control recipes that can be compared with the results of action frequency distribution graphs, we can still make comparisons of the graphs across subjects, blocks, or even interfaces.

Figures 17 and 18 show the results of applying this measure to the first and the last blocks of trials for TL. As TL reported in the normal startup tasks, his actions were driven by a specific procedure; we expect that this would also hold for the tuning trials. Comparing the correspondence components between the two blocks, we can see that for VA, VA1, VA2, VB2, and HTR1, the distributions become more concentrated over time, the distributions for VB and HTR2 remain almost unchanged, and VB1 was not used in either block of trials. This measure, then, shows evidence suggesting that TL was becoming more proceduralised over the course of the experiment.

Across the subjects, the graphs failed to reveal any interesting findings. This is not surprising for three reasons. First, this measure is very coarse, and thus is not sensitive enough to reveal subtle changes in control strategies. Second, subjects experienced different initial conditions in the tuning tasks which may effect how consistent their control actions look. Third, each subject likely had different strategies for making the transition to the shut down task, which also may have effected the consistency of their control actions. We found that some subjects

would prepare for the shutdown task by, for example, reducing the water level in the reservoirs as the tuning stage drew to a close (as was the case with TL).

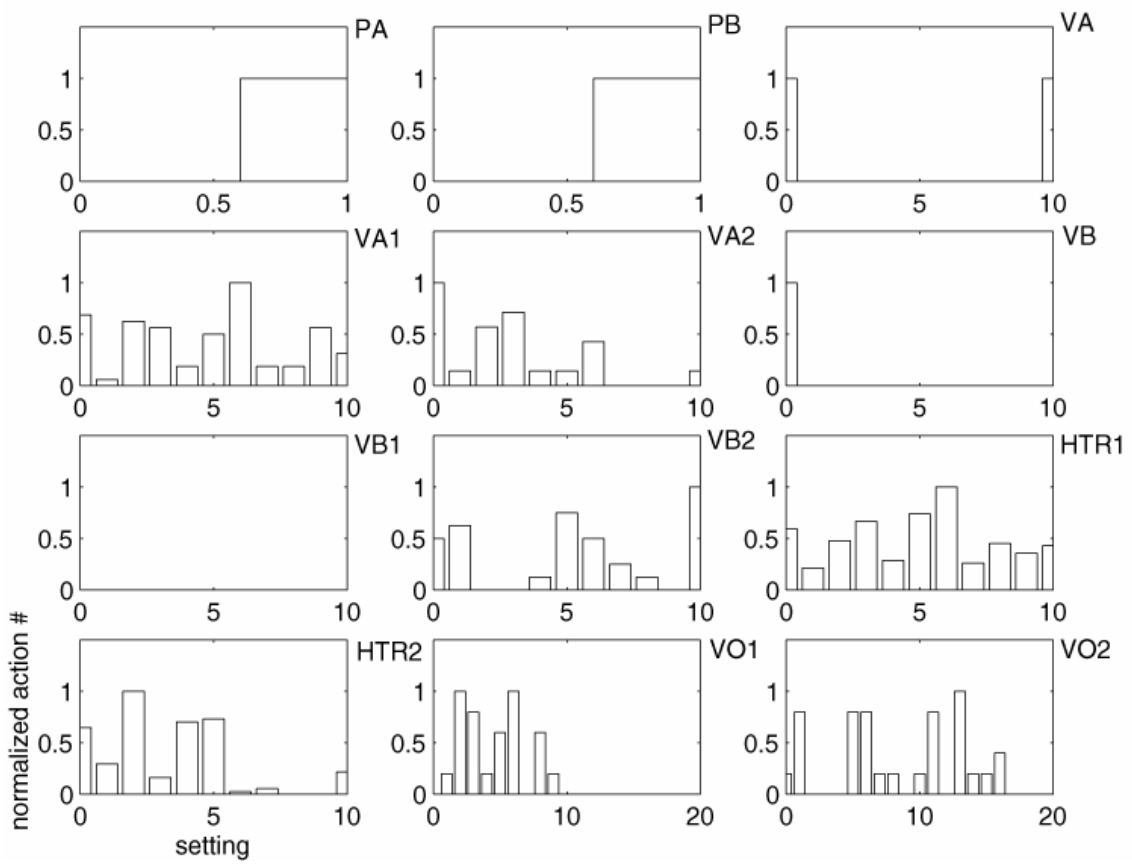


**Figure 17:** Normalised number of actions vs. settings at the first tuning block of TL

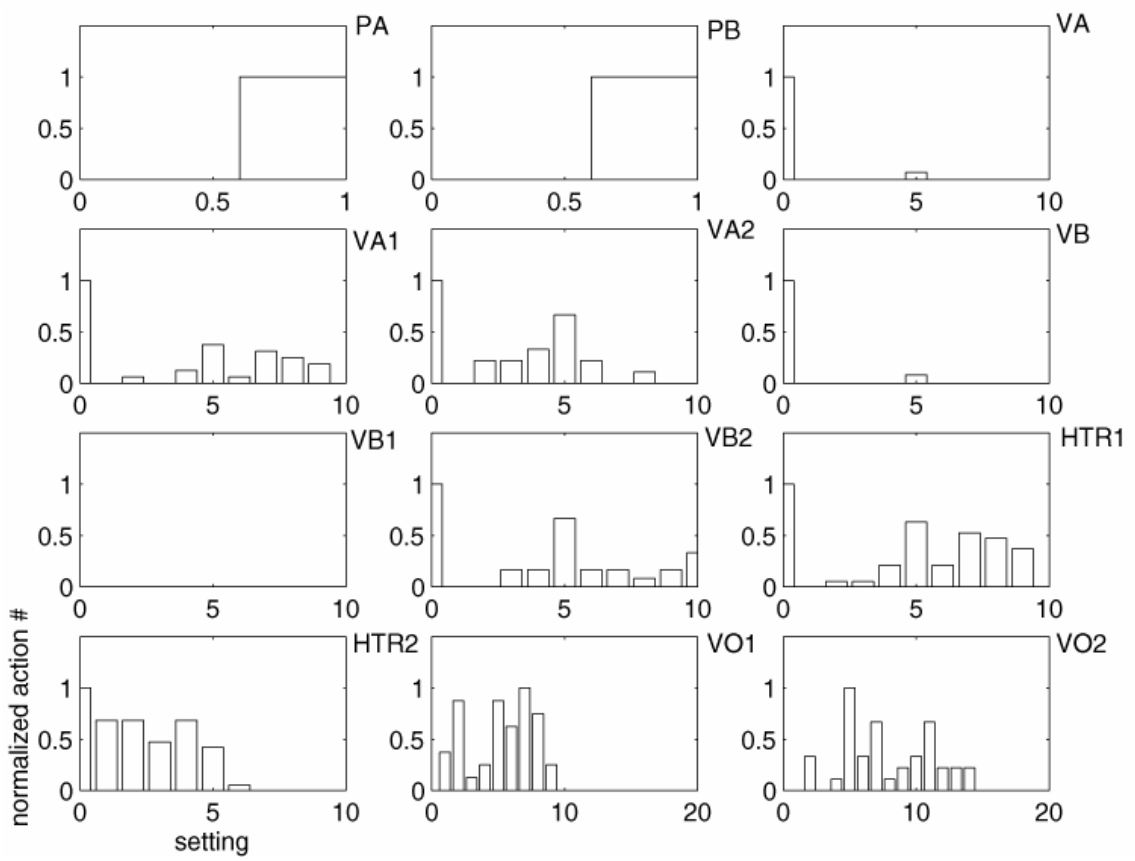


**Figure 18:** Normalised number of actions vs. settings at the last tuning block of TL

**Information theory**

As we pointed out above, action frequency distribution graphs are a very coarse measure. Since inferences can only be made from action frequency graphs by way of visual inspection, so this measure is also not very formal. To make a more fine-grained investigation of subjects' strategies, it would be useful to have a more precise and objective basis for documenting action variability. To elaborate on what was already seen in the action frequency diagrams, we are interested in knowing three things:

- Did subjects use all of the components equally?
- Did subjects choose all the settings equally if the settings are categorised?
- Did subjects exhibit a tendency to consistently choose specific settings for certain components?

Information theory was identified as a potentially viable metric for quantifying the variability in component settings, and so may be able to address the above three issues. Three entropy measures correspond to these questions:

- $H_c$ measures the variability in component usage. If a subject were to act on the 12 components with equal frequency, then $H_c$ is maximised (1). At the other extreme, if a subject were to use only one component, then $H_c$ is minimised (0).
- Similarly, $H_s$ is a measure of variability in component settings[1]. Thus $H_s$ is maximised if all the possible setting values are equally used, and is minimised if only one setting is used.
- The most important and interesting measure in the context of this research is $T_{cs}$ which measures the correlation, or degree of interaction, between specific components and settings. $T_{cs}$ approaches zero if the settings are completely unrelated to components, and $T_{cs}$ approaches 1 if all of the components are consistently put on unique settings.

It should be noted that $H_c$ and $H_s$ alone can not help us discriminate between action-based and function-based control. If a subject were using an action-based strategy to control DURESS II, we would expect his $T_{cs}$ to be high because his procedure would be comprised of precise values for particular components. In contrast, if someone was using a function-based strategy to control DURESS II, we would expect his $T_{cs}$ to be low because the relationship between components and settings would be varied as a function of goals to be achieved. Thus we predicted that the $T_{cs}$ of TL would be lower than that of AV.

The results of information theory ($T_{cs}$) analyses are shown in Figure 19 for all subjects. As expected, TL's $T_{cs}$ is higher than that of AV for almost all blocks.

It is worthwhile to re-emphasise here that we again suffered from the limitations of information theory that were identified last year. In addition to the limitations of information theory measures that were identified in last year's contract report (Yu et al., 1997a), we found increases in $T_{cs}$ are a mix of two factors. While a higher interaction of components and settings would cause an increase in $T_{cs}$, as subjects accumulate experience their control strategies might become more efficient, requiring fewer mediating settings on components. In other words, $T_{cs}$ does demonstrate the interactions between components and settings, but it is generally also increased with experience. With this in mind, it is not so surprising to see that AS, IS, and WL

---

[1] Information Theory measures require that continuous scale measurements be categorised prior to analysis.

experienced a slight increase in $T_{cs}$ over time while ML (the worst performer) experienced no increase in $T_{cs}$.

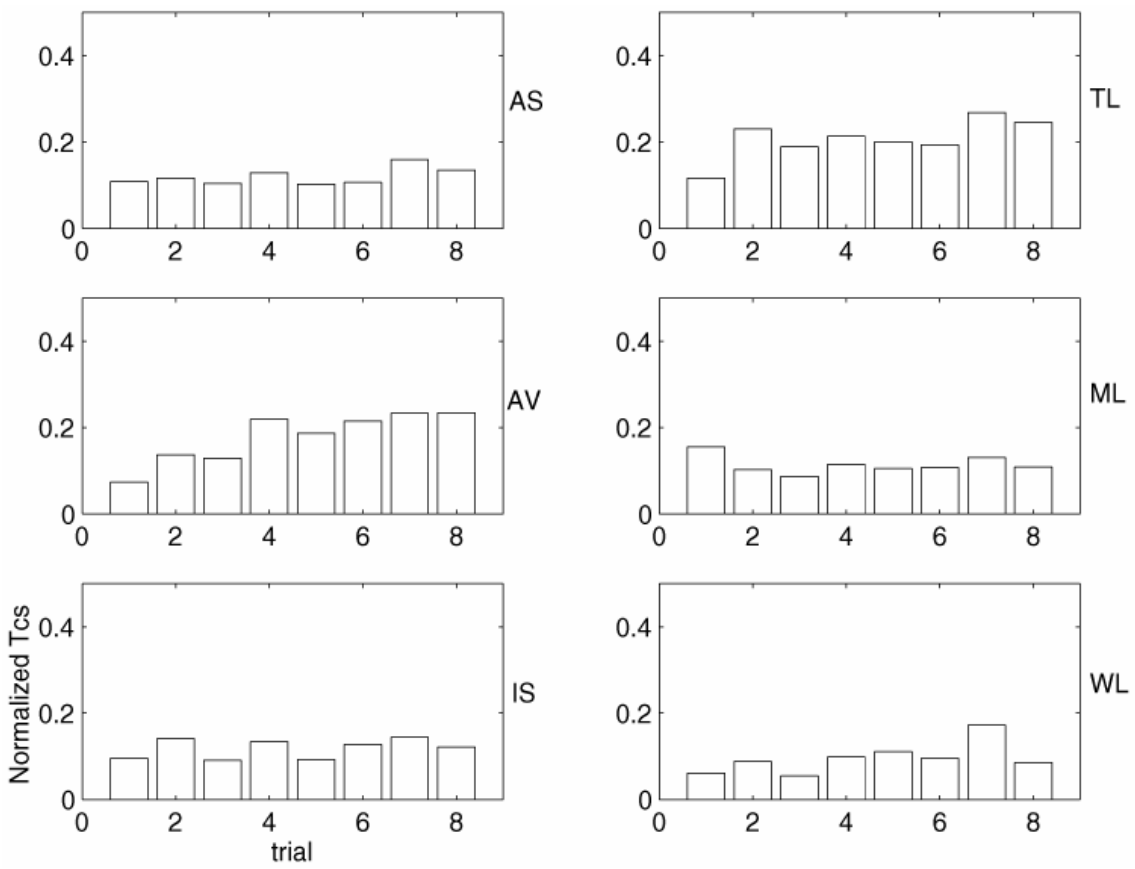


**Figure 19:** Normalised information transfer $T_{cs}$ for all subjects

**Abstraction hierarchy**

One of the theoretical findings of the work performed for JAERI last year was the measures we developed based on the abstraction hierarchy (Rasmussen, 1986) and on a dynamical systems approach to complex systems (Yu et al., 1997b). The measures presented thus far only reveal subjects' behaviour at a very low level. Action frequency distribution graphs studied only subjects' variability at the action level, while information theoretic approaches investigate control strategies at the somewhat higher level of component/setting interactions. We cannot learn a great deal about subjects' information processing behaviour from just these two measures.

Before discussing the results of an abstraction hierarchy analysis for tuning trials, a few key elements of the underlying methodology will be highlighted. The rationale behind the

method is a two dimensional space consisting of an abstraction dimension and a part-whole dimension (see Figure 20). There are five levels of abstraction connected by means-end links (physical form, physical function, generalised function, abstract function, and functional purpose), and there are three levels of resolution for the system connected by part-whole links (system, subsystem, and component). The two dimensions are coupled in practice: at higher levels of abstraction (e.g., functional purpose) operators tend to think of the system at a coarse level of resolution (e.g., system), whereas at lower levels of abstraction (e.g., physical function) operators tend to think of the system at a detailed level of resolution (e.g., component). Therefore, only four cells have been identified as being useful for the present purpose:

- Functional Purpose / System – system outputs
- Abstraction Function / Subsystem – mass and energy balance
- Generalised Function / Component – water flow and heat transfer rates
- Physical Function / Component – component settings

The Physical Form level was not used because spatial appearance and location were not simulated in the DURESS II microworld.

At each of the four levels, a model can be created for the DURESS II system. If a subject used a higher level model (for instance, level 1), we would expect that the subjects actions would be functionally based. This would be evidenced by consistent behaviour (i.e., low variability) at the level of system outputs. On the other hand, if a subject used a low level model (for instance, level 4), then we would expect that her actions would be driven by a procedure. This would be evidenced by consistent behaviour (i.e., low variability) at the action level.

The abstraction/decomposition space provides four different frames of reference to describe system behaviour. For example, in the Abstract Function/Subsystem level, the system behaviour can be represented by a trajectory over time in a multi-dimensional space composed of mass and energy variables. In the rest of this section, we will detail the results of variability of trajectories in each of the four frames of reference.

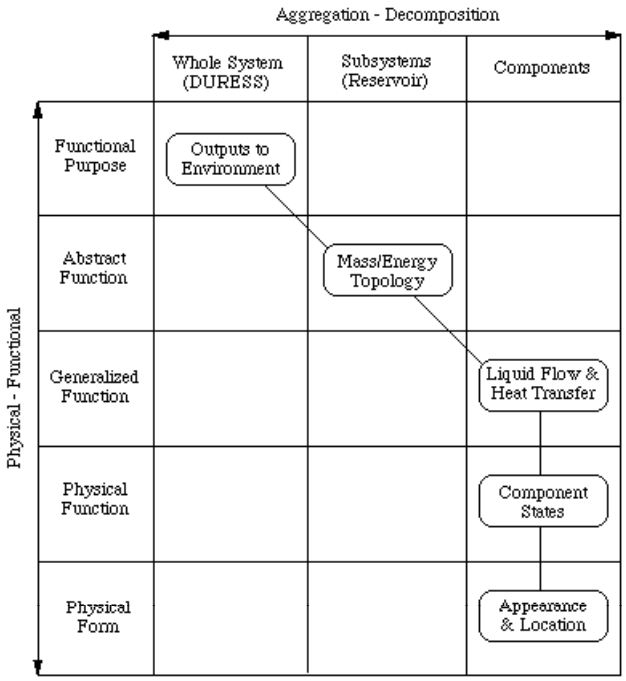


**Figure 20:** Representation of DURESS II in abstraction / decomposition space (from Bisantz & Vicente, 1994)

We start by discussing the results from the goal space level. Note that in this research we consider the trajectories for the two goal variables separately (last year they were treated together, in a two-dimensional space). This has been done because the temperature goal does not change during the tuning task. Subjects' concern with temperature does change from the startup goal of meeting a specific goal to the tuning goal of ensuring that this temperature is not perturbed. To capture this, the system was described using two goal spaces, one for temperature and the other for the output flowrate.

Figure 21 shows the variability of output flowrates by block (of trials) for each subject. For the sake of meaningful comparison across blocks, these trajectories were normalised by the demand change values for each trial. Keeping in mind that TL and AV are generally considered the best performers, from these graphs it is difficult to see any interaction between variability and subject performance. These graphs also do not reveal the strategies employed by subjects. There are at least two reasons for this. One reason is that some subjects tended to set the output valves once when the new output goals came into effect instead of using the output valves to

adjust the temperature or to balance mass and energy. The second reason is that the time constant of output valves was only five seconds, making it very easy for the subjects to control the output flowrates.

Figure 22 illustrates the results for variance of trajectories of temperature goals. As in the normal startup trials, the data for TL and AV are quite similar, both of their variances reducing over the course of the experiment. Therefore, this measure does not allow us to differentiate between TL and AV's control strategies. Note that these graphs are quite consistent with the performance analysis of temperatures. AS and ML, the subjects with the worst performance in their groups, had very unstable performance, while the rest of subjects were quite consistent.

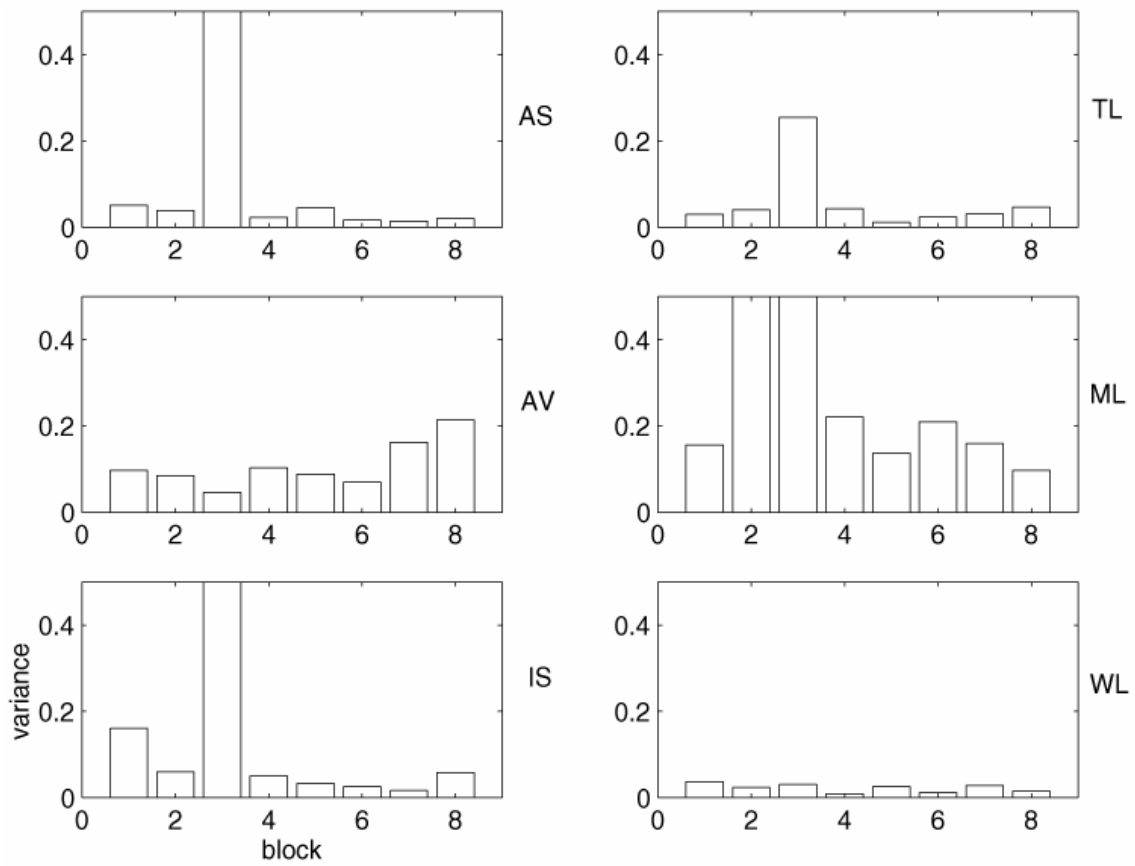


**Figure 21:** Variance of outflow rates for all subjects

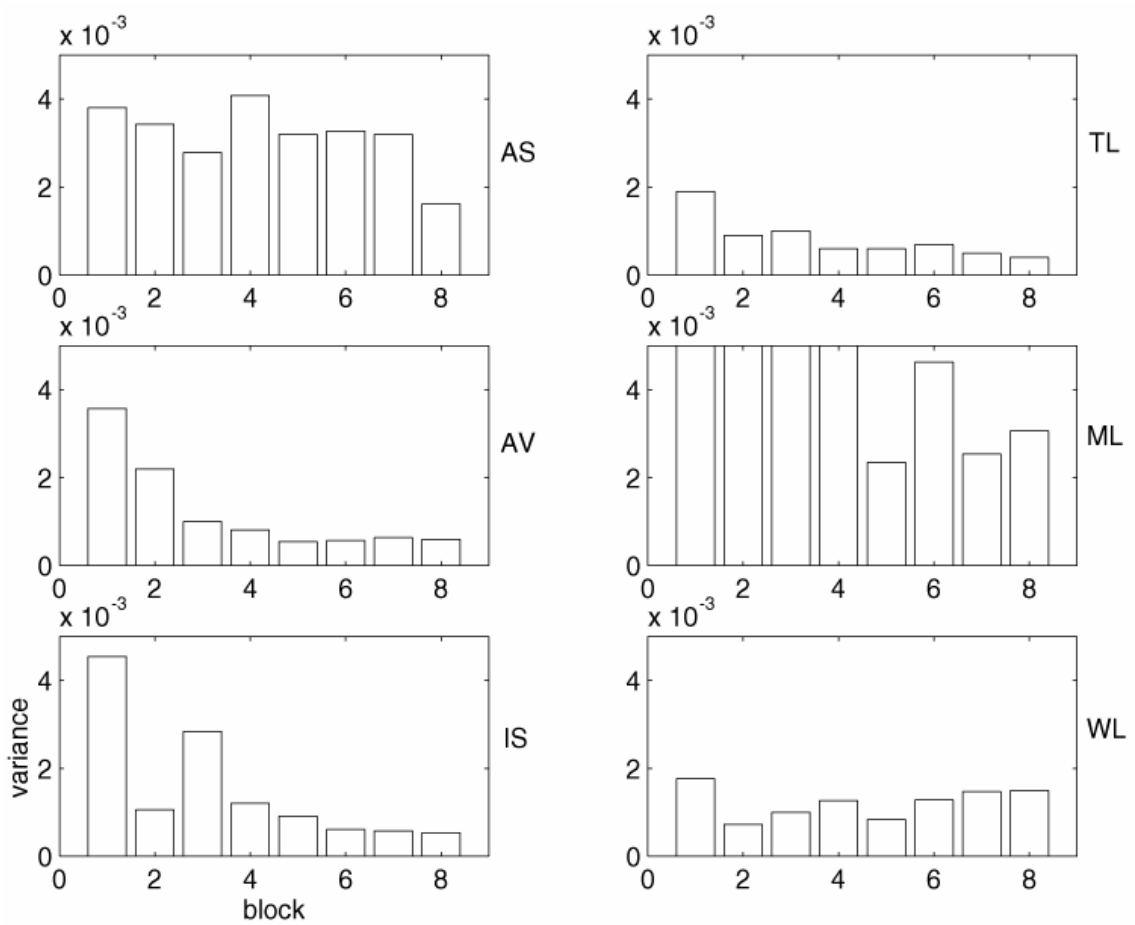


**Figure 22:** Variance of temperature for all subjects

The second set of variance analysis were conducted at the Abstract Function/ Subsystem level. Two different methods were used to calculate this variance. The first was based on trajectories that were not normalised by demand changes (except for a compensation for the fact that different components have different scale values; see Yu et al., 1997a). The results for this analysis are shown in Figure 23. Notice that all subjects' variance reduced over the last four blocks of the experiment, though not necessarily monotonically. Still, the graphs failed to differentiate meaningfully between the control strategies used by the two groups of subjects. Having had the same experience in the analysis of normal startup trials, we are not surprised at this result. There are several factors that effect the values of this variance. In addition to those pointed out in last year's contract report, the different initial conditions in the tuning tasks for each trial for

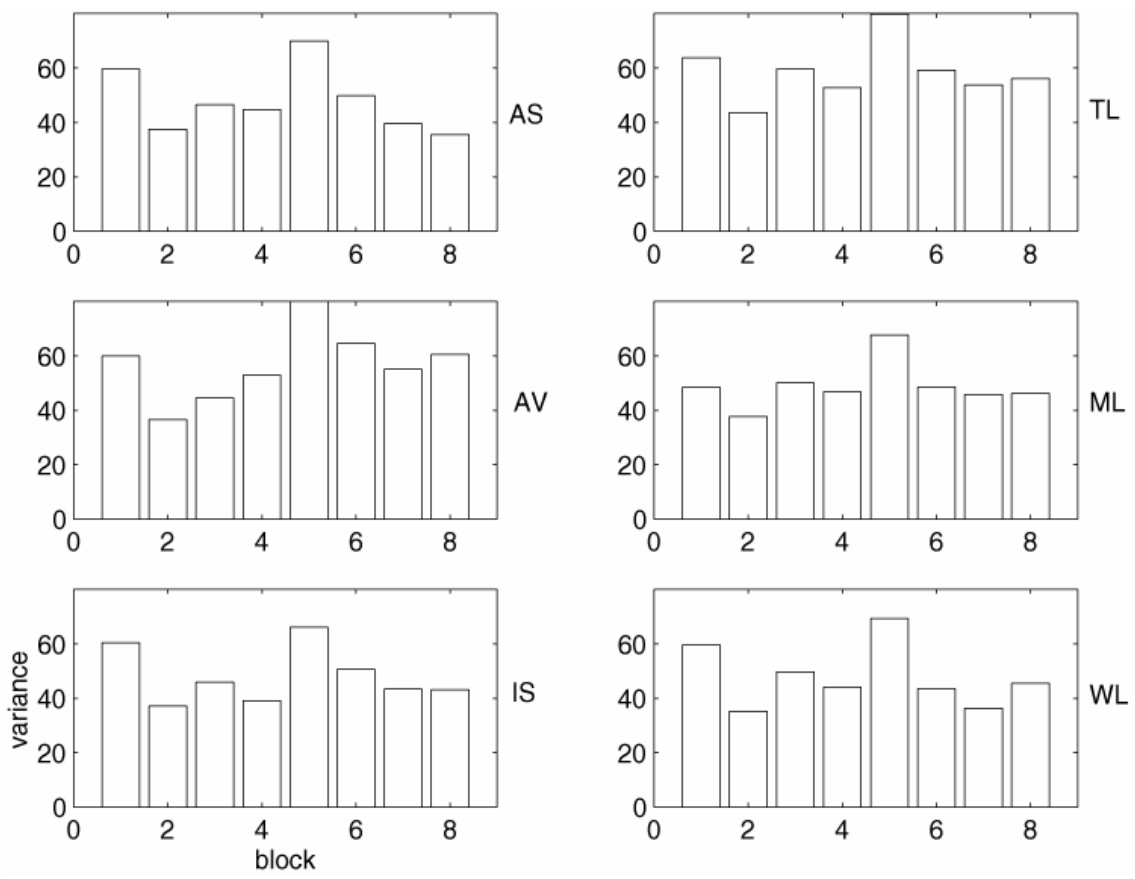


**Figure 23:** Variance of mass and energy for all subjects (normalised by scale only)

each subject may be responsible for the failure of this variance measure to reveal any interesting findings. We will further discuss this point later in this report.

Another method we used to calculate this measure was to explicitly take the changed demand goals into account, thus correcting the trajectories for difference in setpoint values across trials. The method of doing this was somewhat different, as in this case we needed to divide by the mass/energy input and flowrate only since there was no change in temperature goals during the tuning tasks.

The results of this analysis are shown in Figure 24. Unfortunately, these graphs do not seem to reveal much more information than those in Figure 23. While normalising the trajectories in this fashion eliminated the variability caused by the difference in setpoints across trials, it is apparent that the strong effect of the different initial conditions for each of the tuning tasks still adds noise to the results.

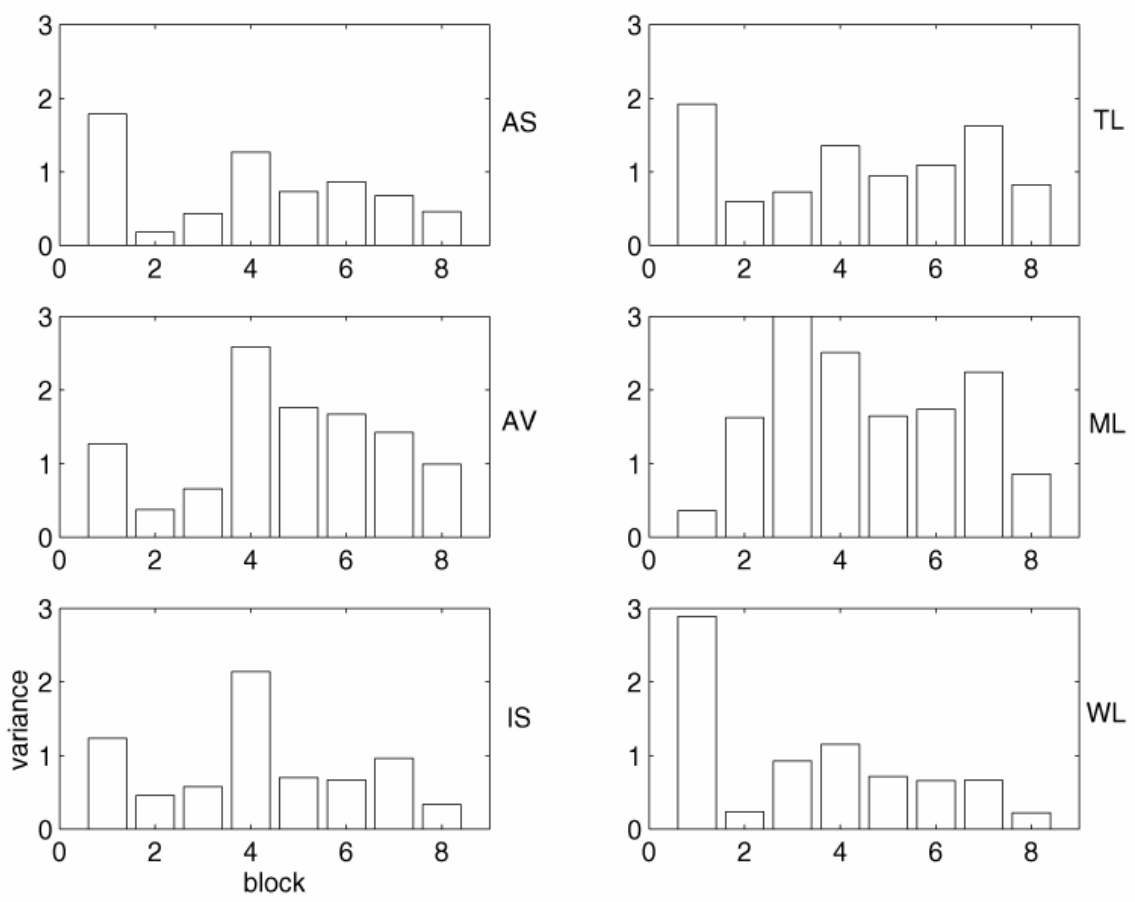


**Figure 24:** Variance of mass and energy for all subjects (normalised by both scale and demand changes)

The third set of variance analyses were calculated at the Generalised Function/ Component level. At this level, the variables of interest are flowrates and heat transfer rates. Similar to the mass and energy variables, subjects cannot act on the flowrates and heat transfer rates directly, but must effect a change in these variables by manipulating the various component settings. Figure 25 shows the results of this analysis. It is interesting to note that the results from this frame of reference are similar to those shown in Figures 23 and 24. Again, these results are not as informative as we would have hoped because of the effect of the different initial conditions in each of the tuning tasks.

The final set of variance analyses were computed at the lowest level, Physical Function/ Component. If we consider the goals as the system state that must be achieved, the variables in the Physical Function/Component level are what the subject must use to

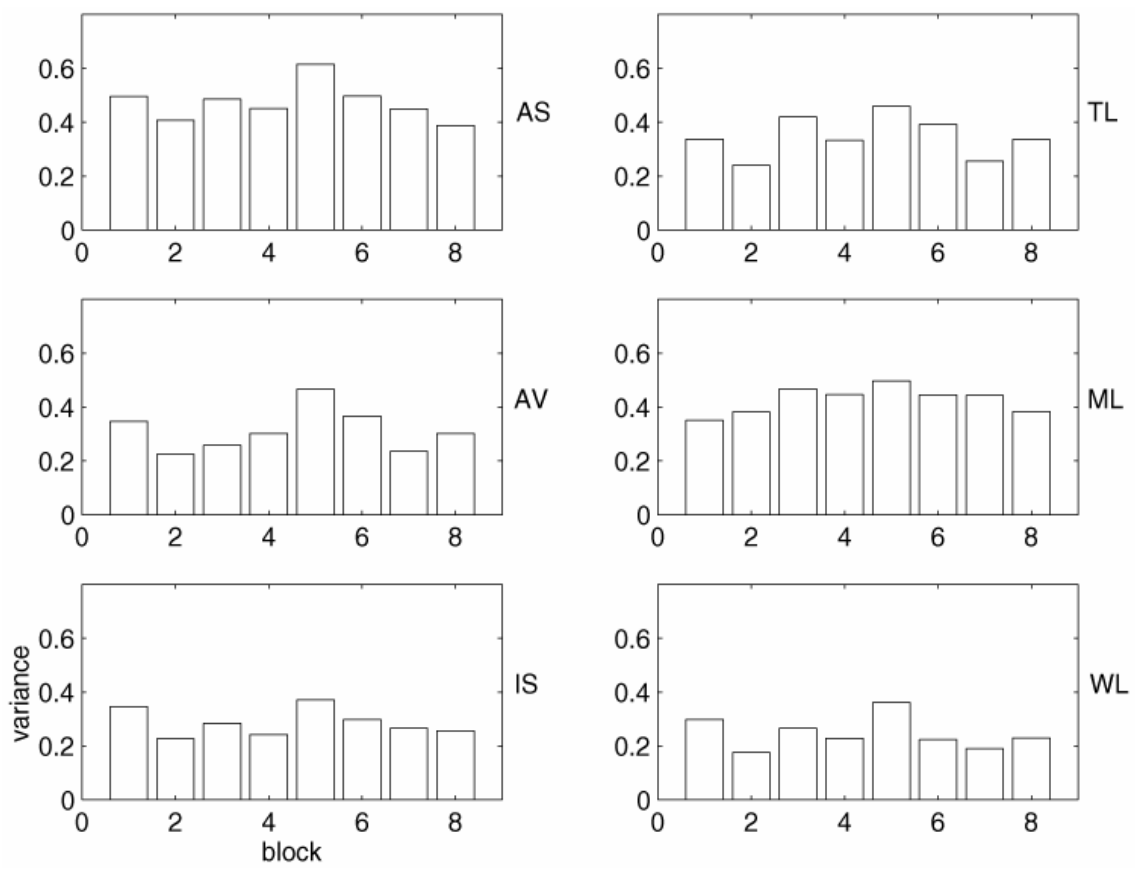


**Figure 25:** Variance of flows and heat transfer for all subjects

achieve these goals. In this analysis, the variance was not normalised by the demand changes for each trial. Such normalisation would be unreasonable as there is no direct relationship between goal values and component settings. Thus, the variability analysis at this level was based on the absolute values of component setting in each action by the compensation of scales only (see Yu et al., 1997a).

The results of this analysis are shown in Figure 26. AS and WL had the lowest values in each groups, while AV and TL had the largest values in each group. These graphs do not give any evidence of strategy differences between the subjects. This is unfortunate, as a similar analysis performed on normal start-up trials showed evidence that TL's behaviour was driven by a fixed procedure while AV's behaviour was more goal directed. Again, we believe that in this case this measure was again contaminated by the differences in initial conditions across tuning trials.

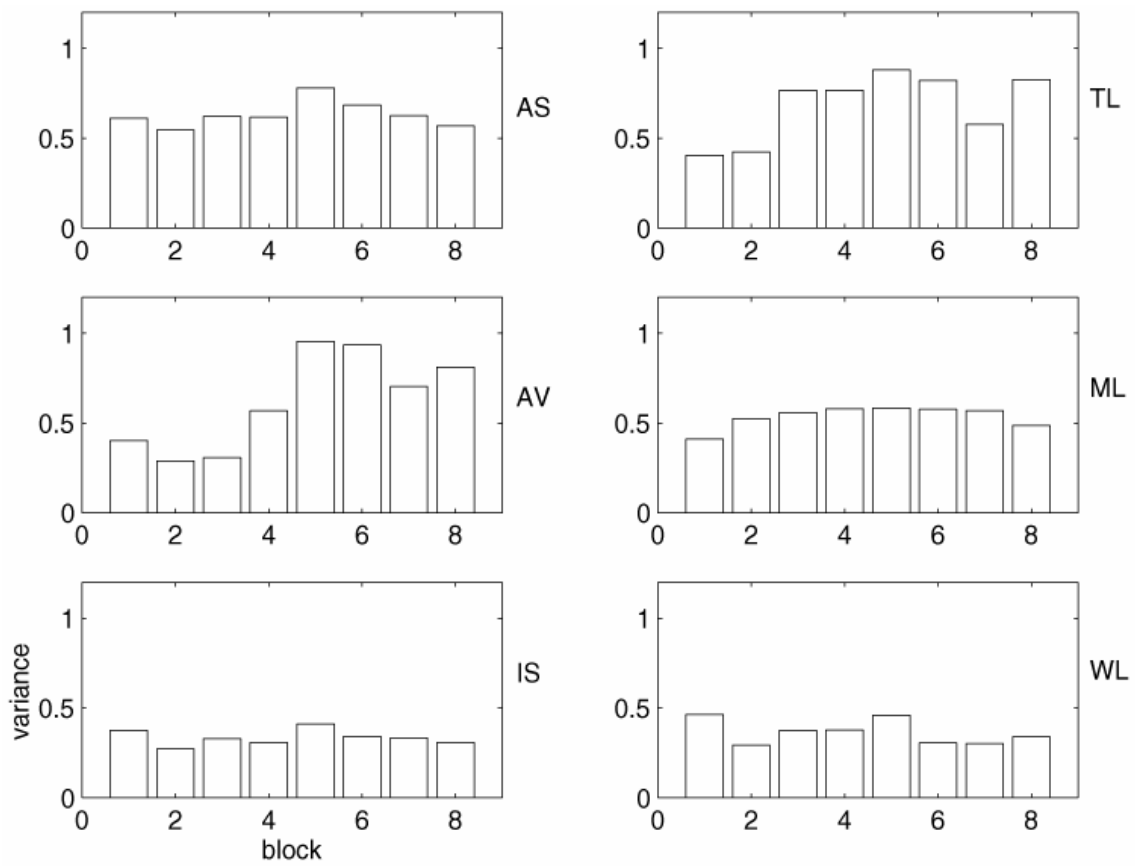


**Figure 26:** Variance of component settings for all subjects (in action space)

**Discussion**

In this section, we tried to investigate the characteristics of information processing for each subject, to learn how the subjects adapted to the DURESS II system over a long term experience with control tasks. Although there were no control recipes to guide our hypotheses (as there were in the analysis of the startup task), it is reasonable to assume that similar control strategies to those of the startup task would be reported by subjects performing the tuning task. Accordingly, we would expect TL's behaviour to be strongly procedural while AV's behaviour would be more goal-directed. Just as we were able to show the truth of this assertion for the startup tasks (see Yu et al., 1997a), we expected that these measures would also be sensitive to this assertion for tuning tasks. The analysis of action frequency graphs supported the strategy difference between

TL and AV, while the information theory analyses revealed that, for tuning tasks, TL's control actions were strongly correlated to specific components.

Unfortunately, the various variance analyses were not as fruitful, for two reasons. First of all, subjects' performance on the tuning task was dependent on the final system state from the startup task as well as on their anticipation of the shutdown task. For instance, in the tuning task some subjects prepared for shutdown by reducing the level of water in the reservoirs to facilitate a quick shutdown. The variance analyses presented were sensitive to, and contaminated by, both of these issues. Second, for some of the tuning trials the change in output demand was relatively small, leaving the subjects with little to do during the tuning period. In other words, the tuning trials did not always provide equal opportunity to observe strategy differences.

These two points are important methodological contributions for future experimental design with DURESS II. To remedy the first issue, identical initial conditions could be set for each subject on transition into the tuning task. As well, the tuning period could be changed from a strict five-minute period to an indeterminate period of time, which would reduce subjects' ability to anticipate shutdown. The second issue can be remedied by ensuring that tuning tasks always involve a demand change of at least double the setpoint tolerance (i.e., if subjects were required to meet some output demand, plus or minus two units, all tuning tasks should involve a demand change of at least four units).

### 3.3. Adaptation to System State

In feedback control systems, the control signals are functions of system states. Causal relationships exist between these control signals and system states. Control signals are driven by the evolution of system states, and in return, the control signals cause the system states to change toward the required values. The aim of system design is to define such a "good" relationship between control signals and system states. For a human-machine system, things are not so simple as human actions may be driven by system state, the operator's experience, or other events. In the current research, it would be interesting to explore the question of which system states were highly relied upon by the subjects to control DURESS II. To do this, we will consider measures that reveal the relationship between subjects' actions and the time at which a situation requiring action took place. These measures will focus directly on the relationships between operator actions and system state, and will hopefully help in understanding why subjects performed specific actions when they did by showing how operators adapted to the system state.

**Measures of action-state interaction**

Actions can be driven by external system events and internal events of human operators. External events are mainly the displayed information of the plant to be controlled, that is, the system states. In other words, external events are a part of feedback control. Internal events are the experiences of human operators that can be either relevant or irrelevant to system states, and promote feedforward control. Both feedback and feedforward control are of interest in human factors research. In studying feedback control, an understanding is needed of the kinds of system states that prompt the actions of a human operator, and in studying feedback control an understanding is needed of the impact of the display of external events on the accumulation of experience and expertise by the human operator.

In this section, we will present an analysis of the interaction between operator actions and system states. We will consider three categories of system state that were identified by a dynamical systems analysis and which we believe guide operator actions: time to goal boundaries (TTGB), time to goal contact (TTC), and time to failure boundaries (TTFB). In order to complete the task, subjects must (1) drive the system outputs to goal region while (2) avoiding any potential for the system to reach its failure boundaries, and while (3) maintaining the goal variables within the tolerance regions. If the system is far from the goal region (indicated by a large TTC), then the main task for the operator is to bring the system into the goal region as quickly as possible. On the other hand, if the system is within the goal region, the operator's main task is to keep the goal variable stable enough to remain in the goal regions (i.e., within the goal boundaries).

As before, we divided the actions into five categories (Yu et al., 1997b):

- Goal oriented actions
- Failure avoiding actions
- Fine-tuning actions
- Exploratory actions
- Miscellaneous actions

**Results**

For each subject, we plotted graphs of action frequency vs. TTGB, TTC, and TTFB for the first and last blocks of 20 trials. For TTGB, TTC, and TTFB, a bin size of 15 seconds was used to categorise the continuous time scale that ranges from 0 to 300 seconds (all values above 300

seconds were lumped into the last category). For details of the calculations, see Yu et al. (1997a). If we identify situations with low TTGB and TTFB and larger TTC as critical situations, we would expect to see a great deal of control actions in these situations (thus, high action frequency). Assuming that more proficient subjects get into fewer critical situations, when plotting graphs of action frequency versus TTGB, TTFB, and TTC over a block of trials for these subjects we would expect to see action peaks at low TTGB and TTFB and at high TTC. Further, by comparing graphs from the first and last block of the experiment, we would expect to see higher action peaks on these graphs for the subjects who are well adapted to the system.

Similar to the findings of the startup analysis, there were no strong differences between interface groups. There were, however, slight differences between the results of the first and last blocks. In the interests of brevity we only present the results for the P+F group. The results for the first and last blocks of the P+F group are shown in Figures 27 and 28, respectively.

These graphs show convincingly that all subjects made many more actions at critical times than at any other time. Even in the first block of trials it can be seen that subjects already had enough experience with the system to limit their actions to times when TTGB became small.

The general impression of these graphs was that all subjects made more actions at critical times than at non-critical times. Individually, the graphs of action frequency vs. TTGB suggest that all subjects have obtained experience after the first 44 startup trials, and acted only when TTGB was small. Both AV and IS had higher TTC peaks in the last bin in the last block than in the first block. This indicates that both AV and IS's behaviour became more sensitive to TTC over time. As for TTFB, there was a slight shift to the left in the distributions for all subjects, indicating an increase in the adaptation to

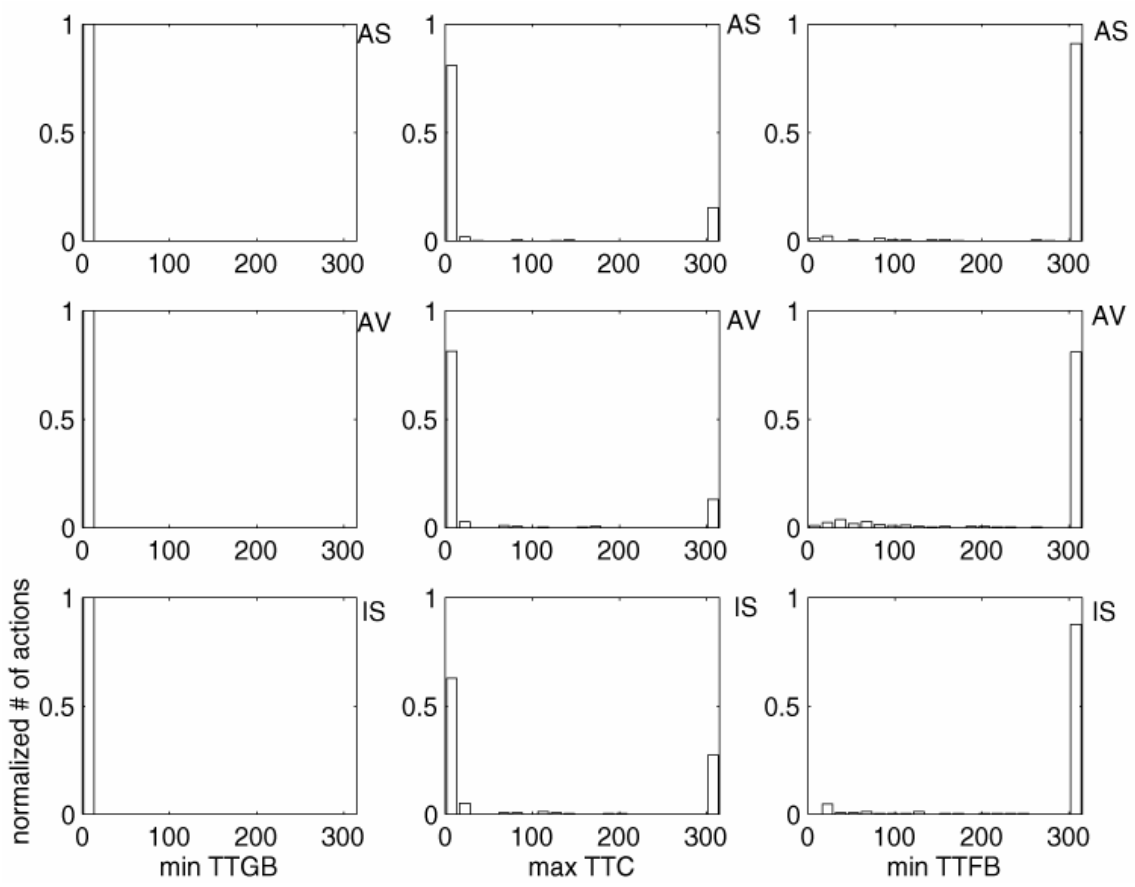


**Figure 27:** Normalised number of actions vs. TTGB, TTC, and TTFB of P+F subjects at the first block

failure boundaries. The data for the P group (except for ML) was consistent with this finding.

**Discussion**

These analyses have shown some evidence that all subjects were adapted to the system states with respect to action frequency versus TTC and TTFB. Maintaining the system within the goal regions seemed to be an easy task, and the subjects were able to do so after some experience. As this was the case, it is not surprising that action frequency vs. TTGB was not as informative as we might have expected (recall that in the startup analysis, this measure was relatively informative). There were no clear differences between the interface groups, and very few differences between the more and less proficient subjects.

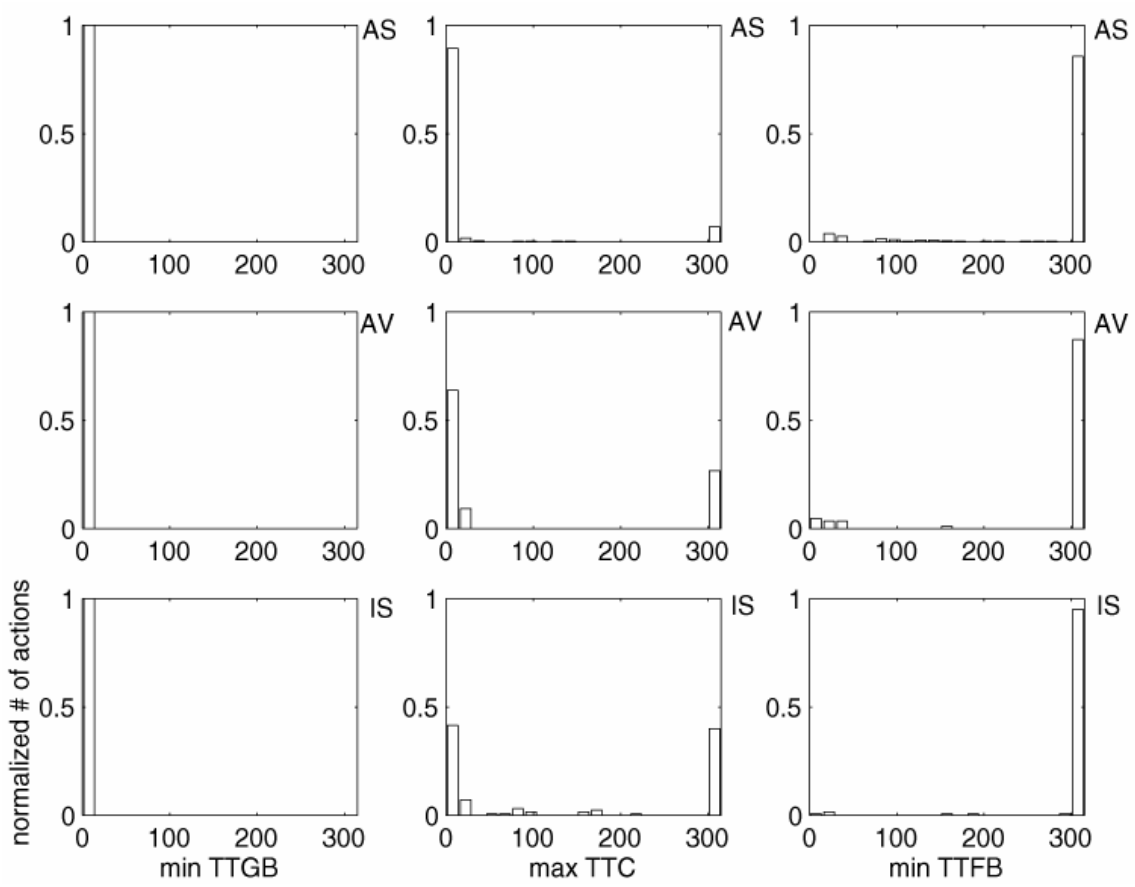


**Figure 28:** Normalised # of actions vs. TTGB, TTC, and TTFB of P+F subjects at the last block

When considered together with the findings of the startup analysis (see Yu et al., 1997a), we believe that in future experiments we could learn more from these analyses if (1) the tuning tasks were made more difficult by ensuring that the change in output demand was large, so that subjects could not easily transition from one goal to another, (2) the tolerances around the goal regions were narrowed, and (3) in the analyses a smaller bin size was used in the area of small TTGB, TTFB, and large TTC.

3.4 Adaptation to System Structure

If we consider the system state as a soft constraint that is defined by the system being controlled, then the system structure can be considered as a hard constraint that subjects must adapt to. In this section, we will present analyses investigating how subjects adapted to the structure of DURESS II while performing the tuning task. Comparisons will be made between subjects and between the two interface groups.

Three frames of reference were developed in our previous work for the evaluation of subjects' adaptation to system structure (Yu et al., 1997a). These were:

- the relationships between mass and energy,
- the relationships between actions versus the ratio of water demand and heater settings, and
- the water volumes in each reservoir for each trial.

The first frame of reference focuses on the dynamic structural condition, while the others emphasise steady state system requirements.

**Co-ordinating mass and energy inventories**

During the tuning task, one of the subjects' goals should have been to maintain the reservoir temperature at the goal setpoint (that does not change) while trying to meet the changed output demand. Since the temperature of the water in a reservoir is actually a function of the energy in the reservoir and the mass of the water, subjects could manipulate the temperature using three methods: (1) using the heaters to change the input energy, (2) using the input and output valves to adjust the mass flow through the reservoir, or (3) using the heaters and the valves simultaneously. To understand how subjects controlled the system during each tuning trial, energy versus mass plots can be constructed for each trial and subject. Since the temperature goal was a constant in each control task, it can be represented as a straight line on the graphs of energy vs. mass. Subjects could satisfy this goal using whatever combinations of mass and energy fulfilled the goal. In the tuning trials, the curve of energy vs. mass for successful trials will start from a point on the goal line. Ideally, the trajectory will never leave the line, but if it does, it must eventually stabilise at another point on the line for a subject to successfully complete the tuning task.

By examining the trajectories of the graphs for each trial, we can gain some insight into how subjects co-ordinated the mass and energy levels to achieve the temperature goals. Through a comparison across subjects, we can learn about the differences in this co-ordination between more and less proficient subjects. By considering the progression of trials for each subject, we can learn about the development of co-ordination strategies. We expect that more proficient subjects would rarely leave the temperature line, and that when they do, they would get back to it again quickly. In terms of adaptation to system structure, we would expect that proficient subjects should exhibit faster and more accurate co-ordination of mass and energy.

No strong differences could be seen in the graphs between subjects and across trials. Perhaps the experience gained in the first 44 trials prior to the addition of the tuning task obscured any differences that might exist between subjects or interfaces. For brevity, we only present the graphs of the first and last trial of IS, shown in Figures 29 and 30. As can be seen, his co-ordination curves did improve, although only slightly, between the first and the last tuning trial.

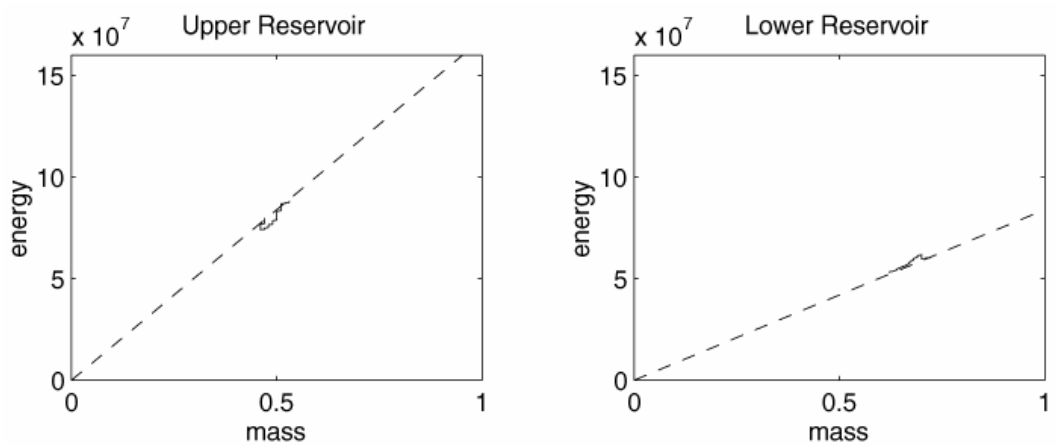


**Figure 29:** Graphs of mass vs. energy of IS at the first trial (trial 45)

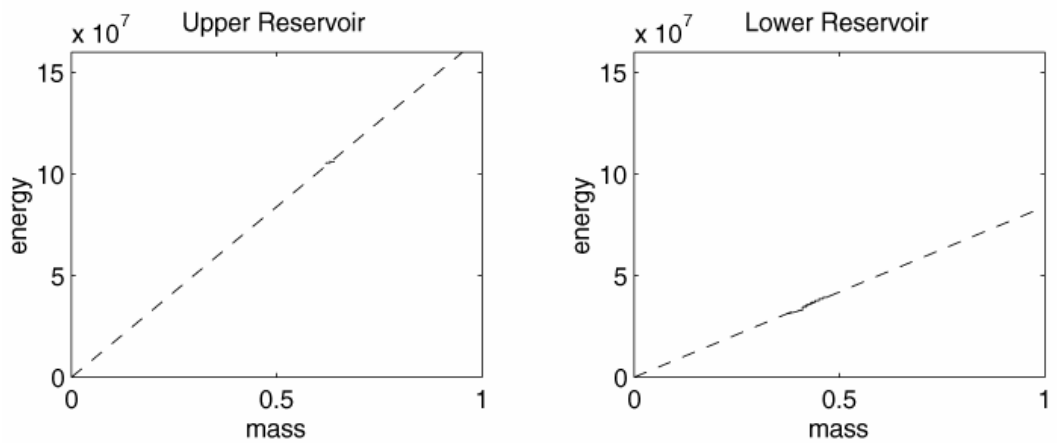


**Figure 30:** Graphs of mass vs. energy of IS at the last trial (trial 217)

**Steady state heater settings**

From the mathematical model driving DURESS II, the relationships between the setpoints of demand goals and heater settings at steady state can be ascertained. For the upper reservoir, the ratio of heater setting and demand is 1:1 (HTR1/D1 = 1/1), and for the lower reservoir, the

ratio is 1:3 (HTR2/D2 = 1/3). This structural constraint is another reference against which to measure subjects' adaptation to system structure. Note that the subjects were not told about this feature of DURESS II, although some of them did come to understand this property of the system with experience (Christoffersen et al., 1994). If subjects learned to adapt to this property, we would expect that they would tend to set the heaters according to these 'ratio rules' to reduce their workload and task completion time.

To illustrate the degree of adaptation of subjects' actions, action frequency was plotted against the ratios of heater settings and demand for each reservoir for a particular block of trials. We would expect that people who had knowledge of the 'ratio rules' should have peaks in the graphs around 1 and for the upper reservoir and 1/3 for the lower reservoir. If subjects were adapting to the system structure over time, we would expect to see the peak of these plots shift to the best values as experience is gained (i.e., across trials).

Figures 31 and 32 show the graphs for the P+F and P groups, respectively, for the first block of tuning trials. WL had a relatively flat distribution that left much room for improvement. The peaks of IS, TL, and ML were particularly prominent, which is consist with their being the only subjects to report using the 'ratio rule' in their control recipes. AS and AV in the P+F group had relatively flat distributions in the first block of trials.

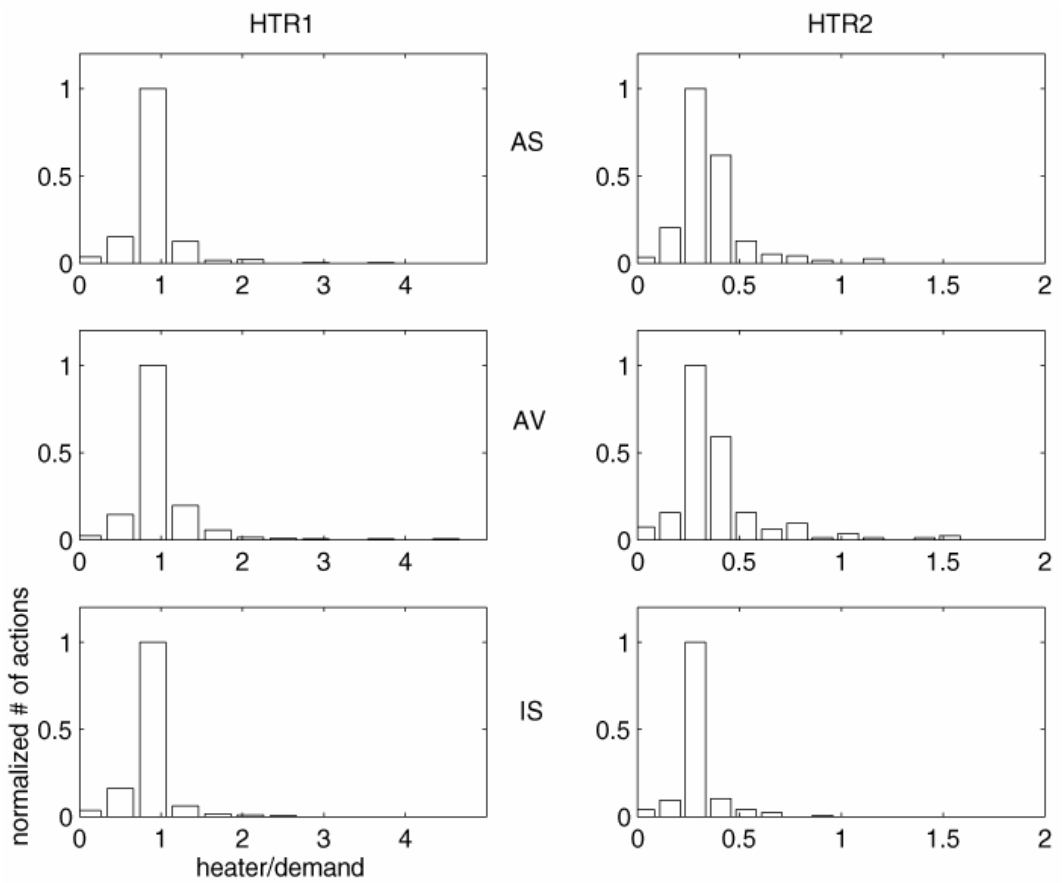


**Figure 31:** Distributions of actions on heaters of P+F subjects at the first block

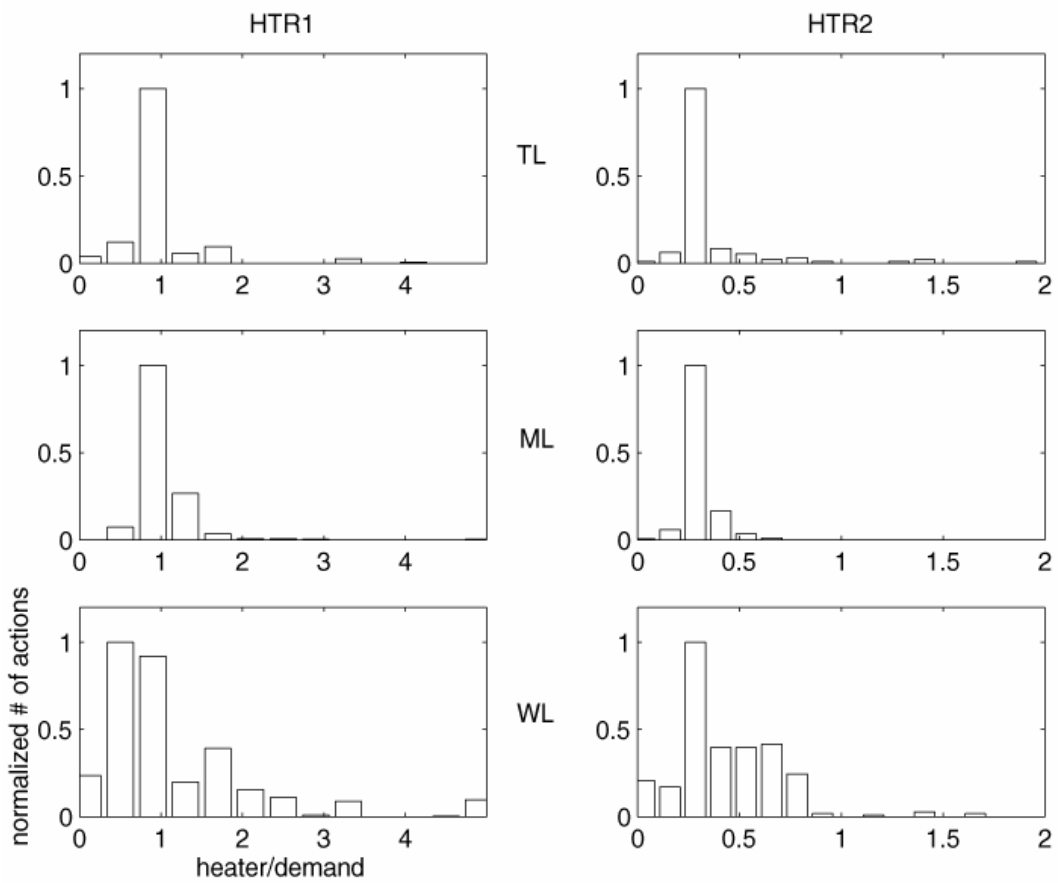


**Figure 32:** Distributions of actions on heaters of P subjects at the first block

Figures 33 and 34 show the graphs for the two groups during the last block of trials. If we compare these graphs to the previous two, some interesting findings stand out. First, the three subjects who had learned the 'ratio rule' had high peaks at the expected ratios of heater setting to demand, indicating that they did indeed follow these rules. Second, the subjects who had much room for improvement in the first block, such as TL and WL, showed signs of adapting to the system structure. Third, similar observations to those made in the analysis of startup trials about AV can be made here (Yu et al., 1997a). As evidenced by the high second peaks in AV's plots and the overall right-skew of the distributions, he was using a very aggressive control strategy that used high heater settings to heat the reservoirs up more quickly and so complete the task faster. No other subject exhibited similar behaviour.

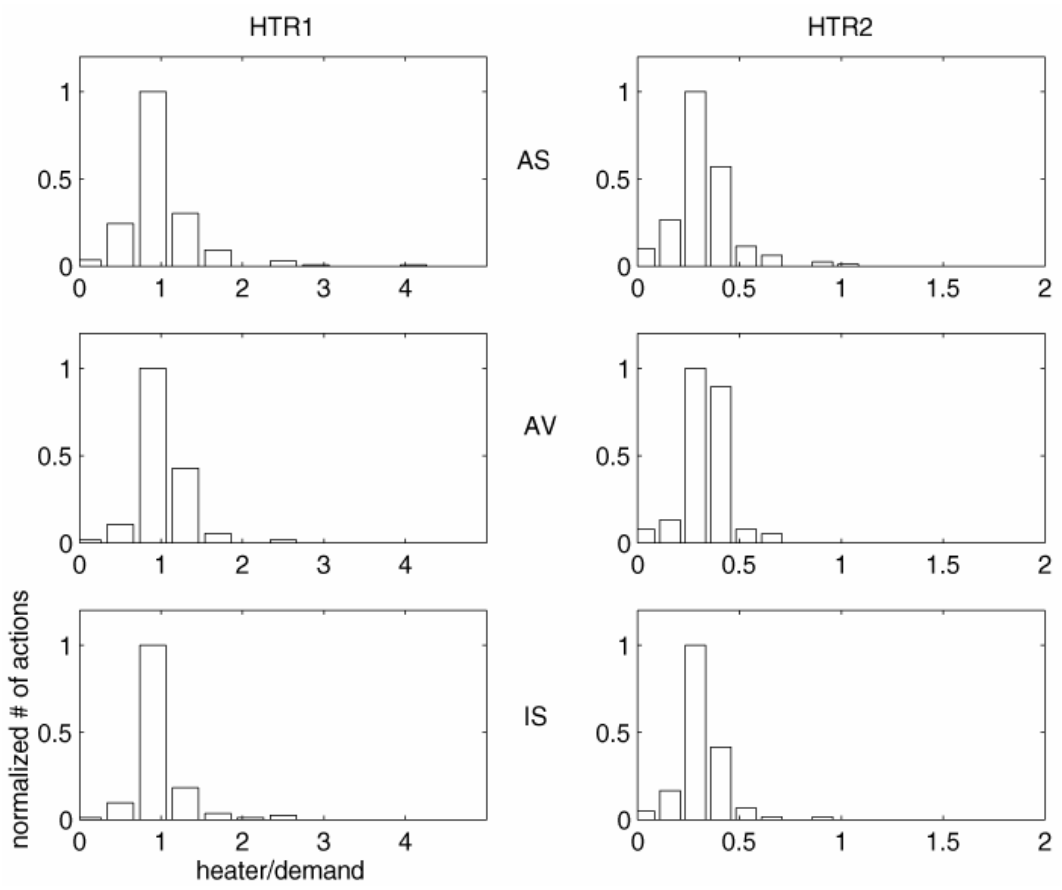


**Figure 33:** Distributions of actions on heaters of P+F subjects at the last block

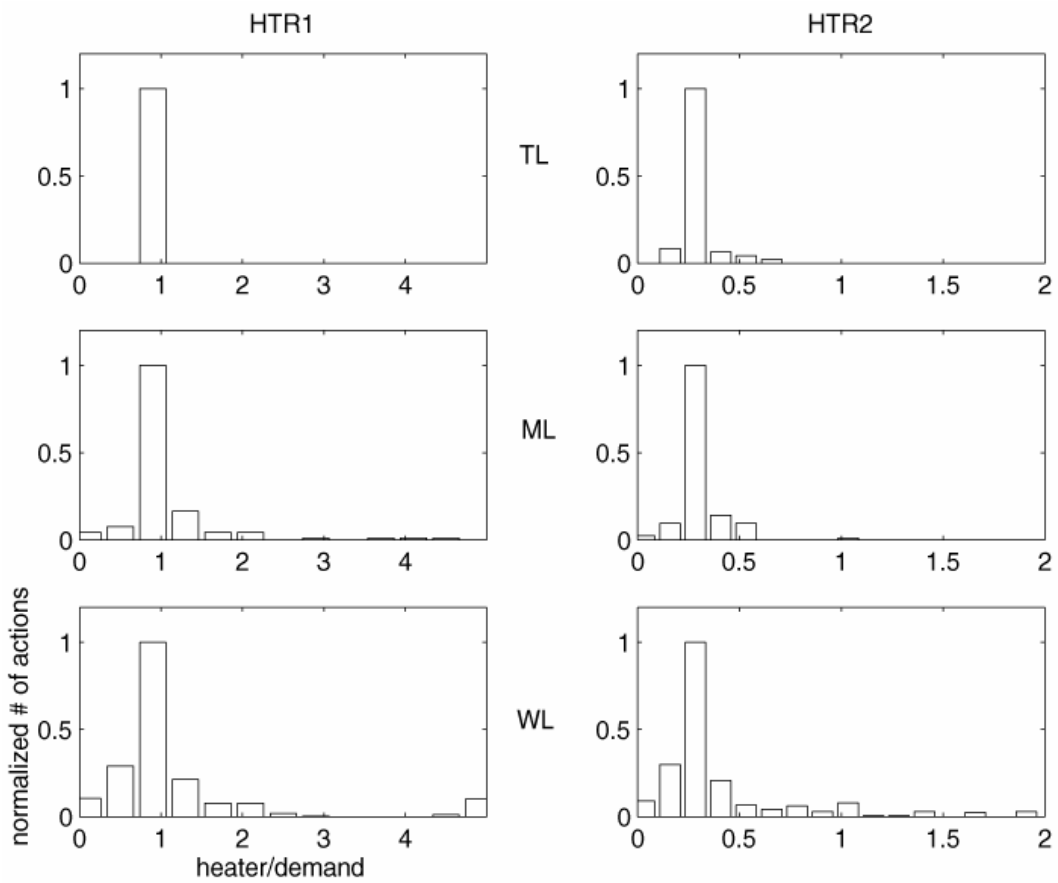


**Figure 34:** Distributions of actions on heaters of P subjects at the last block

**Choice of steady state volumes**

The volume of water in reservoirs can have great impact not only on the ease with which the temperature can be controlled, but also on the speed with which the system can be shutdown after the tuning period. With lower volumes of water in the reservoirs, the temperature of the water in the reservoir is very sensitive to changes in the heater settings (Pawlak & Vicente, 1996). On the other hand, with higher volumes of water, the temperature of the water is more difficult to adjust. In addition, higher volumes generally necessitate a longer period of time to shutdown the plant, while lower volumes facilitate faster shutdown. Since the shutdown task directly follows the tuning period, some subjects were able to anticipate the beginning of shutdown while other subjects were even able to prepare for shutdown by lowering the water levels in the reservoirs.

In our previous study (Yu et al., 1997a), steady-state reservoir volume was shown to be a useful measure to illustrate subjects' adaptation during the startup task. The same analysis here

was applied to the tuning task. The volumes at steady state for each trial were plotted against trial number for each subject. The results are shown in Figures 35 and 36 for the P+F and P groups, respectively. A first impression from these graphs is that the least efficient subject in each group (AS in P+F group, ML in P group) used either very high or very low volumes. Second, we see that AV, the most proficient subject in the P+F group, and TL, the most proficient subject in P group, clearly anticipated and prepared for the shutdown tasks. In some extreme trials, they were able to almost empty the reservoirs before the shutdown task began.

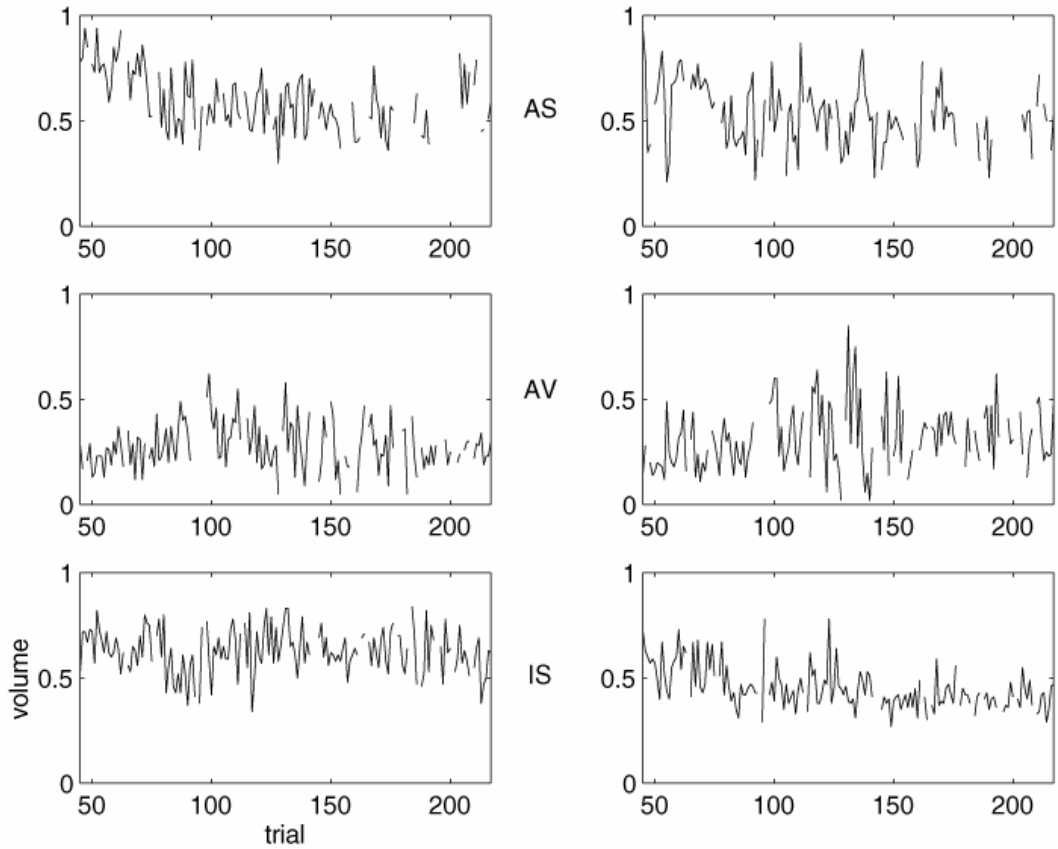


**Figure 35:** Reservoir volumes at steady state of P+F subjects

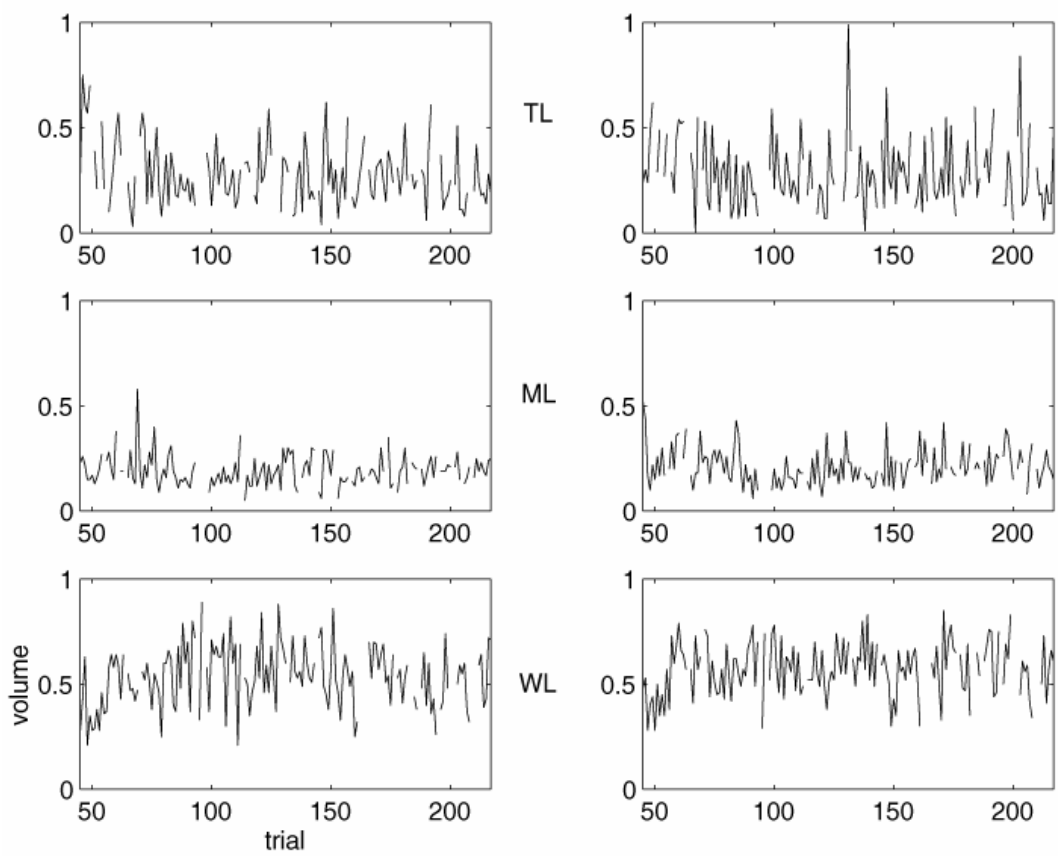


**Figure 36:** Reservoir volumes at steady state of P subjects

Despite the results that were reported above, the results of this analysis are not as informative as those for the startup task. This is not surprising because the volume used during the tuning period is dependent on the steady-state volume chosen during the startup period. While a change in output demand might cause subjects to alter the volume of the reservoir as they work to meet the new demand, it is unlikely that subjects consciously aimed to change the volumes as specific reservoir volumes are not system goals. The results of this analysis were likely also contaminated by subjects' anticipation of and preparation for the shutdown task.

**Discussion**

The identification of structural constraints is very important for the design of control interfaces for large scale control systems. Operators must identify these constraints because they must operate within them. Structural constraints also provide a context in which operators can search for and find efficient ways to control the system. Thus, it is important to study how human operators use these constraints to more efficiently adapt to the system structure.

The most interesting finding from these three measures is that control experience from the startup tasks (for instance, the 'ratio rule') carries on into and is beneficial for the tuning task. Further, where there was room for improvement when startup strategies were applied to the tuning task, subjects adapted. Unfortunately, in the present work analyses of the mass/energy relationship and reservoir water levels failed to reveal interesting results like they did in the startup analysis (Yu et al., 1997a). In our opinion, the problem does not lie with the measure used, but with the tasks they were applied to. Mass (water level) and energy in the tuning tasks were both inherited from the startup tasks and were adjusted for the shutdown tasks. These two facts may muddy the waters of these measures, preventing us from obtaining any clear findings.

### 3.5 Adaptation to Information Processing Limitations

The adaptation behaviour of subjects may be determined by their environment as well as by their own mental and workload limitations. In the DURESS II microworld, the environmental factors primarily include system state and structure, which have been addressed above. Below we present an analysis of long term adaptation to human operators' information processing limitations. (Recall that this type of analysis revealed some outstanding findings when applied to the startup task (Yu et al., 1997a).

Just as in our previous work, the approach taken here has been influenced by the work of Payne, Bettman, & Johnson (1993) on adaptive decision making. The fundamental insight behind this work is that people select their strategies based on a demand-resource trade-off (see Rasmussen, 1986). Accordingly, subjects will adapt their behaviour over the course of an experiment to their own limited capacity for information processing. In this section, we will explore how the feedwater control strategies of the subjects were adapted to their information processing abilities. We would expect that adaptive operators would try to perform the feedwater stream control task effectively while simultaneously trying to minimise the information processing demand needed to complete the trials.

**The feedwater stream configuration subtask**

In DURESS II microworld, there are six input valves (VA, VA1, VA2, VB, VB1, VB2), two pumps (PA, PB), and two output valves (VO1, VO2) that can be manipulated by subjects to achieve the required goals. In most cases these control valves are redundant, and subjects do not have to use all of them to complete the tasks. Therefore, an interesting problem the subjects have to face is to decide which valves and what settings will be used to finish the tasks, that is, to

choose the feedwater stream control configuration. In this research, we primarily focused on the six input valves, the settings of which are decided by output demands. More formally, the subtask of feedwater stream configuration can be formulated by the following function:

$$(VA, VA1, VA2, VB, VB1, VB2) = f(D1, D2)$$

Since the control valves are redundant, this function cannot be solved until further constraints are added. From a normative perspective, we can ask the question, "What criteria should operators use to resolve the remaining degrees of freedom?" From a descriptive perspective, we can ask: "What criteria do operators use to resolve the remaining degrees of freedom?" To answer the first question, the feedwater control task must be formulated more precisely by breaking it down into three nested decisions:

1. How many valves will be used?
2. Which valves will be used?
3. At what quantitative value will each valve be set by the end of the trial?

**Results of normative analysis**

A normative analysis was conducted by Yu et al. (1997a) based on Payne et al.'s (1993) work. A number of system constraints were identified that all subjects had to obey. We postulated that the way in which subjects make each of the above decisions is constrained by two factors: the characteristics of DURESS II and psychological criteria. The criteria for each of the three issues associated with the feedwater stream are summarised in Table 1 (for more detail, see Yu et al., 1997a):

**Table 1:** Constraints on problem resolution in DURESS II

| Problems | System constraints | Psychological constraints |
|---|---|---|
| Number of valves? | feedwater stream capacity | minimise monitoring demand & interaction |
| Which valve? | maximum flowrates & demand pair | stimulus-response compatibility |
| Valve setting? | demand pair | computation |

Based on the system constraints and operator's information processing limitations, we derived a normative algorithm which might provide an answer to the above problems. This algorithm is reproduced below:

**Table 2:** The normative algorithm to control DURESS II

```
ASSIGN      VA = 10
ASSIGN      VB = 10
READ        D1
READ        D2
ASSIGN      VA1 = D1
COMPARE     D2 to 10
    if D2 ≤ 10
        ASSIGN      VB2 = D2
    else (D2 > 10)
        ASSIGN      VB2 = 10
        SUBTRACT    D2 – 10
        ASSIGN      VA2 = (D2 – 10)
```

The algorithm described in Table 2 can be taken as an ideal way to control DURESS II if the normative constraints are considered in the extraction of control methods. In order to see if the subjects adapted their behaviour to approach this way of controlling DURESS II, we calculated an error score to measure the difference between the valves settings used by the subjects in each trial and the valve settings predicted by the algorithm. An error score of zero will mean that the method the subject used completely agreed with the ideal algorithm, while the

larger the error is, the further a subject's method was away from the ideal one. We plotted the error score versus trial number for each subject. The results are shown in Figure 37.

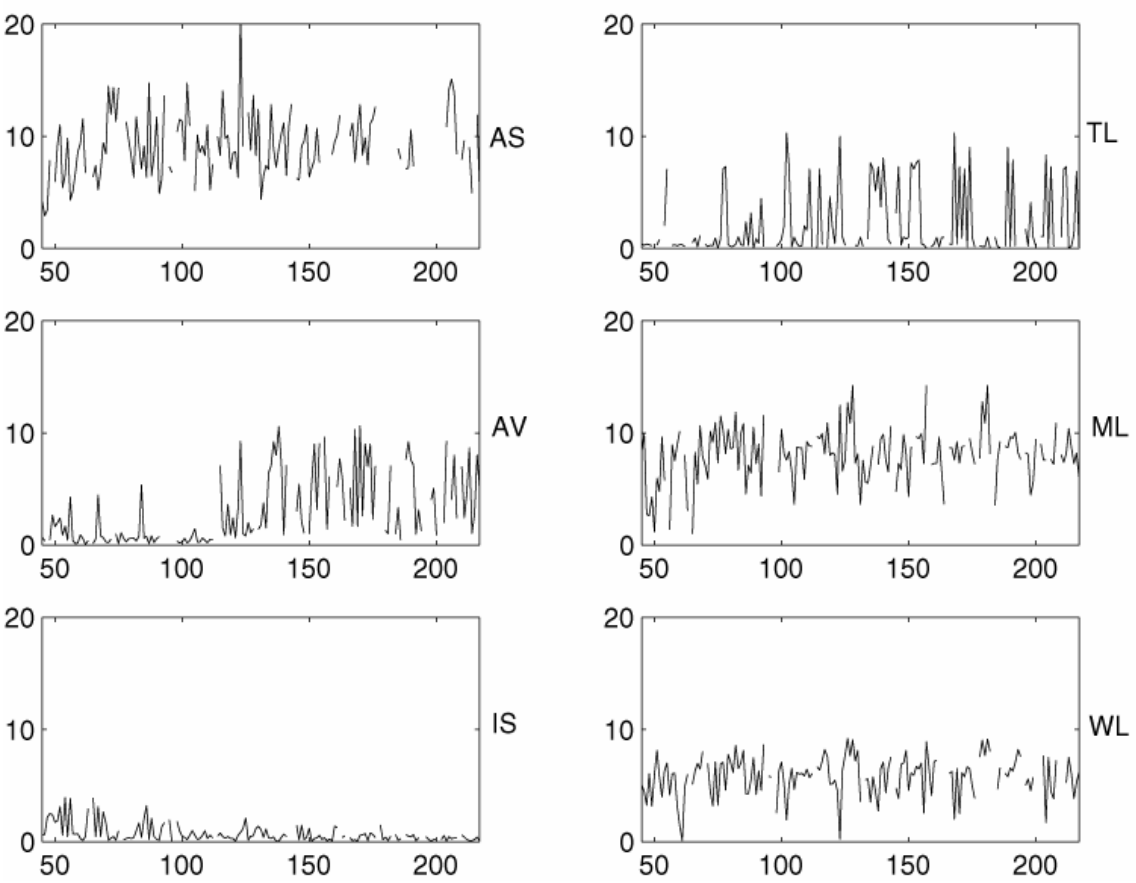


**Figure 37:** Error between practical valve settings and those predicted by Table 2

IS' control method was quite consistent with what was predicted by the normative algorithm. He also seemed to be more and more consistent with his method over the experiment. AV demonstrated good agreement to the normative algorithm during the first half of the experiment, but strayed from this for most trials in the second half probably because he began to anticipate the shutdown task. TL did not show evidence of following the normative algorithm closely at all, also probably because of his anticipation of and preparation for the shutdown task. The data for AS, ML and WL clearly show that they did not follow the normative control strategy.

The subjects who followed the normative strategy used the same control strategies during the tuning period as during their startup sessions. As it turns out, only the people with the best performance in their groups (AV and IS in P+F group, TL in P group) used the normative

algorithm. Thus it seems there is some connection between subjects' performance and the extent to which their behaviour was adapted to system constraints and their information processing limitations. As an aside, these results also confirmed the findings of the startup analysis.

**Results of descriptive analysis**

In order to dig deeper into the control strategies that the other subjects were following, we started by looking at the first of the three decisions outlined above by finding how many valves were used in each trial. Taking into consideration the fact that valve settings are a function of demand pairs, we categorised the demand pairs into three mutually exclusive and exhaustive cases:

- Mode 1: If $D1+D2 \leq 10$, then the minimum number of valves is 3.
- Mode 2: If $D1+D2 > 10$ & $D1 \leq 10$ & $D2 \leq 10$, then the minimum number of valves is 4.
- Mode 3: If $D1+D2 > 10$ & $D2 > 10$, then the minimum number of valves is 5.

Figures 38, 39 and 40 show the number of valves used by subjects in tuning trials of mode 1, 2, and 3. Beginning with Figure 38, there are some interesting findings from these graphs. First, AV and TL clearly anticipated the shutdown tasks in a number of trials, as indicated by their use of less than three valves in these trials. If they were not preparing for the shutdown task, they might have used up to four valves in these trials. Second, subjects following normative algorithm used 4 valves in almost all trials, while the rest of the subjects used the maximum number of valves. Third, since no one used the minimum number of valves necessary to complete the tasks, there was no sign of adaptation in terms of minimising the number of valves used. This might be because the use of the minimum number of valves implies an asymmetrical control configuration, while the symmetric structure of DURESS II might discourage such configurations.

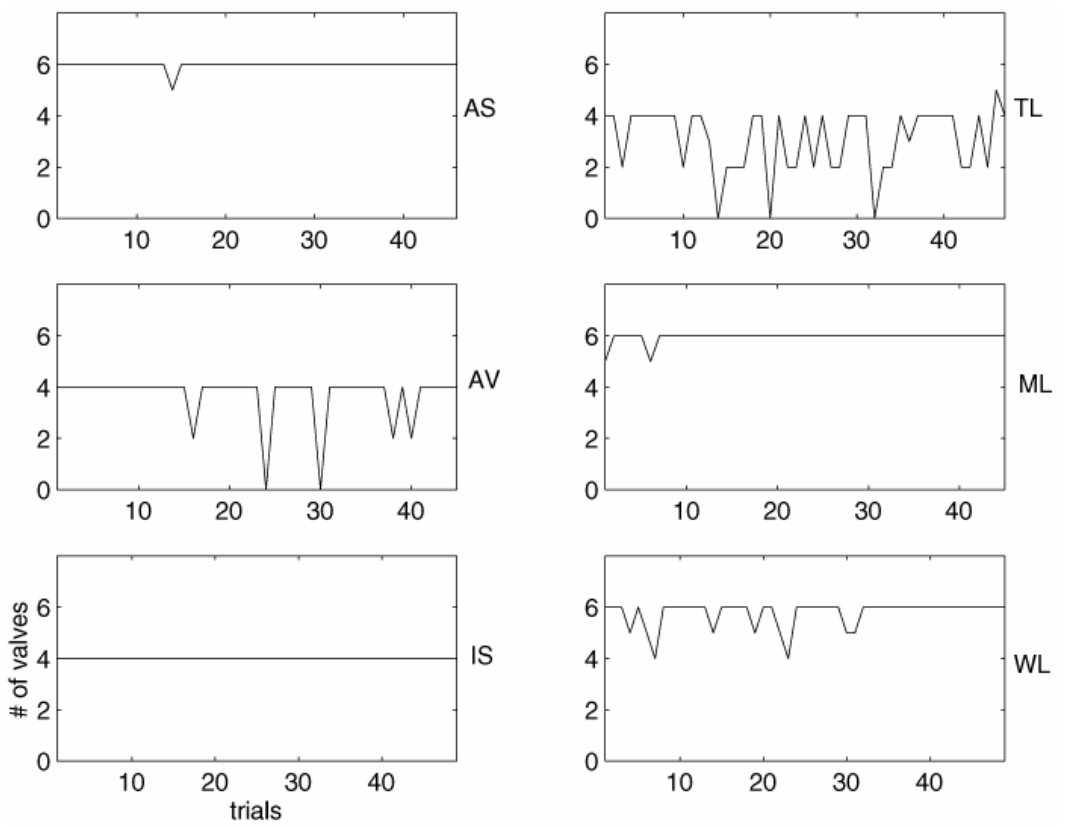


**Figure 38:** Number of valves used at trials with demand pairs of mode 1

Figure 39 shows the results for trials with mode 2 demand pairs. The minimum number of valves required to complete tuning task for this mode is 4. As in Figure 38, we can first see that there are some regions in the curves of AV and TL which might be indicative of their anticipation of shutdown. Second, the subjects with better performance (AS, IS and TL) used the minimum number of valves control strategy in contrast with the poorer performers who used the maximum number of valves control strategy. Third, there were no signs of adaptation to system constraints in these graphs.

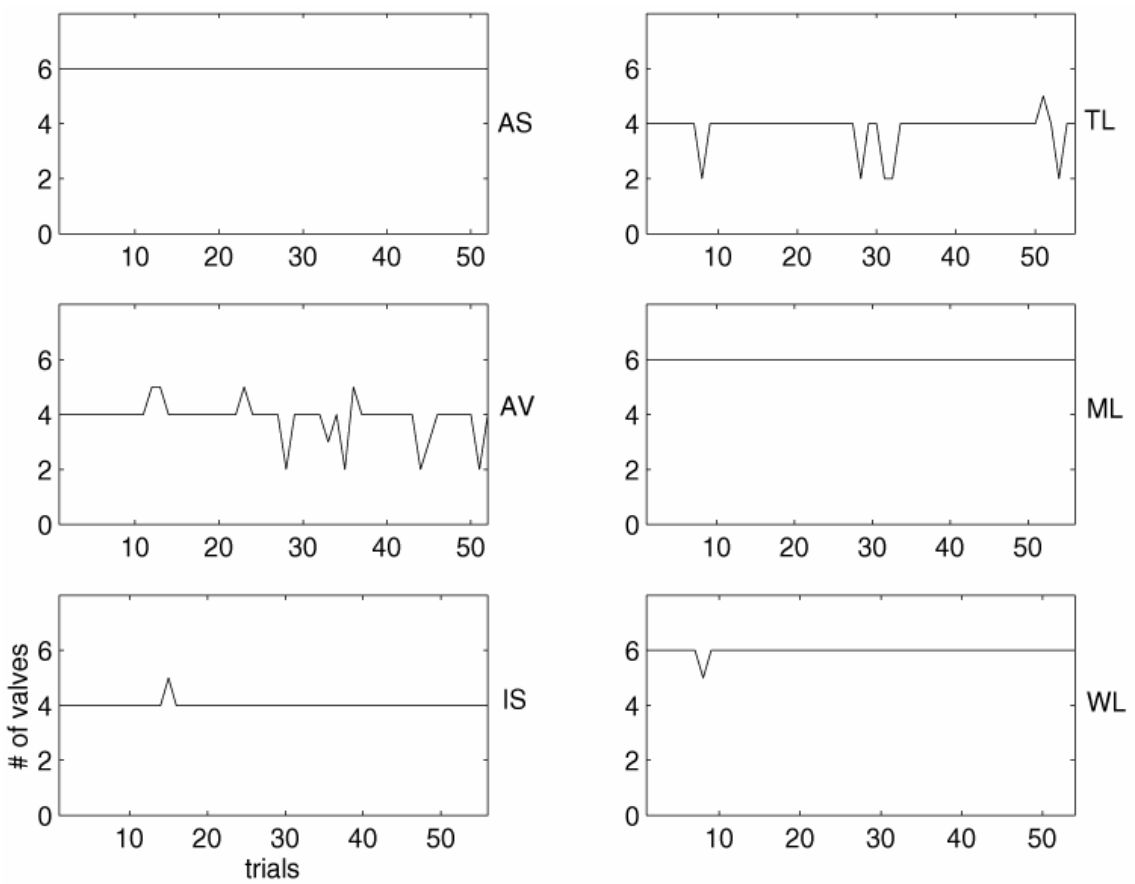


**Figure 39:** Number of valves used at trials with demand pairs of mode 2

Figure 40 shows the data for the trials with mode 3 demand pairs. The minimum number of valves needed for this task is 5. As in the above two figures, we again see that AV and TL anticipated shutdown by the fact that the number of valves they used was often less than the minimum requirement. Since they were using fewer valves than was necessary to complete the task, they must have been emptying the reservoirs. Second, we can also see that the subjects with the best performance in their respective groups used an economic control strategy, while the rest of the subjects rarely used this strategy. Last, these results again show no sign of subjects' adaptation to system constraints.

In summary, the mode control strategies for each subjects show that AV, IS and TL used an economic control strategy in mode 2 and 3 trials, while in mode 1 trials they seemed to balance the information processing effort with the symmetry of the system structure. These three subjects had the best performance in their groups, and therefore it seems that there is some connection between performance and the control strategy the subjects employed. These findings

are quite consistent with what we saw in the analysis of startup tasks. It appears that the control strategy is independent of the tasks that need to be done.

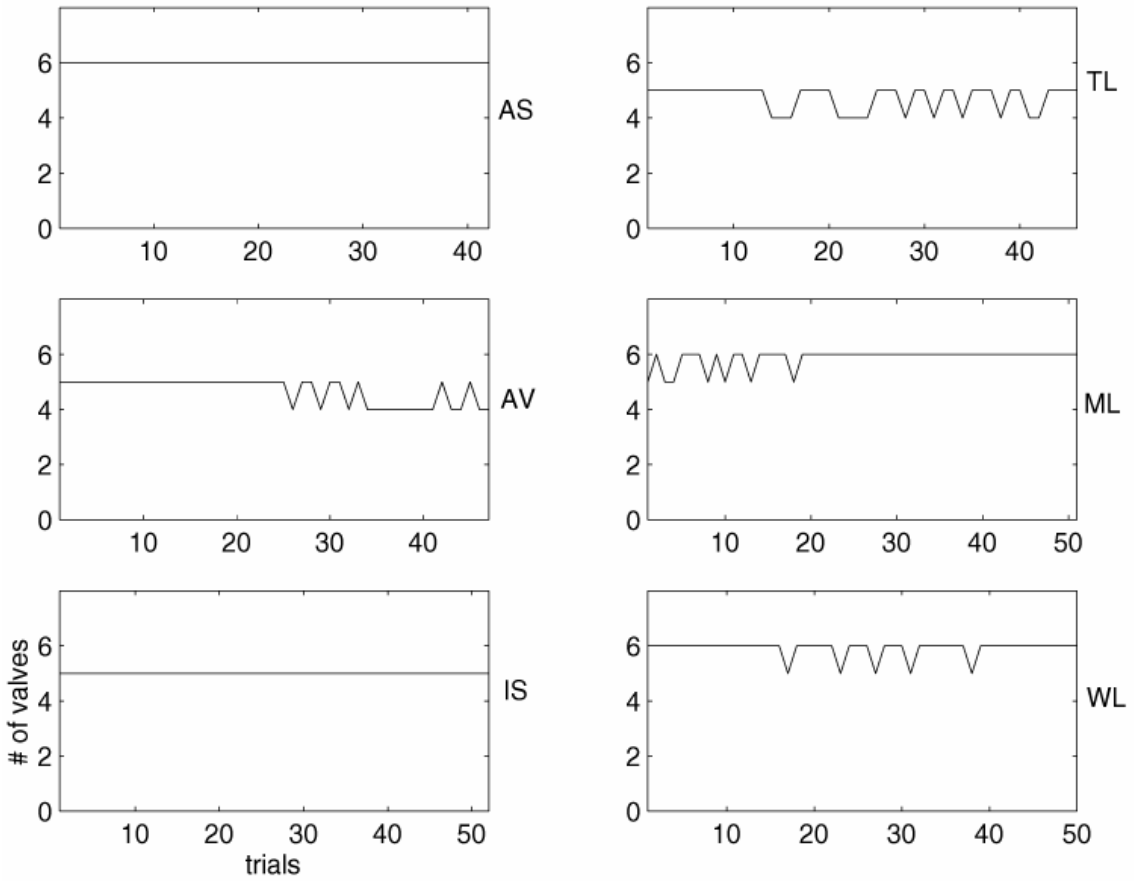


**Figure 40:** Number of valves used at trials with demand pairs of mode 3

Having considered the decision of how many valves were used, the next issue to consider is which valves were used by each subject in trials with different demand pair modes. This analysis was only pertinent to AV, IS and TL who used less than the maximum number of valves to control the system. The results of this analysis are straightforward: When these subjects used only four valves, VA2 and VB1 (the crossover valves) were always the valves that were not used. When they used five valves, VB1 was always not used. These findings are in perfect agreement with the stimulus-response compatibility criterion identified in Table 1.

The final issue is to determine which valve settings were used by subjects in trials with different demand pair modes. As was already seen, the performance of AV, IS and TL is described well by the normative strategy. For the rest of the subjects we used the algorithms inductively identified for the startup tasks in last year's project (see Yu et al., 1997a). Although

there were no verbal protocols recorded during the tuning tasks, and while we have not made an effort to extract the subjects' control strategies during the tuning trials, we still believe that these algorithms apply to the tuning task given the finding that control strategy is independent on control tasks. The algorithms for ML and WL are listed in Tables 3 and 4, respectively.

To see how well these algorithms account for subjects' behaviour, we calculated the error between the actual valve settings and those predicted by the algorithms listed in Tables 3 and 4. The results are shown in Figures 41 and 42. As we expected, the findings from these graphs are quite similar to those we found for normal startup trials (Yu et al., 1997a). Figure 41 shows that the algorithm in Table 3 provided a very good fit to the data of ML, especially late in the experiment. The fit to the data of the other subjects was quite poor, except in the case of WL who followed the algorithm in Table 3 for certain trials. This is because in some conditions, the two algorithms WL used share some similarities.

**Table 3:** The algorithm inductively developed to account for ML's strategies

```
READ            D1
READ            D2
ADD             D1 + D2
DIVIDE (D1 + D2) / 2
ASSIGN          VA = (D1 + D2) / 2
ASSIGN          VB = (D1 + D2) / 2
COMPARE         D2 to 10
if D2 ≤ 10
        ASSIGN          VA1 = D1
        ASSIGN          VB1 = D1
        ASSIGN          VA2 = D2
        ASSIGN          VB2 = D2
Else
        DIVIDE          D1 / 2
        ASSIGN          VA1 = D1 / 2
        ASSIGN          VB1 = D1 / 2
        DIVIDE          D2 / 2
        ASSIGN          VA2 = D2 / 2
ASSIGN          VB2 = D2 / 2
```

**Table 4:** The algorithm inductively developed to account for WL's strategies

```
ASSIGN      VA = 10
ASSIGN      VB = 10
READ        D1
READ        D2
COMPARE     D1 to 1
COMPARE     D2 to 2
if D1 = 1
        ASSIGN      VA1 = 1
        DIVIDE D2 / 2
        ASSIGN      VA2 = D2 / 2
        ASSIGN      VB2 = D2 / 2
if D2 = 1
        ASSIGN      VA2 = 1
        DIVIDE D1 / 2
        ASSIGN      VA1 = D1 / 2
        ASSIGN      VB1 = D1 / 2
else (D1 & D2 > 1)
        DIVIDE D1 / 2
        ASSIGN      VA1 = D1 / 2
        ASSIGN      VB1 = D1 / 2
        DIVIDE D2 / 2
        ASSIGN      VA2 = D2 / 2
        ASSIGN      VB2 = D2 / 2
```

Figure 42 shows the error between the actual valve settings and those predicted by Table 4. This algorithm provided an excellent fit to the data of WL, but a very poor fit to the data of the rest of the subjects. The sole exception was ML who had a similar control strategy in certain conditions.

Note that we again were not able to extract a consistent algorithm for AS, who clearly did not follow any of the control strategies listed in Tables 2, 3, and 4.

**Discussion**

In the framework of the DURESS II system structure, there are many degrees of freedom for solving problems, that is, the tasks that need to be done. Consequently, there are many different ways of performing the very same task. In order to achieve a specific solution, it is up to the human operators to choose what kind of constraints to use to reduce the degrees of freedom. In this analysis of the DURESS II system, we have focused on the feedwater stream control configuration. As we have seen, although the valve settings are functions of demand pairs, these functions are under-specified. We hypothesised that operators will use their information processing constraints to determine which control strategy to use, and that these

strategies will adapt to the operators' information processing limitations to minimise the information processing demands placed on them.

The analysis we have presented in this section revealed some important findings. First, a normative analysis was conducted and a normative algorithm was extracted. This algorithm can lead to a minimum effort solution to the task. We found that three of the subjects (AV, IS and TL) showed signs of following this normative algorithm. If we take into consideration their anticipation of shutdown, the behaviour of these subjects was quite consistent. These subjects adapted to their information processing limitations.

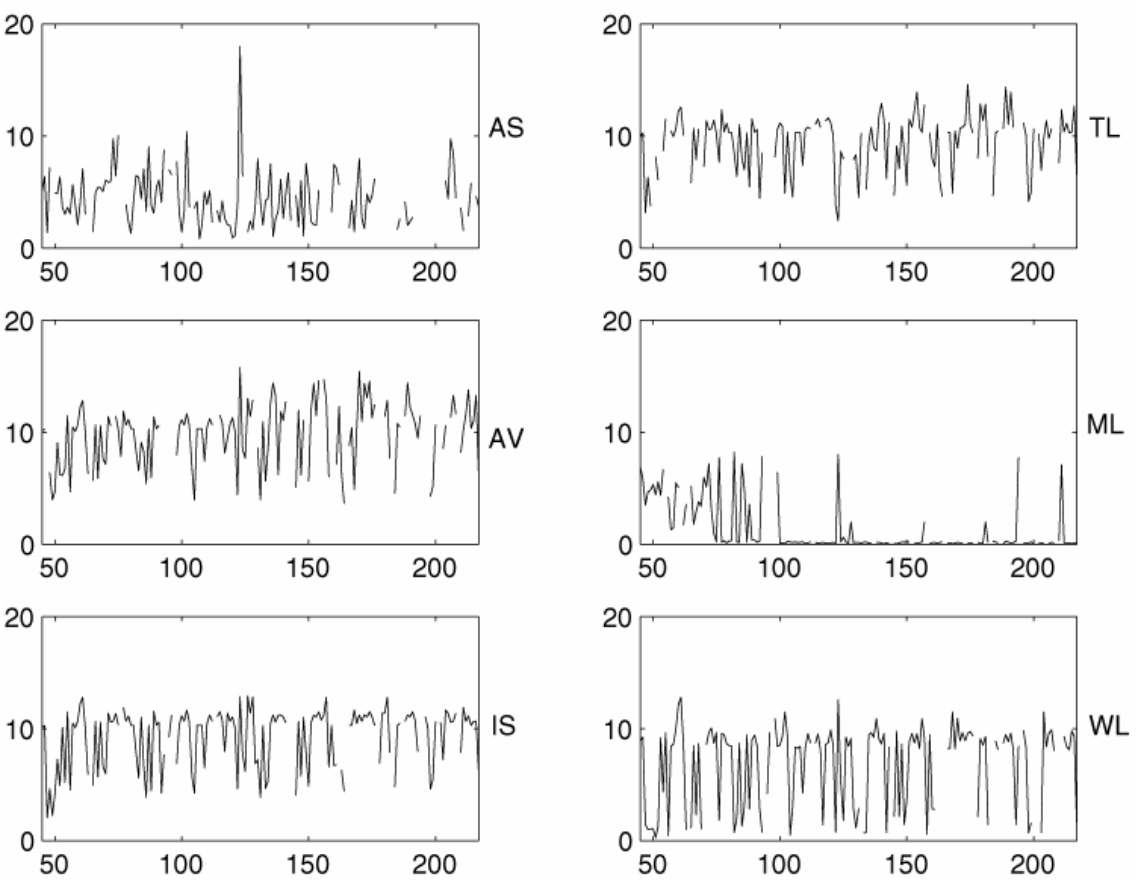


**Figure 41:** Error between practical valve settings and those predicted by Table 3

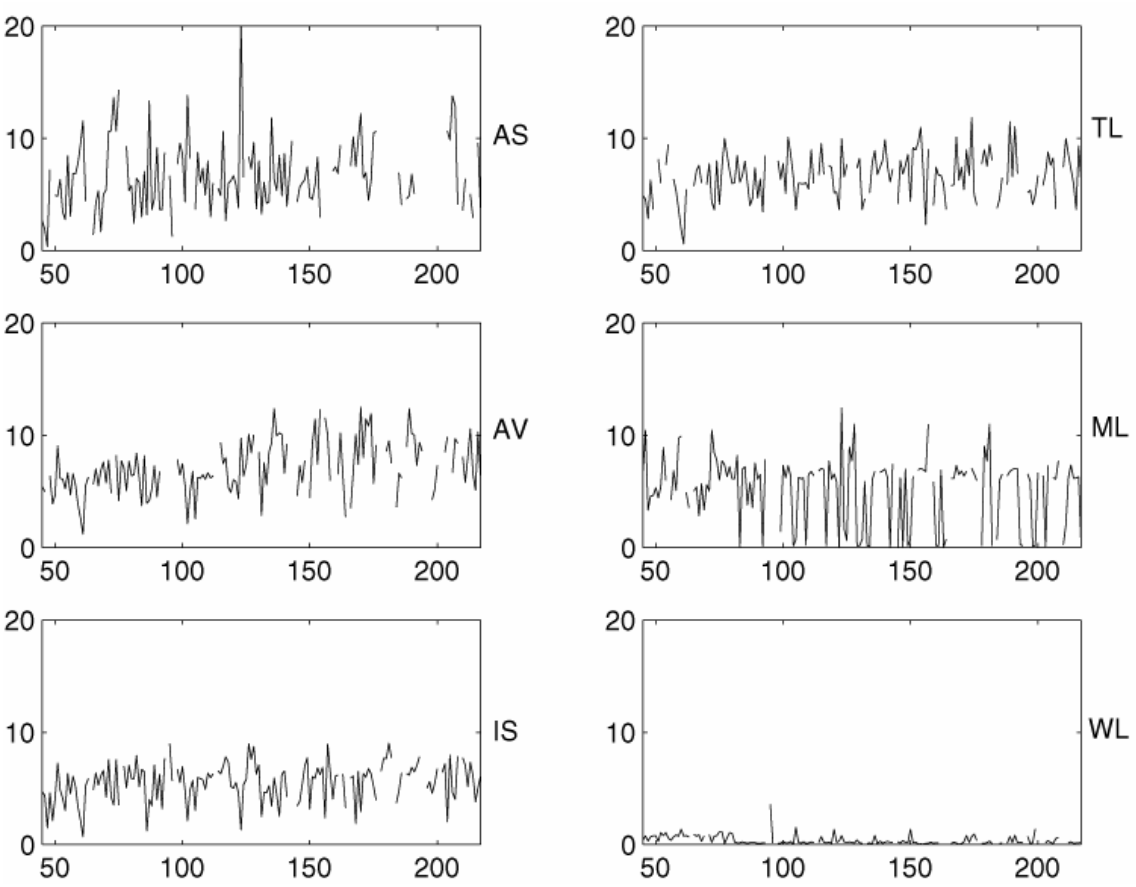


**Figure 42:** Error between practical valve settings and those predicted by Table 4

Second, a descriptive analysis was also conducted, and descriptive algorithms were extracted based the data of ML and WL. The two algorithms provided very good fits to the behavioural data of ML and WL, but poor fit to the data of other subjects. It should be noted that these two algorithms show a poor degree of adaptation to minimal effort criteria. Both ML and WL were not sensitive to the differences between trial modes, and they used strategies that required a larger number of EIPs (Elementary Information Processes) than was necessary. Thus, while they were able to perform the task successfully, they did not do so in a cognitively economic fashion.

Third, there seems to be some connection between performance and the feedwater control configuration the subjects used to control DURESS II. AV, IS and TL, who exhibited signs of adaptation to information processing limitations, were generally the most proficient performers, while the less proficient subjects failed to show these signs.

These findings from the analysis of the tuning task are quite consistent with what we have already obtained from the analysis of startup tasks (Yu et al., 1997a). This suggests that the control configuration the subjects used was independent of the tasks they were required to complete.

## 4. ANALYSIS OF FAULT MANAGEMENT

The previous section detailed our efforts in analysing the tuning tasks of the six-month longitudinal experiment conducted for JAERI. The aim of that work was to understand how subjects achieved the system goals under a perturbation of one of these goals, and how they adapted to these perturbations. A second type of perturbation that was investigated in the longitudinal experiment was the perturbation of system structure in the form of unanticipated system faults. In this section, we will consider subjects' adaptation to the perturbation of system structure more closely.

In the longitudinal experiment, subjects were periodically made to complete trials that contained faults. These faults were unanticipated by the subjects as they did not know on which trial or at what point in a trial a fault would occur. In all trials, subjects were to be prepared for faults, which they were asked to detect, diagnose, and compensate for so that they could achieve the system goals despite the perturbation of system structure. Since these tasks involved a new and partially unknown system structure, they were more cognitively challenging than normal trials. A study of these fault tasks will reveal how the subjects handled the faults and how they adapted to the changed system structure. This study should also lead to a deeper insight into subjects' knowledge of system structure and function.

### 4.1 Fault Management Trials

Before discussing our new analyses of fault trials, the nature of these trials will be briefly reviewed. For more details, see Christoffersen et al. (1994).

After the introductory session of the experiment, fault trials were periodically presented to the subjects. Subjects were not told what faults would occur or when they would occur. A number of types of faults were built into the DURESS II microworld, all of which are believed to be representative of faults that would occur in an actual process control plant. There are two general classes of faults: routine and non-routine. Routine faults were designed to be relatively simple in nature, and were intended to be analogous to recurring failures in any industrial system. Three faults of this type were used in this experiment.

1. Valve blockage – one of the six input valves becomes blocked so that no water could pass through it.
2. Heater failure – one of the two heaters experiences a partial failure, so that the actual heating rate was a decreased or increased proportion of what would be expected given the heater setting.
3. Reservoir leak – one of the two reservoirs develops a leak so that the actual output flowrate is greater than what would be expected given the output valve setting.

Subjects experienced each type of these routine faults three times during the experiment.

The non-routine faults were designed to be more complex than the routine faults. They consisted of some combination of two routine faults in quick succession within one trial, but potentially spanning the startup and tuning tasks. Non-routine faults were designed to be analogous to the rare, unanticipated occurrences in a system that are difficult to compensate for. There were three different non-routine faults:

1. A reservoir leak combined with an increase in the temperature of the input water.
2. A valve blockage combined with an external heat source.
3. A heater failure combined with an increase in the temperature of input water.

Subjects experienced each of the non-routine faults once in the late stages of the experiment.

4.2 Previous Findings

Some interesting findings were obtained from the fault management trials (Christoffersen et al., 1994). These findings are briefly summarised below:

- For routine fault trials in terms of fault detection and compensation times, subjects in the P+F group seemed to have slightly better performance than those in the P group. There was a very strong advantage in favour of P+F group for diagnosis scores, which can be attributed to the effect of the richer and more systematic information displayed on the P+F interface.
- For non-routine trials, the performance of P+F group was better in the various phases of fault management. However, this result was moderated by individual differences.

In this research, we have used the state space measures to further detail the analysis of fault trials. Several of the measures developed for normal and tuning trials were identified as potentially useful for the analysis of routine fault trials and the non-routine fault trials. As our intent was to focus on fault management behaviour, the data used for each fault trial in the following analyses begins at the point the fault was introduced into the system. The results will

be presented in chronological order, organised by interface group, with comparisons across subjects, trials, and interface groups. For the sake of brevity, not all of the results will be included here.

4.3 State Space Analysis: Routine Fault Management

Routine fault trials contain only one of the three faults: valve blockage, heater failure, and reservoir leakage. All subjects experienced each type of faults three times in the course of the experiment. Results obtained from the analysis of nine routine fault trials are presented below.

**Trial 64**

This was the first fault trial for all subjects. During this trial, VA1 became blocked 4 minutes after the beginning of the trial, during the startup task. As a result, the flow through VA1 dropped to zero. All of the six subjects had already reached steady state when the fault occurred, and all six were able to detect the fault. In the P+F group, only IS diagnosed the fault and successfully completed the task. In the P group, ML and WL completed the task.

The graphs of distance to the goals are shown in Figure 43 and 44 for the P+F and P groups, respectively. All of the graphs start from zero, suggesting that at the time when the fault occurred, the system was within the goal boundaries. The graphs for AS, AV and TL were stopped before 5 minutes, indicating that their trials terminated prematurely due to the system blowing up. IS had the least divergence from the goal boundaries and a very short compensation time, while the successful subjects in the P group (ML and WL) seemed to have a more difficult time compensating.

The graphs of time to goal boundaries are shown in Figures 45 – 47 for each of the three successful subjects (IS, ML, WL) in this trial. Three interesting results stand out. First, as evidenced by the fact that the time to goal boundaries for all of these subjects never fluctuated from the maximum, none of these subjects adjusted the output flowrate despite the fact that there was an input problem with the system. Second, there were clear disturbances of temperature due to the blockage of input valve VA1. WL's temperature was the most severely disturbed, indicating that he had a very difficult time recovering the system. Third, since there were no disturbances on the lower reservoir for IS, he used a non-interactive control strategy in which VA1 did not feed reservoir 2.

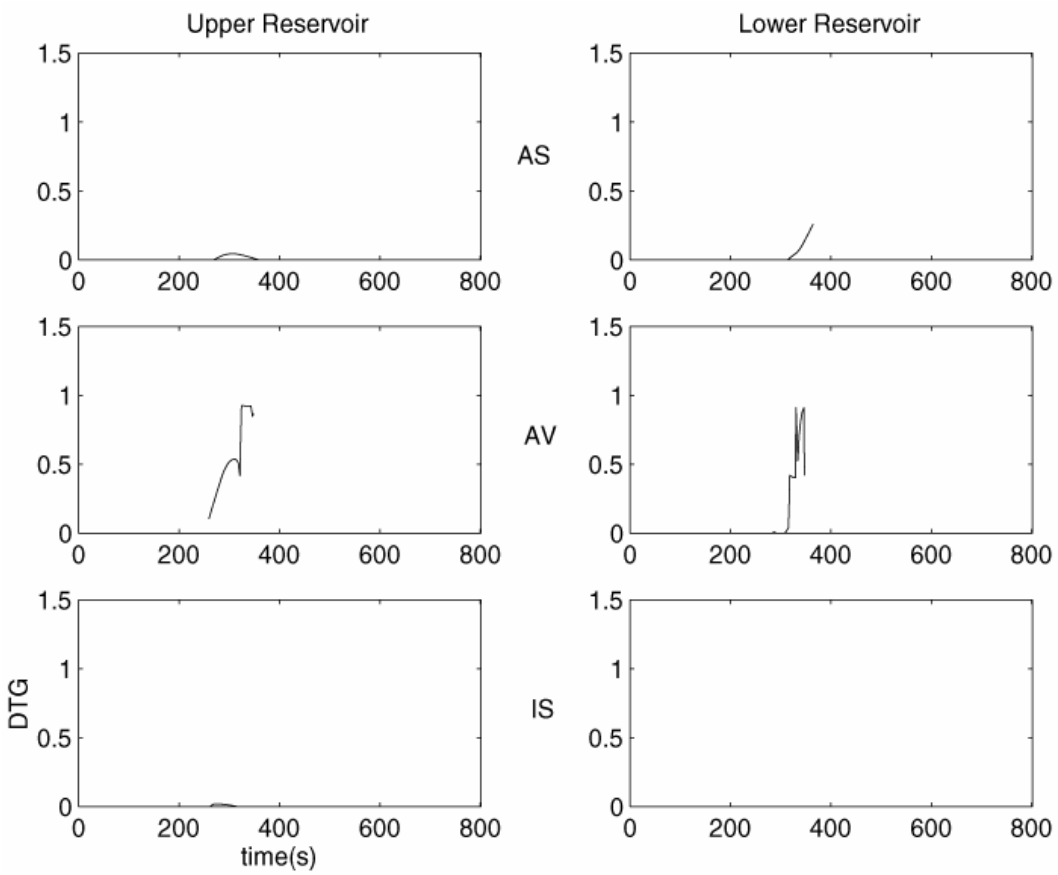


**Figure 43:** Distance to the goals for P+F group after the fault occurred (trial 64)

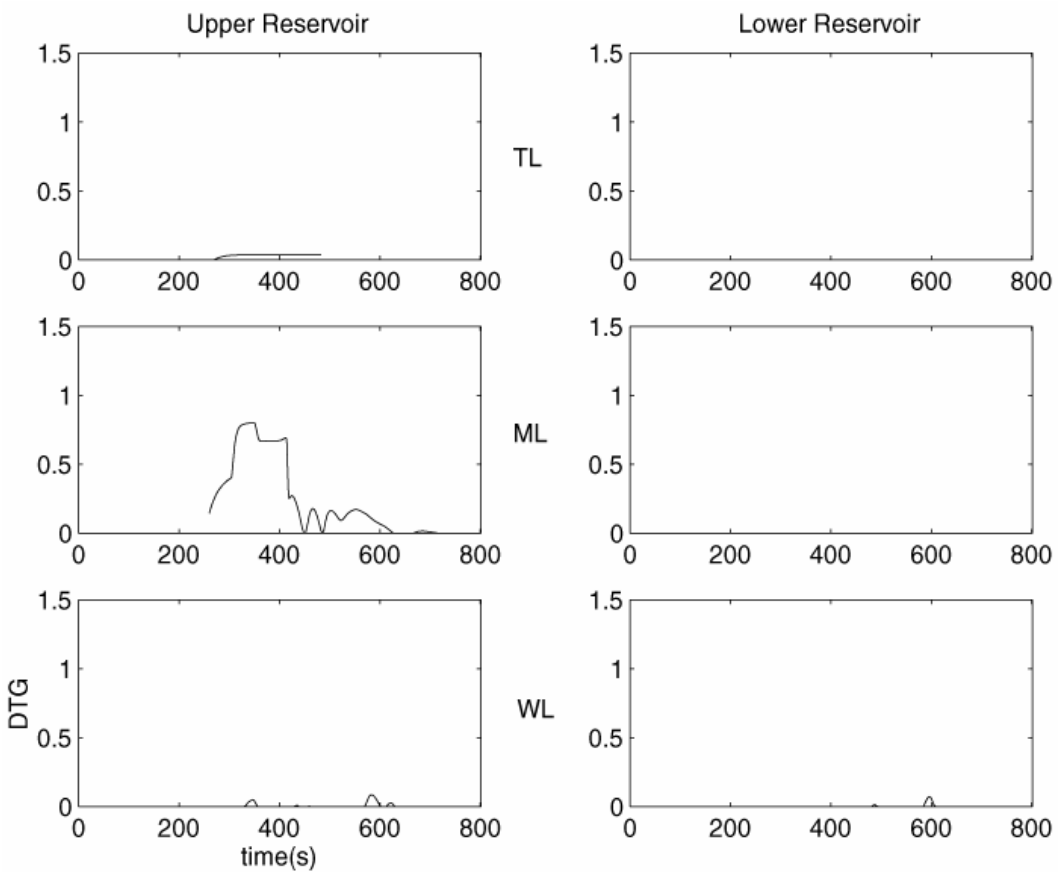


**Figure 44:** Distance to the goals for P group after the fault occurred (trial 64)

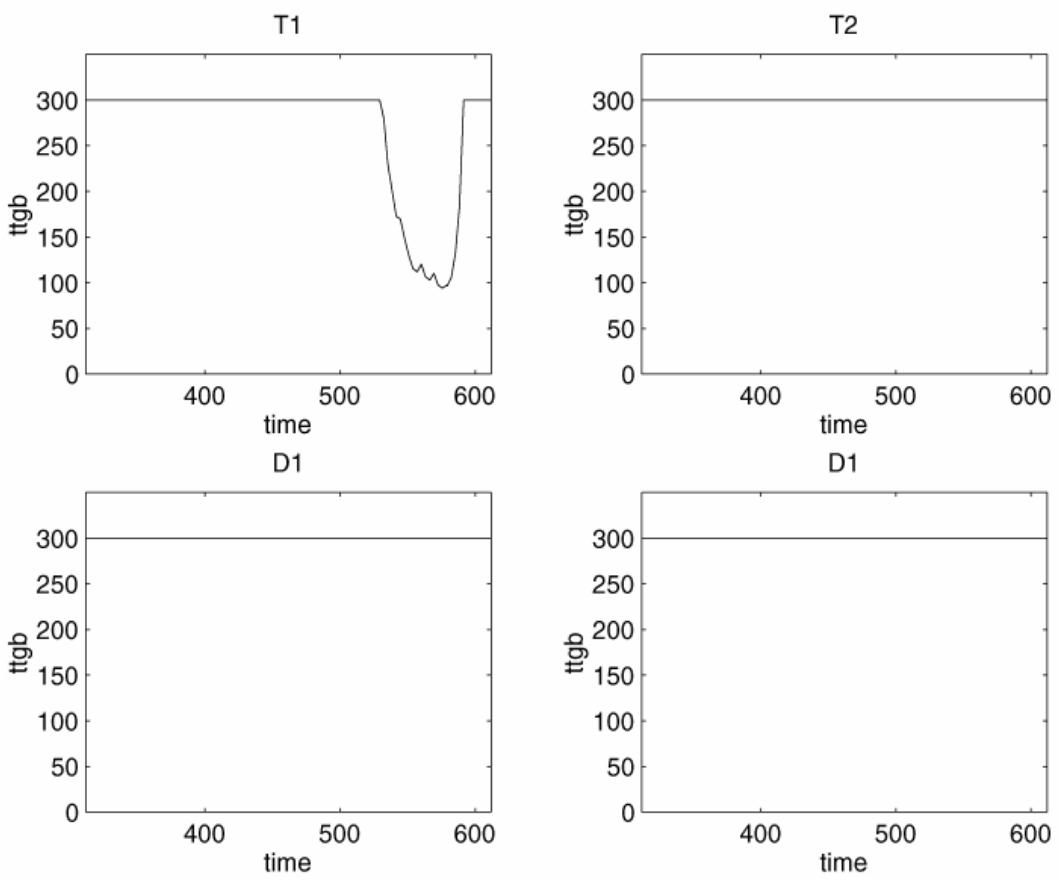


**Figure 45:** Time to goal boundaries for IS (trial 64)

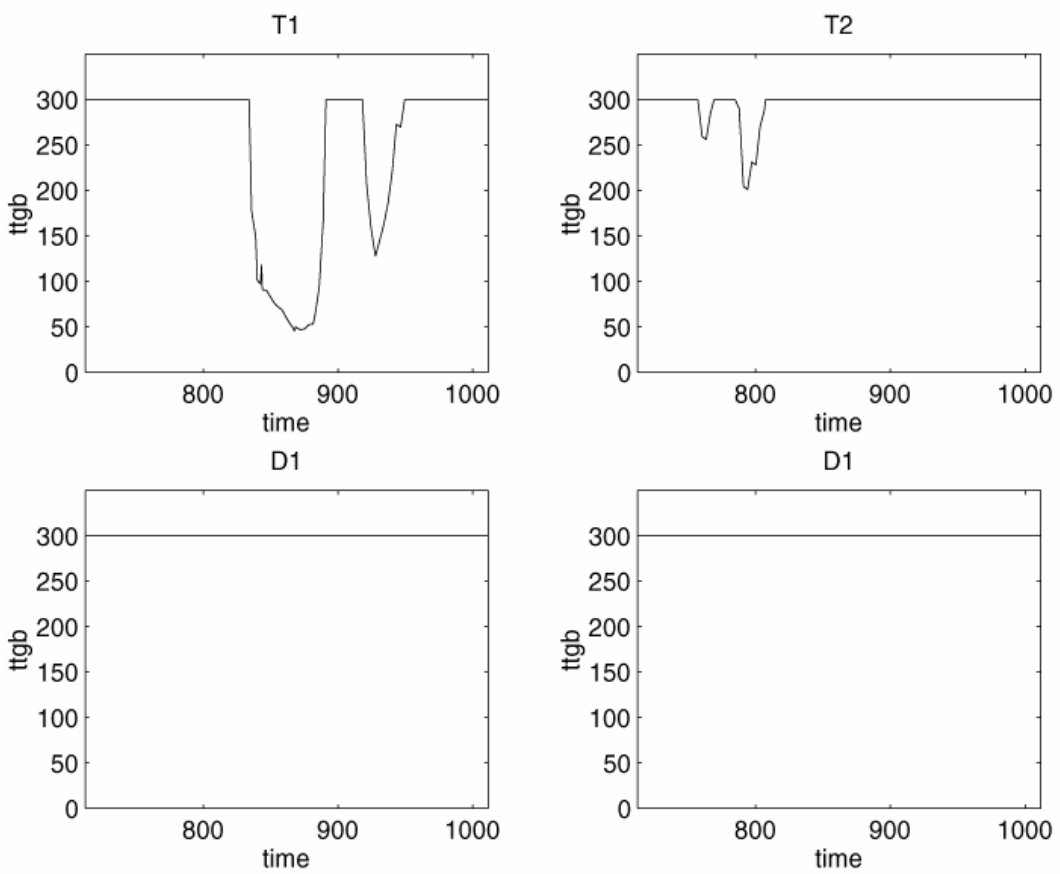


**Figure 46:** Time to goal boundaries for ML (trial 64)

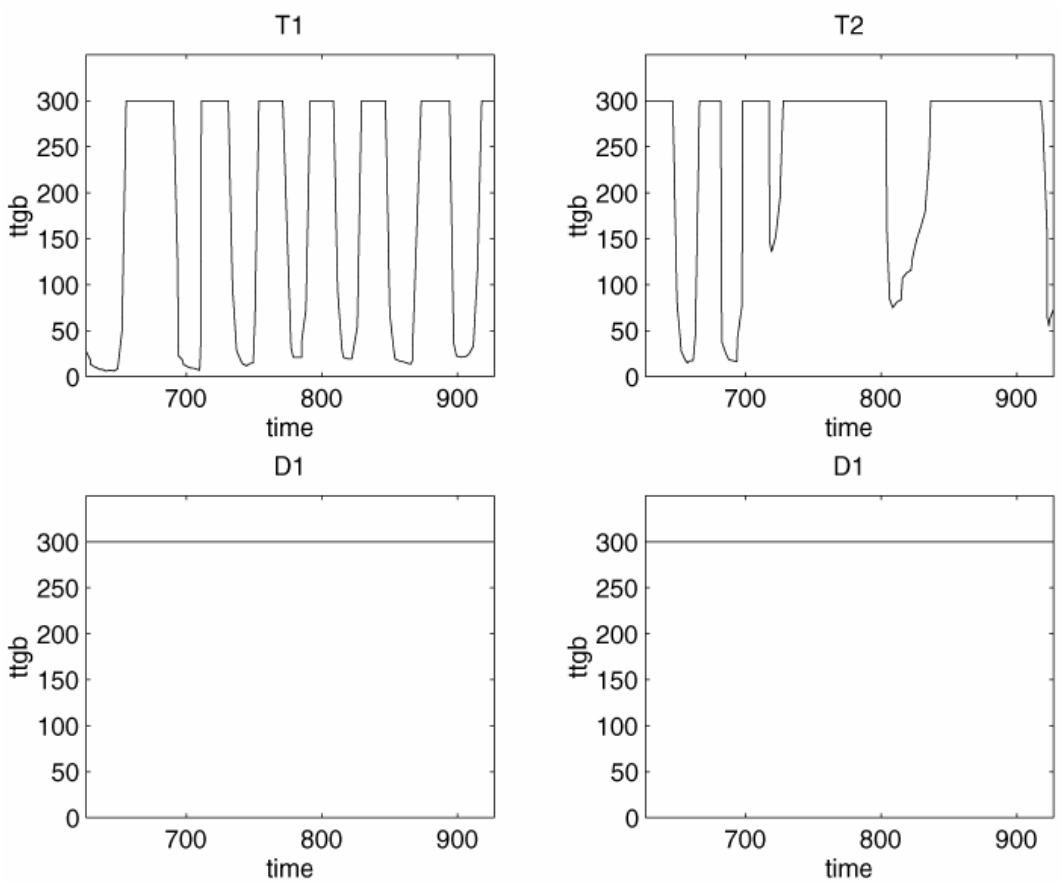


**Figure 47:** Time to goal boundaries for WL (trial 64)

Four quantitative measures were applied to the three successful subjects. The results are shown in Tables 5-8. We can see that IS had lower lengths of trajectories in the goal space, smaller area under distance to the goals, higher area under the curves of time to goal boundaries, and used a fewer number of control actions than those of ML and WL.

**Table 5:** Lengths of trajectories in goal space for IS, ML and WL (trial 64)

| Subjects | Reservoir 1 | Reservoir 2 | Reservoir 1 & 2 |
|---|---|---|---|
| IS | 0.1415 | 0. 0885 | 0.2007 |
| ML | 4.4645 | 0.2425 | 4.5013 |
| WL | 2.1290 | 1.7690 | 3.0589 |

**Table 6:** Area under distance to the goals for IS, ML and WL (trial 64)

| Subjects | Reservoir 1 | Reservoir 2 | Reservoir 1 & 2 |
|---|---|---|---|
| IS | 0.6698 | 0 | 0.6698 |
| ML | 118.1605 | 0 | 118.1605 |
| WL | 2.6931 | 1.0390 | 3.3798 |

**Table 7:** Area under curves of time to goal boundaries for IS, ML and WL (trial 64)

| Subjects | T1 | T2 | D1 | D2 |
|---|---|---|---|---|
| IS | 0.8952 | 0.8120 | 0.9760 | 0.9902 |
| ML | 0.8305 | 0.9110 | 0.9765 | 0.9794 |
| WL | 0.5814 | 0.7745 | 0.9799 | 0.9856 |

**Table 8:** Numbers of control actions for IS, ML and WL (trial 64)

| Subjects | IS | ML | WL |
|---|---|---|---|
| Number of actions | 31 | 54 | 142 |

In summary, from the results of the successful subjects in this trial, we can see that the successful P+F subjects had shorter fault detection, diagnosis, and compensation times than the

successful P subjects. Together with the previous findings based on the process tracing method (see Christoffersen et al., 1994), it seems that the P+F interface supports fault management better than the P interface, due to both the information content and form of the P+F interface.

**Trial 94**

This was the second fault trial for all subjects. During this trial, there was a gradual water leakage (4 unit/second) from the lower reservoir at 2 minutes after the beginning of the trial. This fault occurred during the startup task.

The fault occurred earlier in this trial than in the first fault, which made it more challenging than the first fault as some subjects were still trying to achieve the steady state at the time of the fault. In the P+F group, IS was the fastest at detecting the fault, and compensated quickly. Note that IS had already reached steady state when the fault occurred, and this greatly facilitated his handling of the fault. AS took much longer to detect the fault because he was busy stabilising the temperature goals at the time the fault occurred. He took some iterative control actions until he eventually diagnosed the root cause of the fault and compensated accordingly. AV detected the fault quickly, but only diagnosed the root cause after a lengthy period of time and failed to compensate properly. Only AS and IS completed the task in this group.

In the P group, both TL and WL were trying to stabilise the temperature when the fault occurred. As this was the case, they took quite a while to detect the fault. It is interesting to note that these two subjects compensated for the fault using a trial-and-error strategy as they did not know the nature of the fault. ML was fastest at detecting the fault in this group, but he was not able to diagnose the fault or compensate for it. Only TL and WL completed the task in this group.

The graphs of trajectories in goal space for the P+F and P groups are shown in Figure 48 and 49, respectively. For subjects who had not yet reached steady state, the curves for a successful trial should start somewhere in the two-dimensional space and end at (1,1). Trajectories for the subjects who reached steady state before the onset of the fault begin at (1,1). Two findings stand out. First, AV and ML failed to stabilise the system, and were not able to complete the trial. Second, AS and WL had some trouble compensating since their curves went far away from the goal point (1,1), while IS and TL seemed to compensate easily. The lengths of these trajectories are calculated in Table 9 for the successful subjects. These numbers support the

second finding (above), as we can see that the lengths of trajectories in goal space for IS and TL are significantly shorter than those of AS and WL.

**Table 9:** Lengths of TGS for AS, IS, TL and WL (trial 94)

| Subjects | Reservoir 1 | Reservoir 2 | Reservoir 1 & 2 |
|---|---|---|---|
| AS | 0.2723 | 4.9447 | 5.0429 |
| IS | 0.3020 | 0.5917 | 0.7324 |
| TL | 0.2105 | 0.4800 | 0.5856 |
| WL | 2.2307 | 3.0965 | 4.4699 |

The graphs of distance to the goals are shown for the two groups of subjects in Figures 50 and 51, respectively. These graphs are very informative as well. They are consistent with the findings we made in the graphs of trajectories in goal space. In addition, we can see the time when the subjects detected the fault. For example, in the P+F group, both AV and IS detected the fault quickly, and AS was the slowest to detect the fault. By calculating the area under these curves, we can see that the areas of IS and TL are significantly smaller than those of AS and WL.

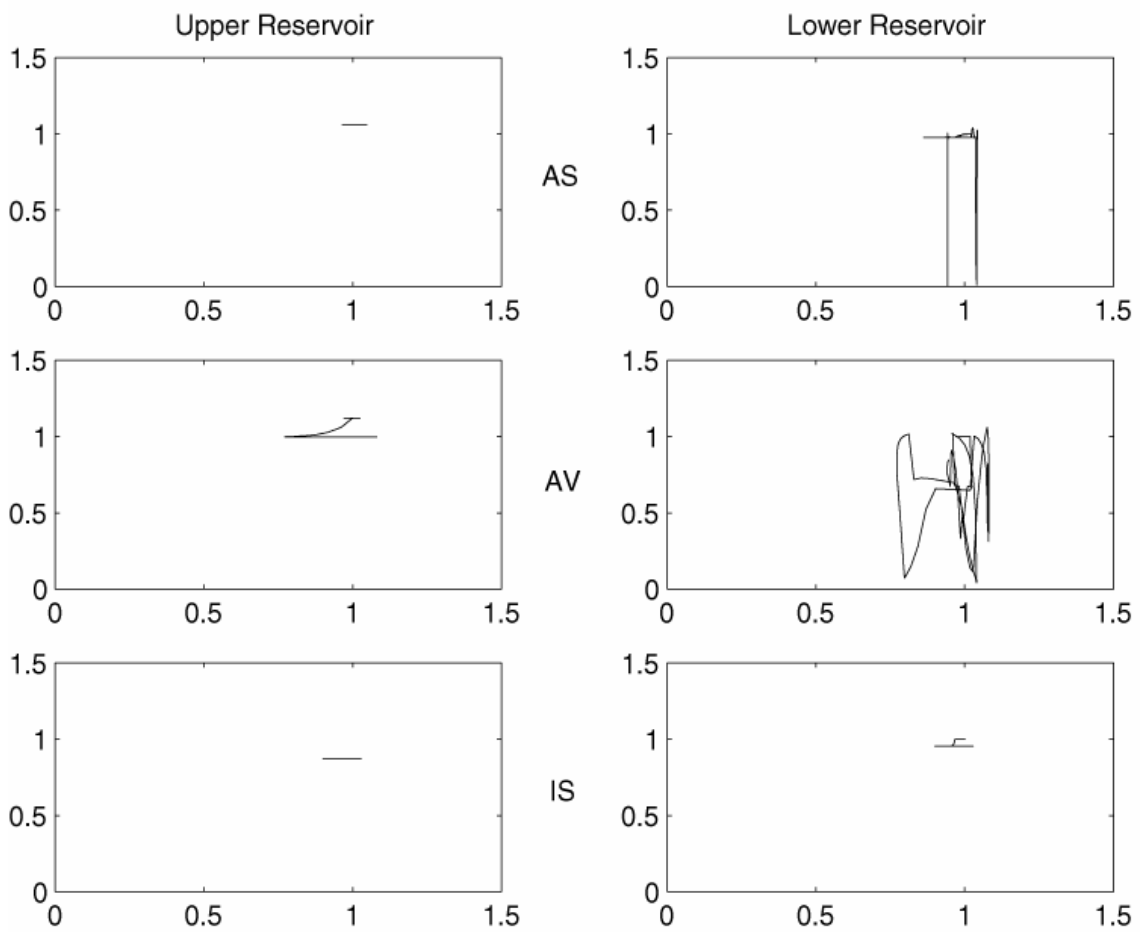


**Figure 48:** Graphs of trajectories in goal space for P+F subjects (trial 94)

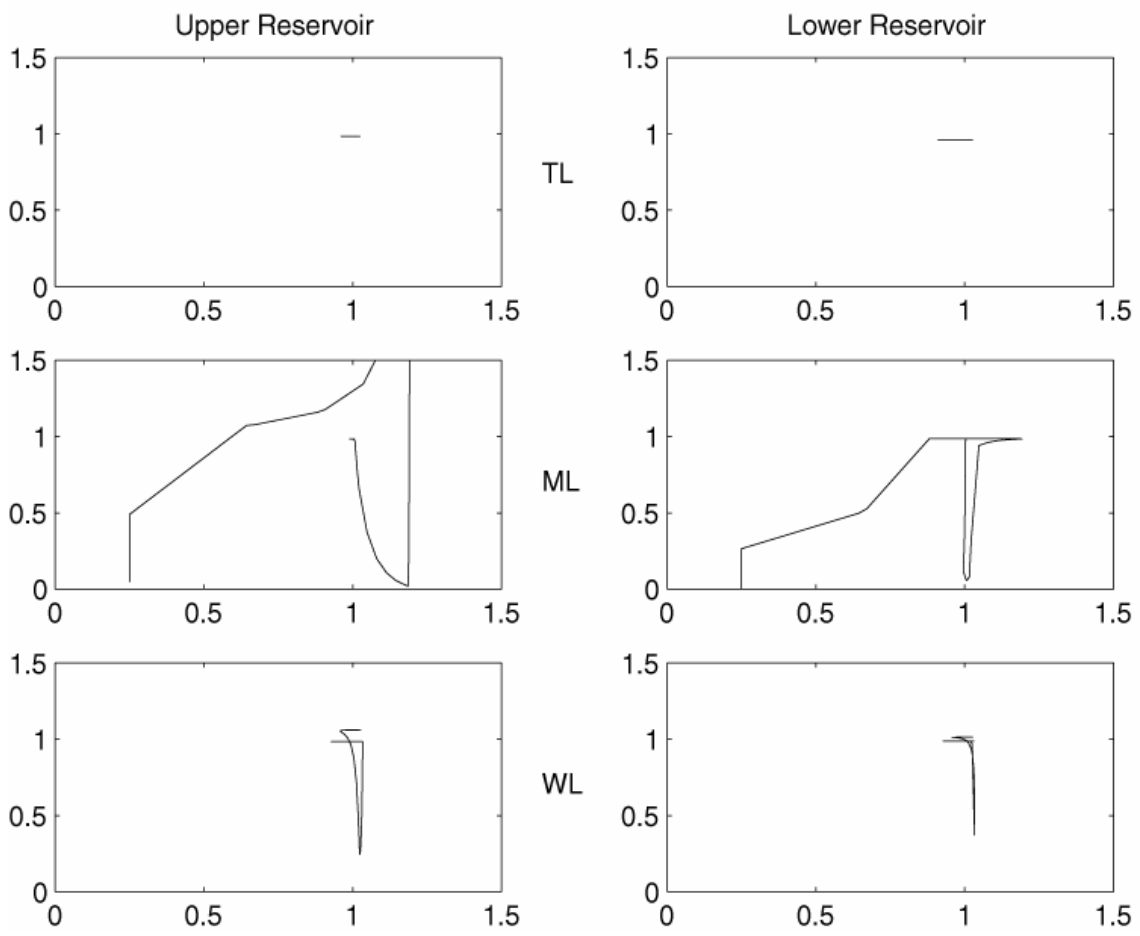


**Figure 49:** Graphs of trajectories in goal space for P subjects (trial 94)

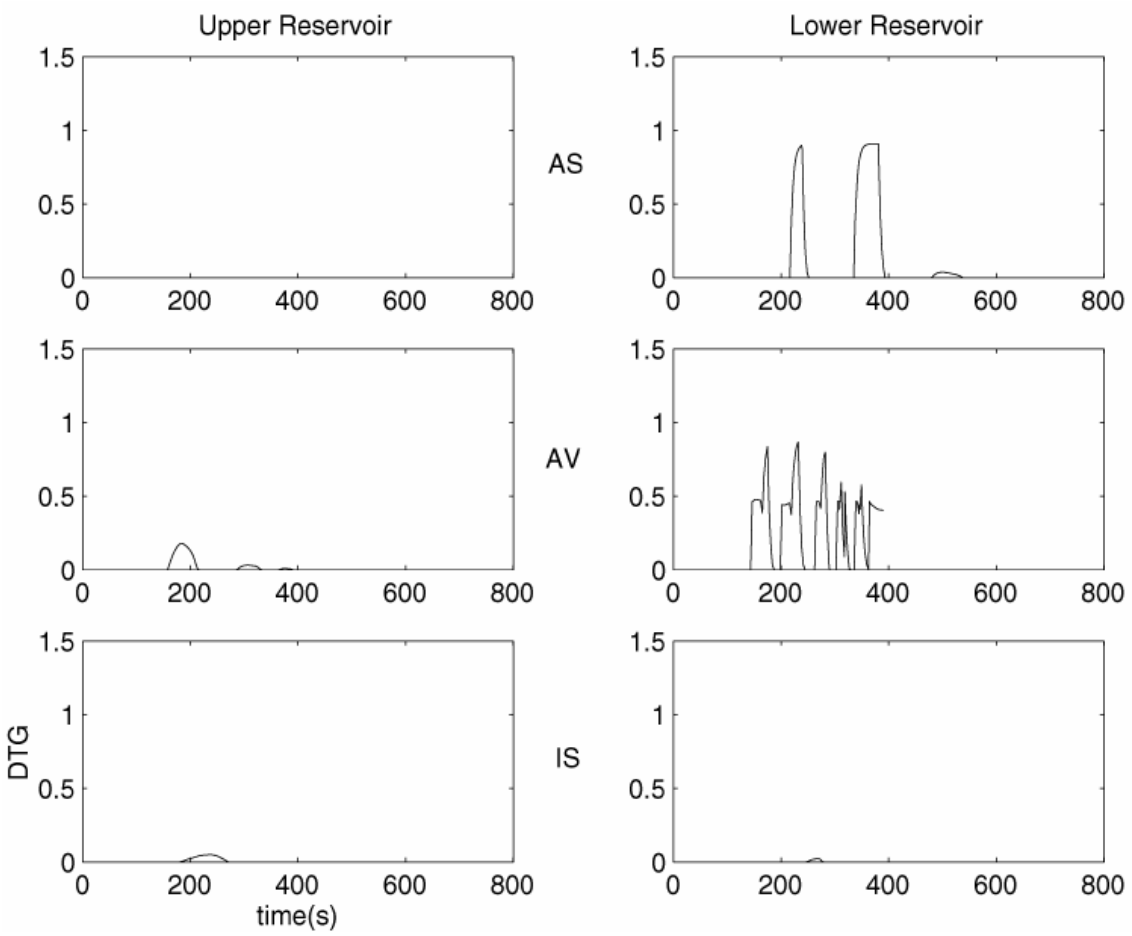


**Figure 50:** Graphs of distance to the goals for P+F subjects (trial 94)

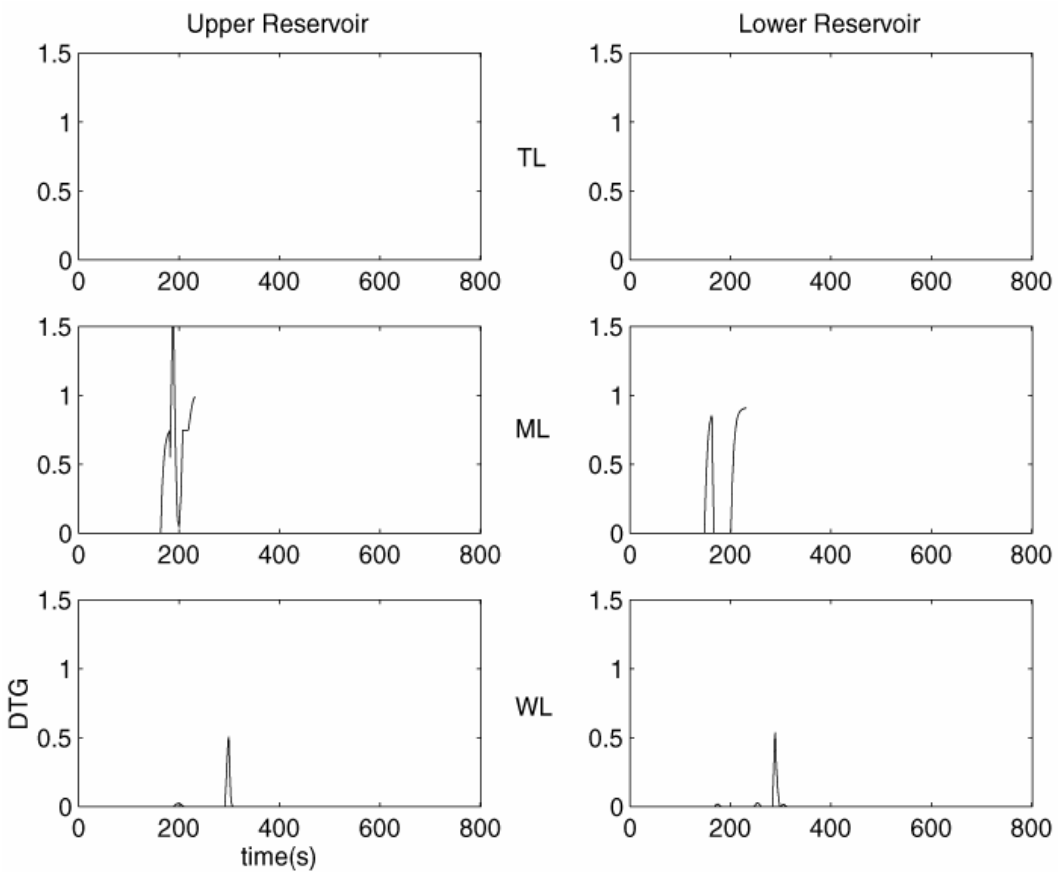


**Figure 51:** Graphs of distance to the goals for P subjects (trial 94)

For brevity, we present only the graphs of time to goal boundaries for WL in Figure 52. The curves show that WL had difficulty meeting the temperature goals. It is interesting to note that he also used the output valves to compensate after the fault occurred (indicated by lower valve of time to goal boundaries for D1 and D2).

The area under time to goal boundaries was calculated for subjects who successfully completed the task. The results are shown in Table 10. There is little difference between subjects within each group, but there seems to be some difference between interface groups in favour of the P+F group. The values of time to goal boundaries of P+F subjects were larger than those of P subjects, indicating that the P+F subjects eventually achieved a greater stability than the P subjects.

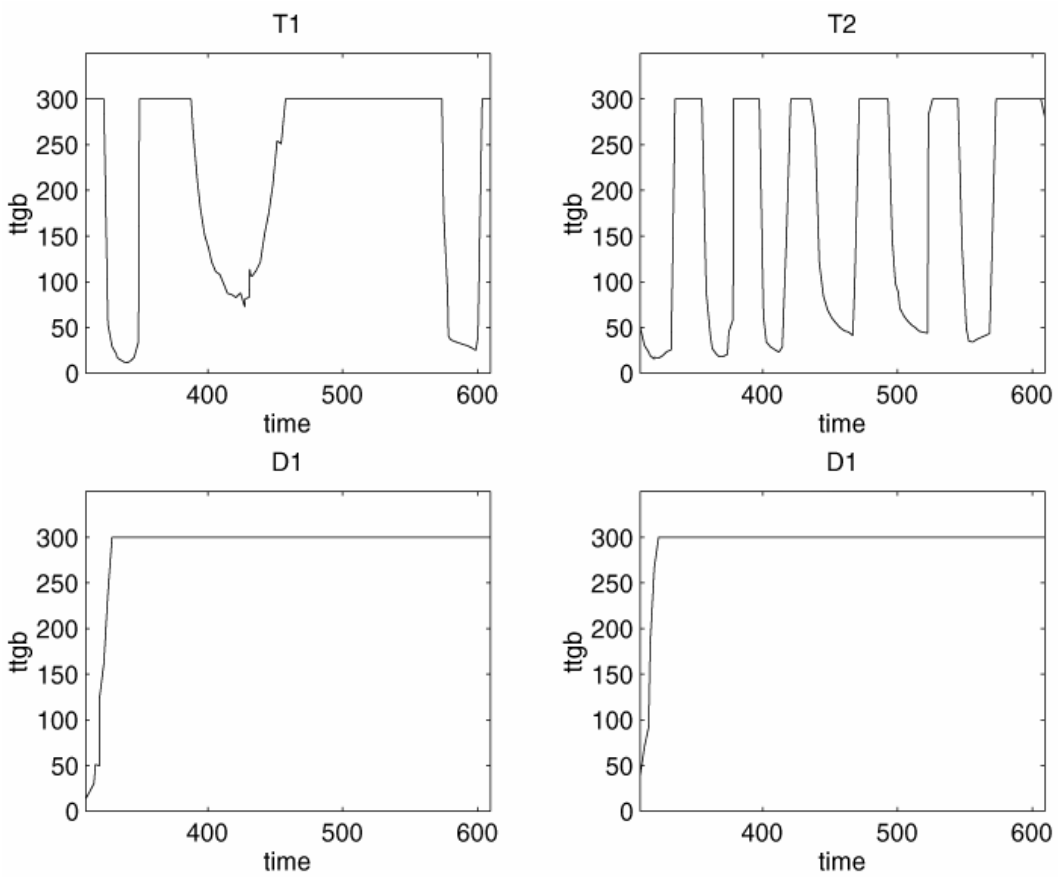


**Figure 52:** Graphs of time to goal boundaries for WL (trial 94)

**Table 10:** Area under time to goal boundaries for AS, IS, TL and WL

| Subjects | T1 | T2 | D1 | D2 |
|---|---|---|---|---|
| AS | 0.8934 | 0.7077 | 0.9696 | 0.9575 |
| IS | 0.8026 | 0.8155 | 0.9859 | 0.9746 |
| TL | 0.7554 | 0.7985 | 0.9657 | 0.9800 |
| WL | 0.6845 | 0.5939 | 0.9497 | 0.9483 |

The numbers of actions made by the subjects after the onset of the fault are summarised in Table 11. There seems to be very little difference between the interface groups. However, the subjects who spent more time on compensating used more actions, such as AS and WL.

**Table 11:** Numbers of control actions for IS, ML and WL

| Subjects | AS | IS | TL | WL |
|---|---|---|---|---|
| Number of actions | 91 | 46 | 34 | 83 |

This second fault was about as difficult to detect, diagnose, and compensate for as the first fault. However, as it occurred earlier in the trial, slow performers who were not yet able to stabilise the system before the onset of the fault suffered. While still struggling to stabilise the system, they had to detect, diagnose, and compensate for this fault in order to get the system under control. Several interesting findings were obtained in the analysis. First, the P+F subjects had a greater tendency to explore the nature of the fault, and were more accurate in detection and diagnosis. P subjects paid less attention to determining the nature of the fault, even if they were eventually able to compensate for it. This may be caused by the difference in interface between the two groups. The P+F interface displays rich information to facilitate fault detection, while the P interface lacks such information, so that the subjects must passively rely on action-response effects. Secondly, there are some interface differences in time to goal boundaries in favour of the P+F subjects. For the rest of the measures, although the individual differences are clear, it is hard to isolate any interface differences. Finally, it should be noted that slow performers were at a disadvantage with this fault as they effectively experienced a more severe fault than the subjects who were able to stabilise the system before the onset of the fault.

**Trial 97**

This was the third fault trial for all subjects. One minute after the start of the trial, the heat transfer from H1 dropped exponentially to half of the heater setting. The heater remained operational throughout the trial, but would only produce heat at half the original rate. Note that this fault occurred earlier in the trial than the previous two faults, and caught more subjects prior to stabilising the system.

In the P+F group, all of the subjects detected, diagnosed, and compensated for the fault. IS was the fastest at detecting the fault, while AS was the slowest. In the P group, TL and ML detected the fault faster than WL did. Across interface groups, the average time to detect the fault for the P subjects was smaller than for the P+F subjects. Using the process tracing method,

Christoffersen et al. (1994) revealed that the P+F subjects detected the fault based on the information provided by the interface, while the P subjects relied on the rules they developed.

To gauge the subjects' performance, we plotted their trajectories in goal space. For brevity, we only present here the quantitative results of these measures. The results are shown in Table 12. As they took the shortest time to compensate for the fault, AV and TL were the best within the respective groups. Across interface groups, the P+F group generally performed better than the P group.

**Table 12:** Lengths of trajectories in goal space for AS, IS, TL and WL (trial 97)

| Subjects | Reservoir 1 | Reservoir 2 | Reservoir 1 & 2 |
|---|---|---|---|
| AS | 0.6965 | 0.3053 | 0.8262 |
| AV | 0.6841 | 0.2330 | 0.7629 |
| IS | 1.1715 | 0.7317 | 1.4958 |
| TL | 0.4314 | 1.1300 | 1.4286 |
| ML | 8.9669 | 1.6674 | 9.6016 |
| WL | 1.2062 | 1.4973 | 2.3328 |

The graphs of distance to the goals are shown in Figures 53 and 54. Here, too, there is a clear interface effect. P+F subjects generally had smaller distance to the goals over time than the P subjects. These graphs also show compensation time. The P+F subjects took similar amounts of time to compensate. In the P group, TL took more time to compensate than ML and WL.

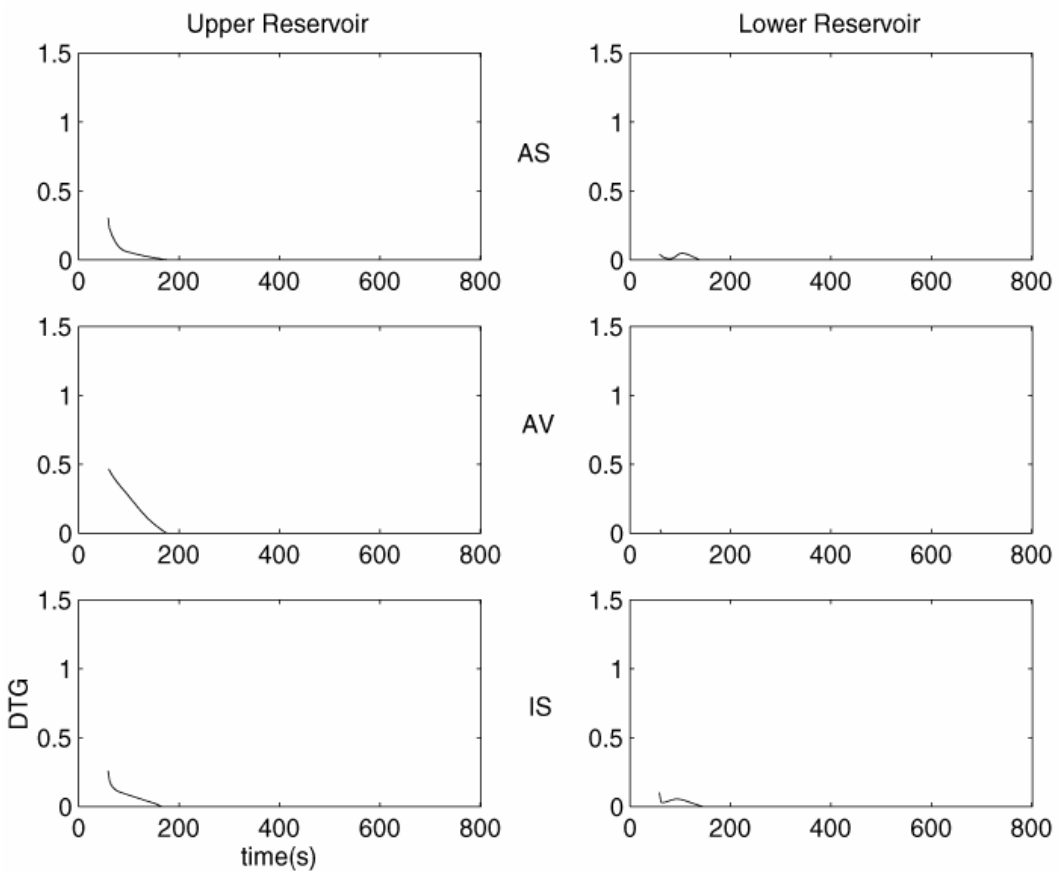


**Figure 53:** Graphs of distance to the goals for P+F group (trial 97)

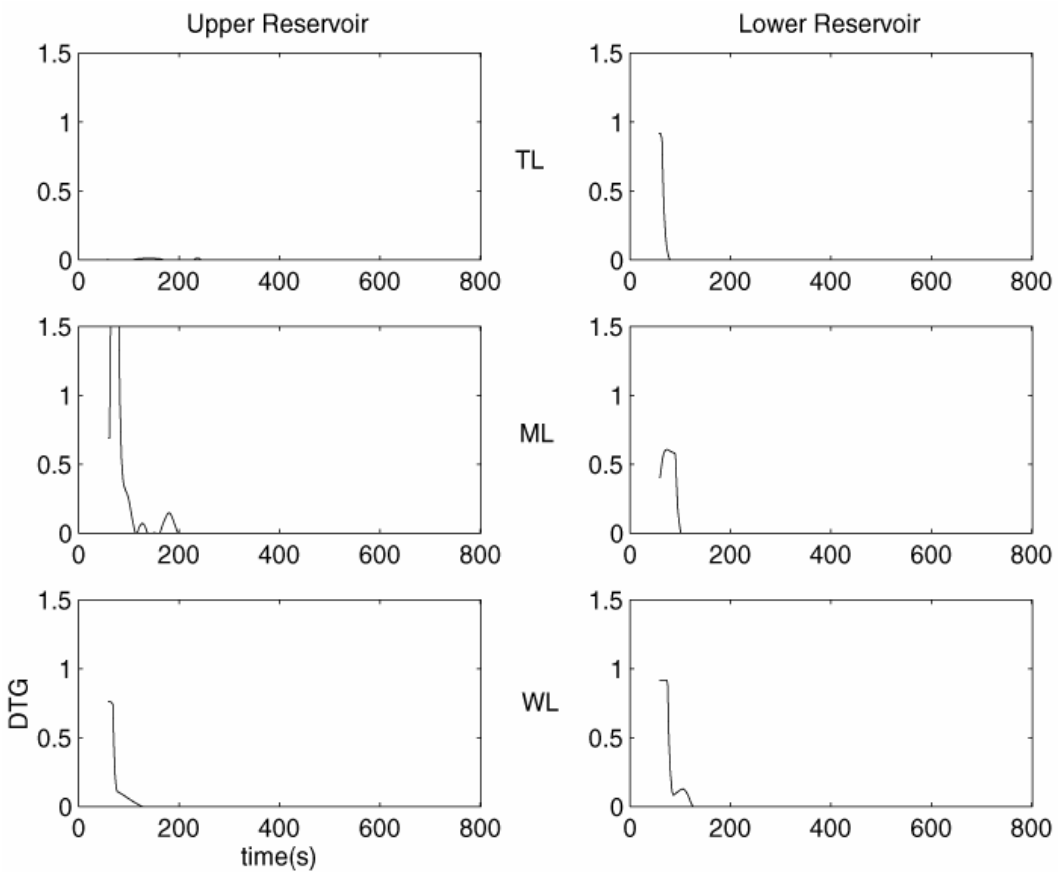


**Figure 54:** Graphs of distance to the goals for P group (trial 97)

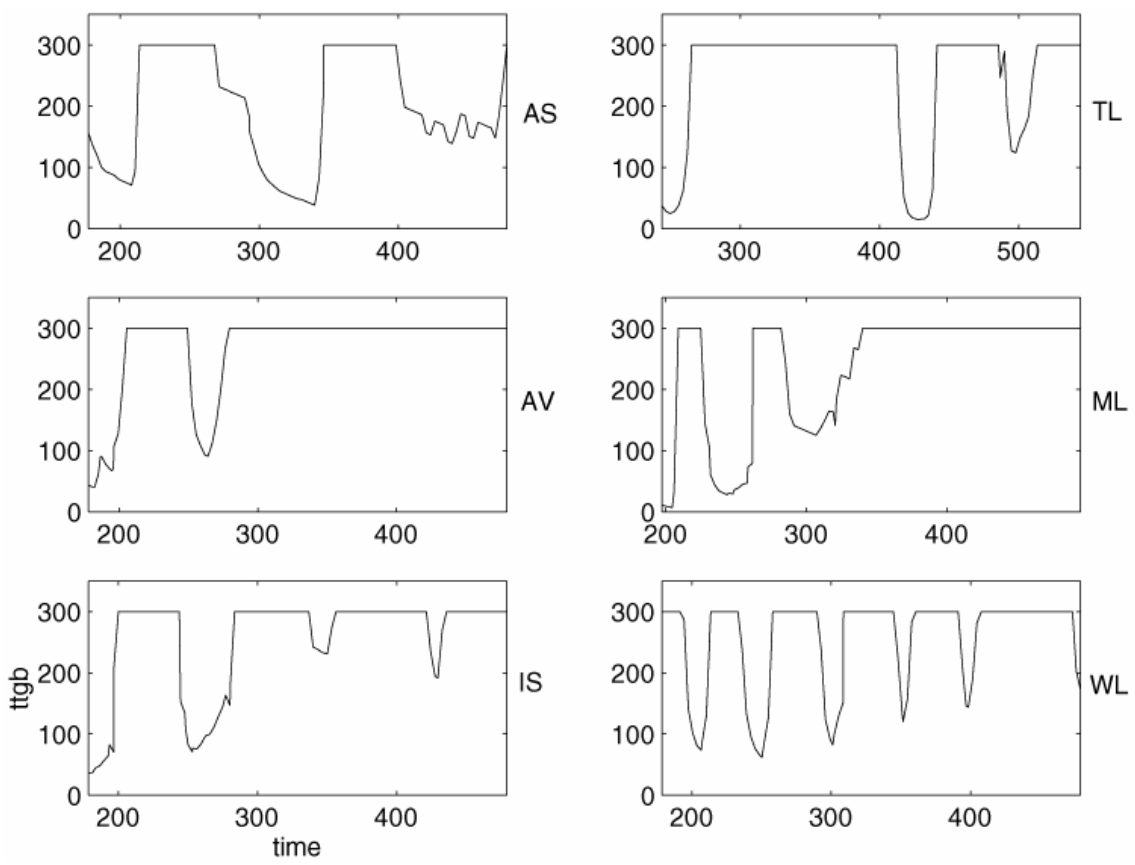


**Figure 55:** Graphs of time to goal boundaries for T1 (trial 97)

For the graphs of time to goal boundaries, we will focus on the temperature of the upper reservoir only, since this fault directly impacted only this reservoir. The results are shown in Figure 55. Again, AV and TL are the most stable in upper reservoir temperature in their respective groups. The area under these curves was calculated (see Table 13). AV had the most stable upper reservoir temperature.

**Table 13:** Area under the curves of time to goal boundaries for T1 (trial 97)

| Subjects | AS | AV | IS | TL | ML | WL |
|---|---|---|---|---|---|---|
| ATTGB | 0.6631 | 0.8944 | 0.8231 | 0.8451 | 0.7977 | 0.8236 |

A count of the numbers of actions the subjects made after the fault is presented in Table 14. AV and TL used the fewest actions in their groups. Across interfaces, the P subjects generally used

fewer actions after the fault occurred. This might be because these subjects detected the fault more quickly than the P+F subjects.

**Table 14:** Numbers of control actions

| Subjects | AS | AV | IS | TL | ML | WL |
|---|---|---|---|---|---|---|
| Number of actions | 40 | 38 | 43 | 28 | 34 | 41 |

In summary, this was the most challenging fault among the three faults introduced so far for two reasons. First, this fault involved one of the heaters. Since heaters have a longer time constant than flows, heater faults are more difficult to detect, diagnose, and compensate for. Second, this fault occurred early in the trial when the subjects were still busily manipulating the system components to bring DURESS to a steady state. Given that the subject's systems had not yet reached steady state but were in a state of flux, the change initiated by this fault was more difficult to detect.

Several important results emerge from the analysis. The most interesting one is that P subjects detected the fault faster than the P+F subjects. Recall the findings obtained by Christoffersen et al. (1994), the P subjects started using the rules they developed to detect the fault, while the P+F subjects relied on the information provided by the interface they were given. The numbers of actions used after the fault also supports this finding. If we compare the best two subjects in each group, however, the P+F subjects kept the temperature of the upper reservoir more consistent.

**Trial 113**

This was the fourth fault trial for all subjects. During this trial, the lower reservoir gradually started leaking at 7 units/second 4 minutes into the trial. This was the second time the subjects encountered this type of fault, but this occurrence was more severe than the first. Again, this fault occurred during the startup task. When it occurred, all subjects were maintaining the system in a steady state, and all of them detected, diagnosed, and compensated for the fault successfully. The compensation time for the P+F subjects was generally shorter than that of the P subjects.

The graphs of trajectories in goal space are shown in Figures 56 and 57 for the P+F and P groups, respectively. It seems that both AS and IS in the P+F group had some

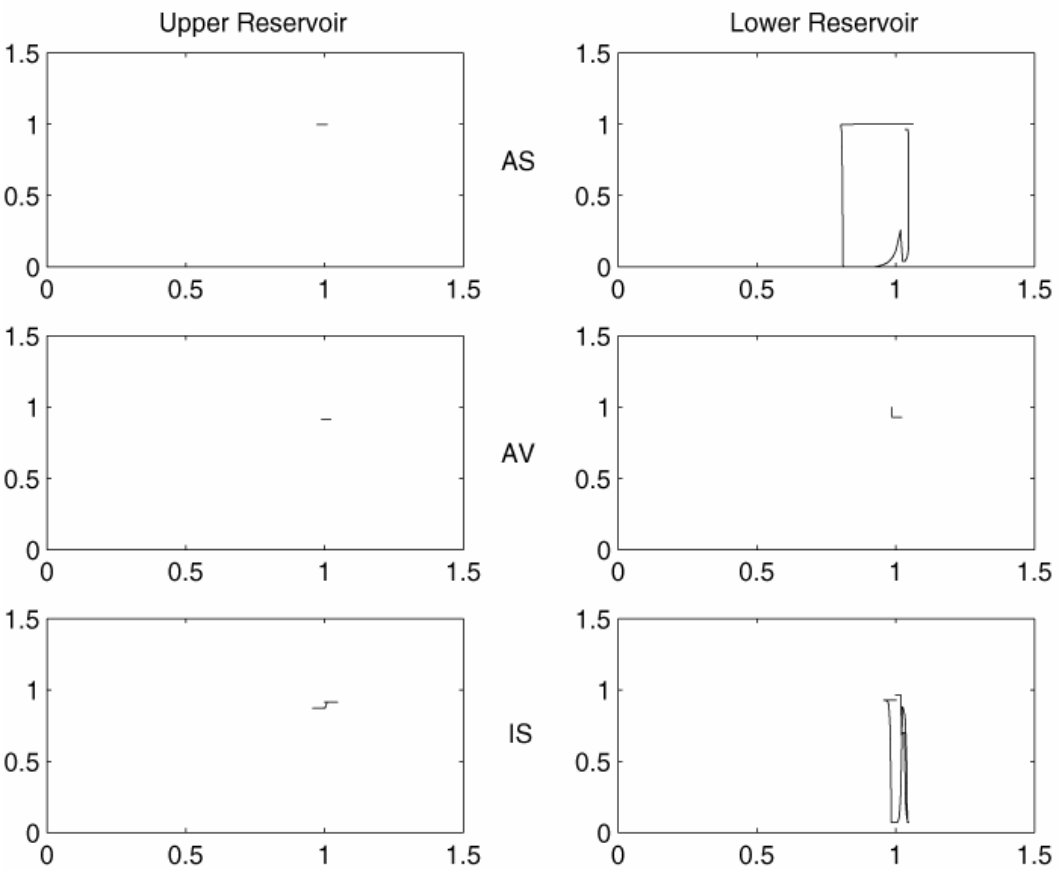


**Figure 56:** Graphs of trajectories in goal space for P+F group (trial 113)

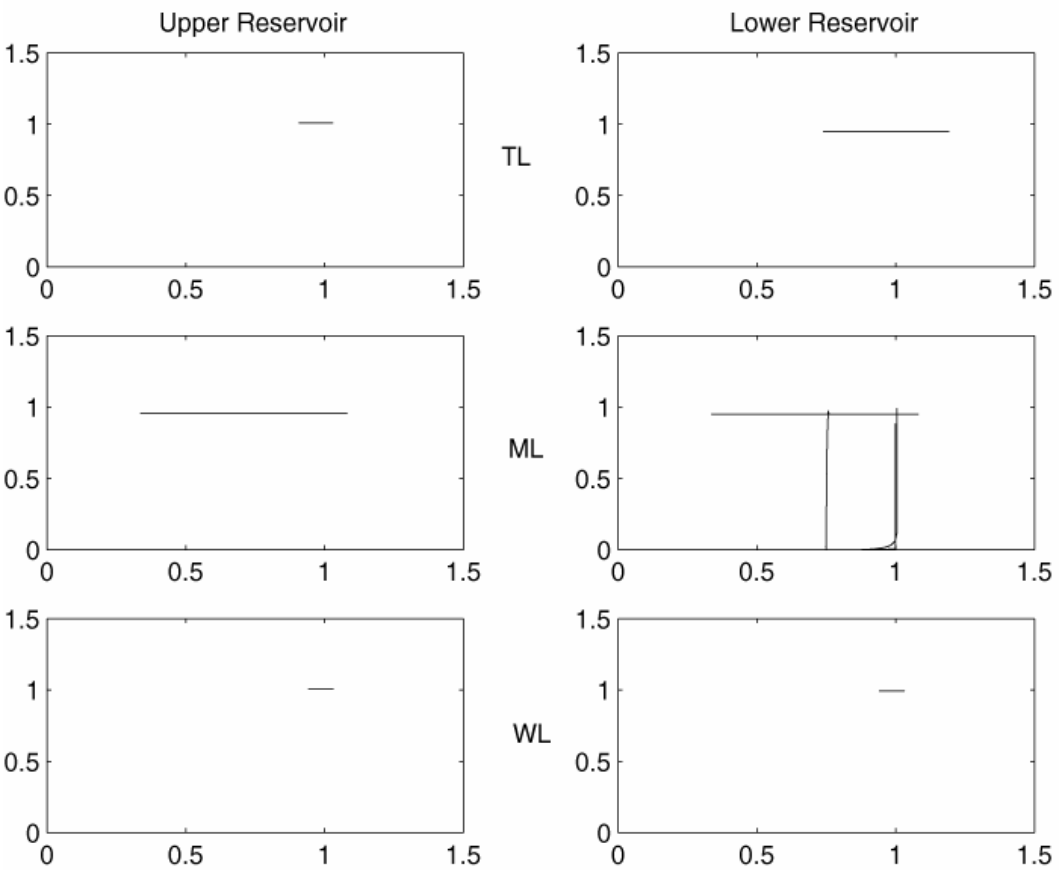


**Figure 57:** Graphs of trajectories in goal space for P group (trial 113)

rouble compensating for this fault, and that AV was the best among the three. In the P group, WL (instead of TL) was the best performer. ML, as in other trials, was the worst performer in this group. If we consider the length of these trajectories, AV's was shortest while ML's was longest.

**Table 15:** Lengths of trajectories in goal space (trial 113)

| Subjects | Reservoir 1 | Reservoir 2 | Reservoir 1 & 2 |
|---|---|---|---|
| AS | 0.1862 | 3.2789 | 3.3222 |
| AV | 0.0385 | 0.4633 | 0.4690 |
| IS | 0.2207 | 4.5934 | 4.6514 |
| TL | 0.8030 | 2.0785 | 2.3816 |
| ML | 1.8210 | 6.1329 | 7.2389 |
| WL | 0.4145 | 0.9950 | 1.1382 |

The graphs of distance to the goals for this trial were also plotted for every subject. Similar findings emerge from these graphs as were seen from the graphs of trajectories in goal space. For brevity, only the areas under the curves of distance to the goals are included in this report (see Table 16). An interesting finding from these data is that the two groups of subjects used different control strategies. The output demand for the lower reservoir was 7 units/second in this trial, and the reservoir itself was also leaking at 7 unit/second. This means that if the subjects wanted to balance the mass in the lower reservoir, then they would have to adjust both of the two feedwater streams, which implies that they had to perturb the upper reservoir when they compensated. While it seems that the P subjects did so, the P+F people did not seem to be aware of this.

**Table 16:** Area under distance to the goals (trial 113)

| Subjects | Reservoir 1 | Reservoir 2 | Reservoir 1 & 2 |
|---|---|---|---|
| AS | 0 | 155.9688 | 155.9688 |
| AV | 0 | 1.1521 | 1.1521 |
| IS | 0 | 108.8967 | 108.8967 |
| TL | 1.3921 | 5.9123 | 7.1855 |
| ML | 101.1847 | 354.9577 | 399.6623 |
| WL | 0.0592 | 0 | 0.0592 |

To determine performance in the steady state of this trial, we used ATTGB to measure the degree of stability of the four controlled variables for the last five minutes. The results of an ATTGB analysis are presented at Table 17. AS and WL seem to be the weak subjects in each of the groups.

The graphs of mass versus energy are presented in Figures 58 and 59 for the P+F and P groups, respectively. Not surprisingly, there are some disturbances on the temperature in the upper reservoir due to the subjects' usage of the two feedwater streams to compensate when fault occurred.

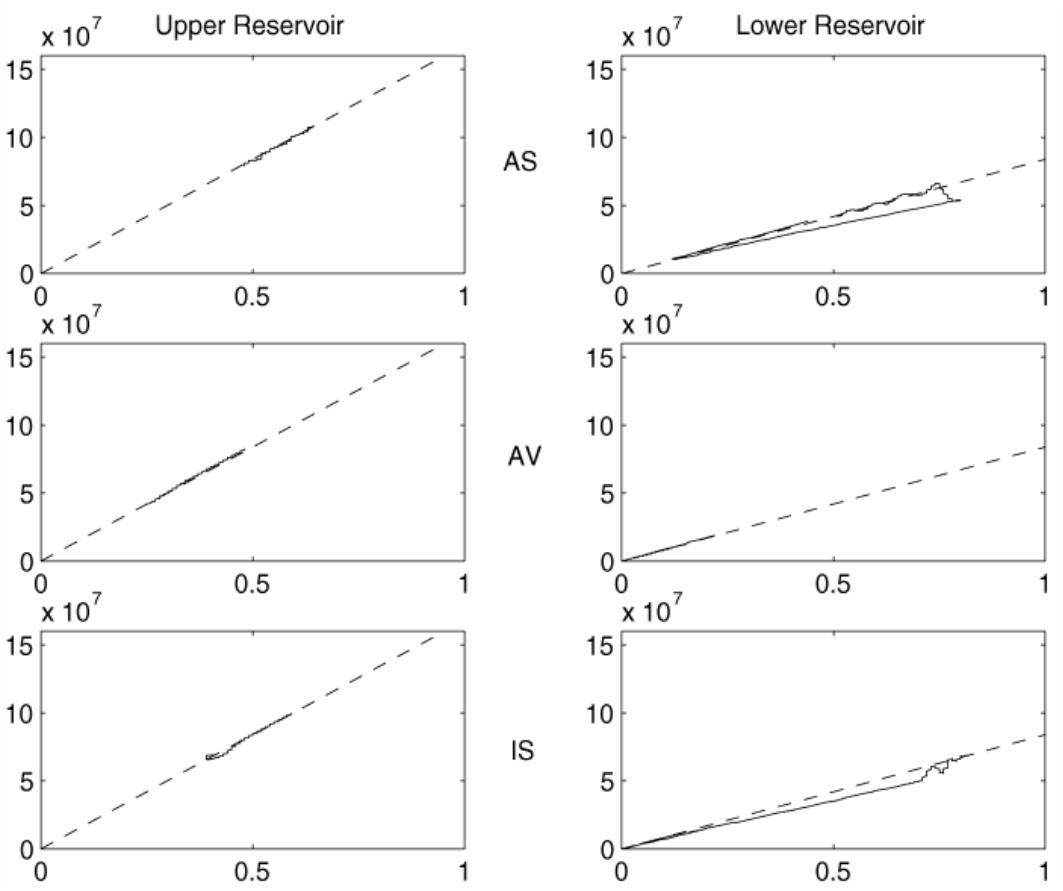


**Figure 58:** Graphs of mass versus energy for P+F group (trial 113)

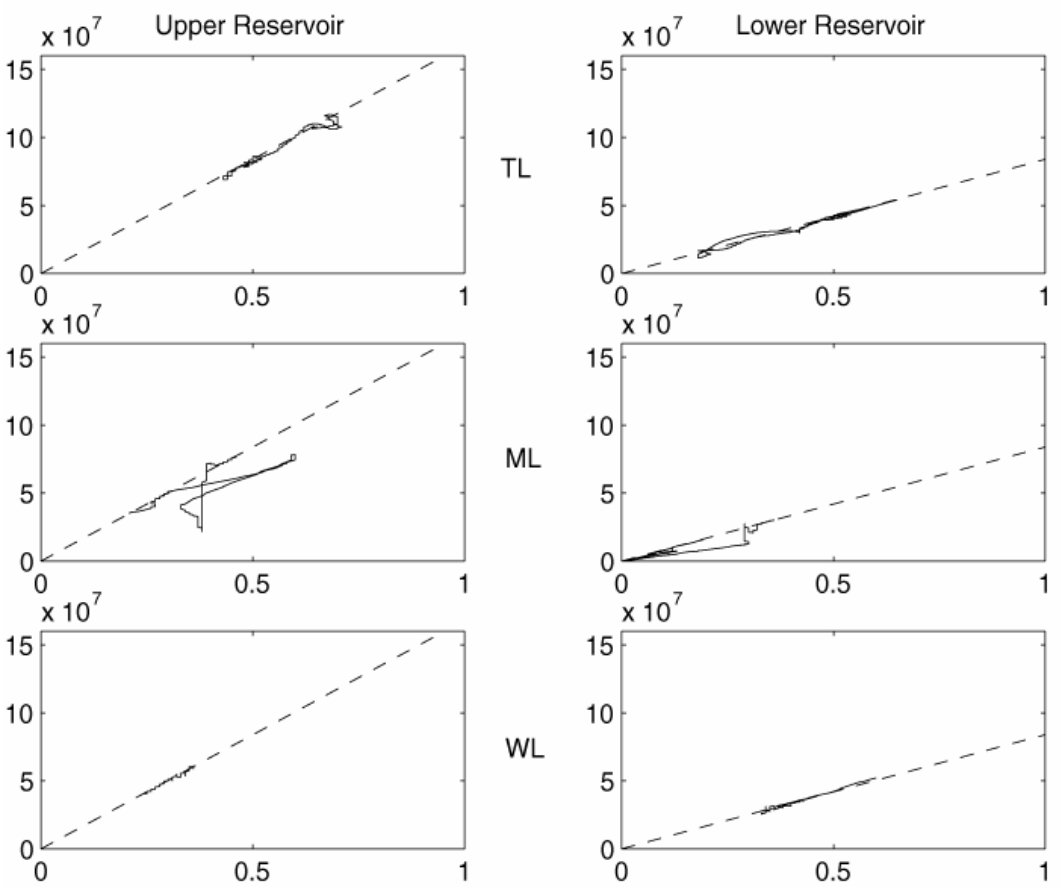


**Figure 59:** Graphs of mass versus energy for P group (trial 113)

**Table 17:** Area under TTGB for AS, IS, TL and WL (trial 113)

| Subjects | T1 | T2 | D1 | D2 |
|---|---|---|---|---|
| AS | 0.8220 | 0.6596 | 0.9680 | 0.9549 |
| AV | 0.8802 | 0.8940 | 0.9700 | 0.9727 |
| IS | 0.8040 | 0.7947 | 0.9864 | 0.9798 |
| TL | 0.6510 | 0.7768 | 0.9842 | 0.9895 |
| ML | 1.0000 | 0.8860 | 0.9875 | 0.9885 |
| WL | 0.8019 | 0.6559 | 0.9731 | 0.9747 |

The numbers of control actions the subjects made after the fault occurred (Table 18) show that there were differences between the interface groups. The P+F subjects, especially AS and AV, made very few control actions to compensation.

**Table 18:** Numbers of control actions (trial 113)

| Subjects | AS | AV | IS | TL | ML | WL |
|---|---|---|---|---|---|---|
| Number of actions | 12 | 12 | 63 | 80 | 77 | 46 |

This was the second fault trial with reservoir leakage, with the reservoir leaking at a faster rate in this fault than in the previous fault of the same type. Nonetheless, there was some improvement in subjects' performance. First, all of the subjects completed the trial successfully, while in the previous fault trial of the same type, two of them failed to do so. Second, the compensation time was shortened for most of the subjects in this fault compared to the previous one. It is understandable that some of the measures indicated that the subjects' performance was not better than before, given that this fault task was more difficult than the previous one.

**Trial 144**

This was the second fault trial involving a heater malfunction. The lower heater began to output at 150% the set rate 8 minutes after the beginning of the trial. For all subjects, this fault occurred when they had entered the steady state portion of the tuning task.

No verbal protocols were recorded in this fault trial. Some of the measures developed for the tuning trials were applied to this fault, but no interesting results were found. This can probably be attributed to the design of the experiment, as the impact of the fault was not significant enough to affect system behaviour. To illustrate this, we plotted graphs of the temperature perturbation after the occurrence of the fault (see Figure 60). There is no change in the temperature of lower reservoir for the subjects.

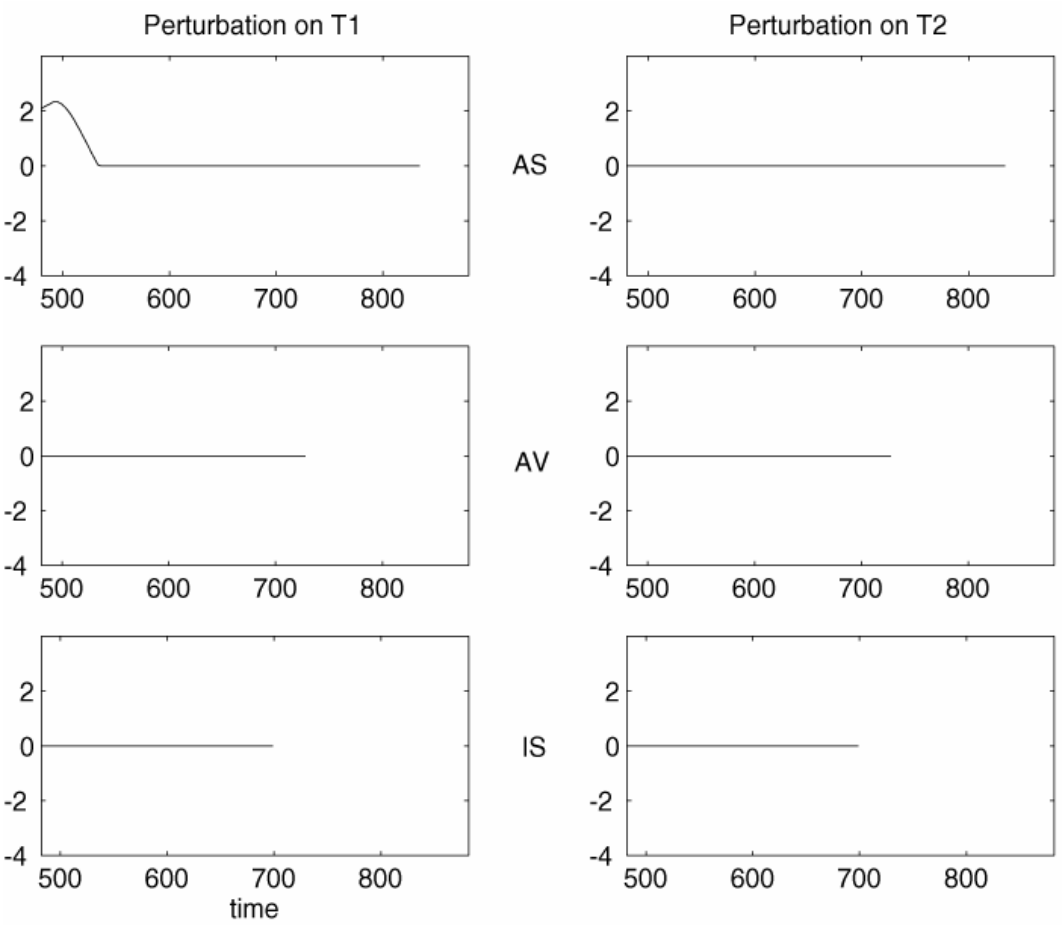


**Figure 60:** Temperature perturbation (trial 144)

**Trial 165**

This was the third fault trial involving reservoir leakage. Nine minutes into the trial the lower reservoir started to leak at a rate of 4 units/second. For all subjects, the fault occurred during the steady state portion of the tuning task.

In the P+F group, IS was the fastest to detect the fault, in spite of the fact that he was not able to diagnose the location of the fault. He was the only subject in the P+F group that was able to compensate successfully for the fault. AS was the slowest to detect the fault, but he seemed to understand the location of the fault. AV did detect the fault, but nonetheless was not able to compensate as he was already emptying the reservoir in anticipation of the shutdown task. In the P group, TL was the fastest at detecting the fault, ML was much slower, and WL did not detect the fault at all. Although all three of the P subjects were not able to diagnose this fault correctly,

two of them were able to successfully compensate for it. One of these subjects, WL, did not even notice that a fault had taken place!

As there was no change in the temperature goals for the tuning task, it is useful to understand how the temperatures of the reservoirs were perturbed by the fault. We plotted the curves of temperatures over time and then calculated the area under these curves. The results of this analysis for subjects who successfully completed the trial are presented in Table 19. Notice that ML was the only subject to experience any perturbations.

**Table 19:** Area under the curves of perturbation on temperature for IS, ML and WL (trial 165)

| Subjects | Upper Reservoir | Lower Reservoir |
|---|---|---|
| IS | 0 | 0 |
| ML | 1.4952 | 10.5734 |
| WL | 0 | 0 |

To see how quickly subjects recovered from the disturbance of temperature and how severely these disturbances were, the time taken to bring the temperature back to within the goal boundaries and the maximum temperature deviation was calculated (see Table 20). Not surprisingly, ML has the worst performance with respect to this measure as well.

**Table 20:** Recover time and largest deviation

| | Recover time | | Largest deviation | |
|---|---|---|---|---|
| Subjects | Reservoir 1 | Reservoir 2 | Reservoir 1 | Reservoir 2 |
| IS | 0 | 0 | 0.8600 | 1.6500 |
| ML | 387.6117 | 347.7907 | 5.4900 | 20.0000 |
| WL | 0 | 0 | 1.7400 | 1.8100 |

The graphs of time to goal boundaries for the three subjects who successfully compensated for this fault are shown in Figures 61 – 63. WL's temperature instability indicate that he was the worst performer among these subjects. IS and ML had comparable performance in terms of this measure. A calculation of the area under these curves (Table 21) gives further evidence for these statements.

**Table 21:** Area under the curves of TTGB for temperature

| Subjects | T1 | T2 |
|---|---|---|
| IS | 0.8988 | 0.8104 |
| ML | 0.7031 | 0.8999 |
| WL | 0.6497 | 0.7546 |

The last measure we used in this trial was the numbers of control actions used by successful subjects (Table 22). IS made very few control actions to complete the task, when compared with ML and WL.

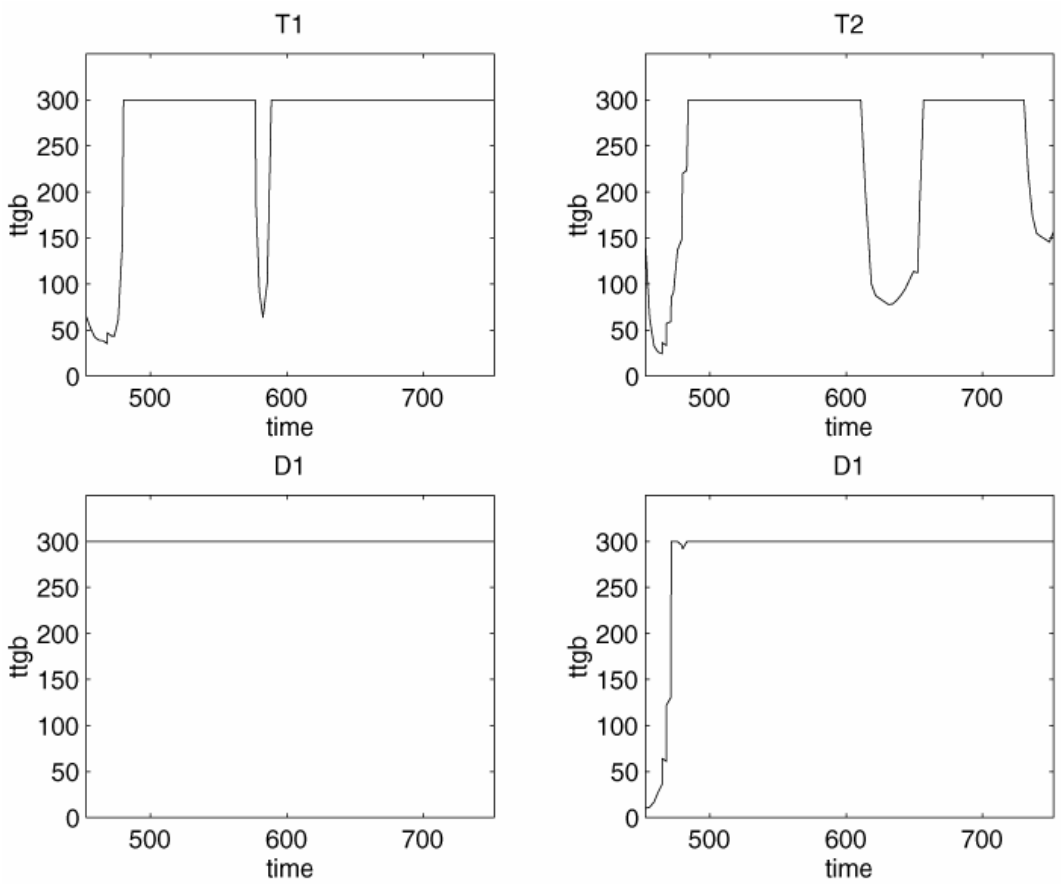


**Figure 61:** Graphs of time to goal boundaries for IS (trial 165)

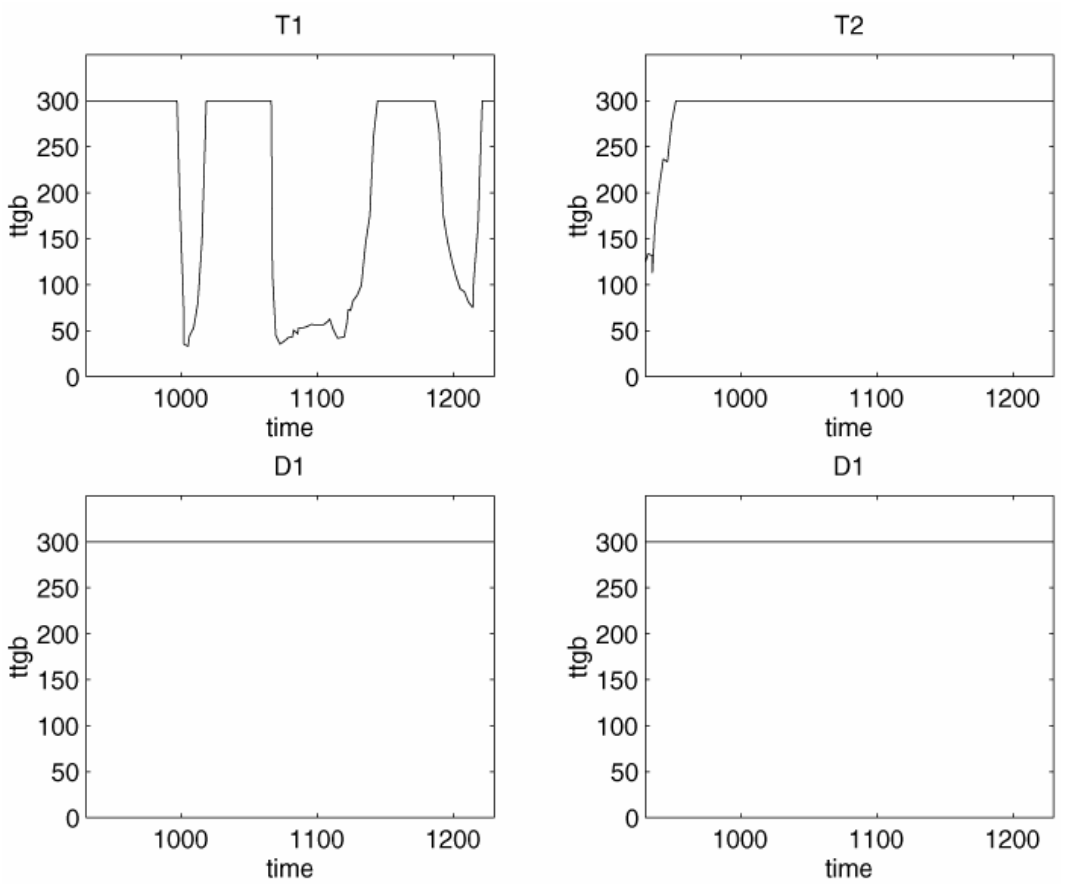


**Figure 62:** Graphs of time to goal boundaries for ML (trial 165)

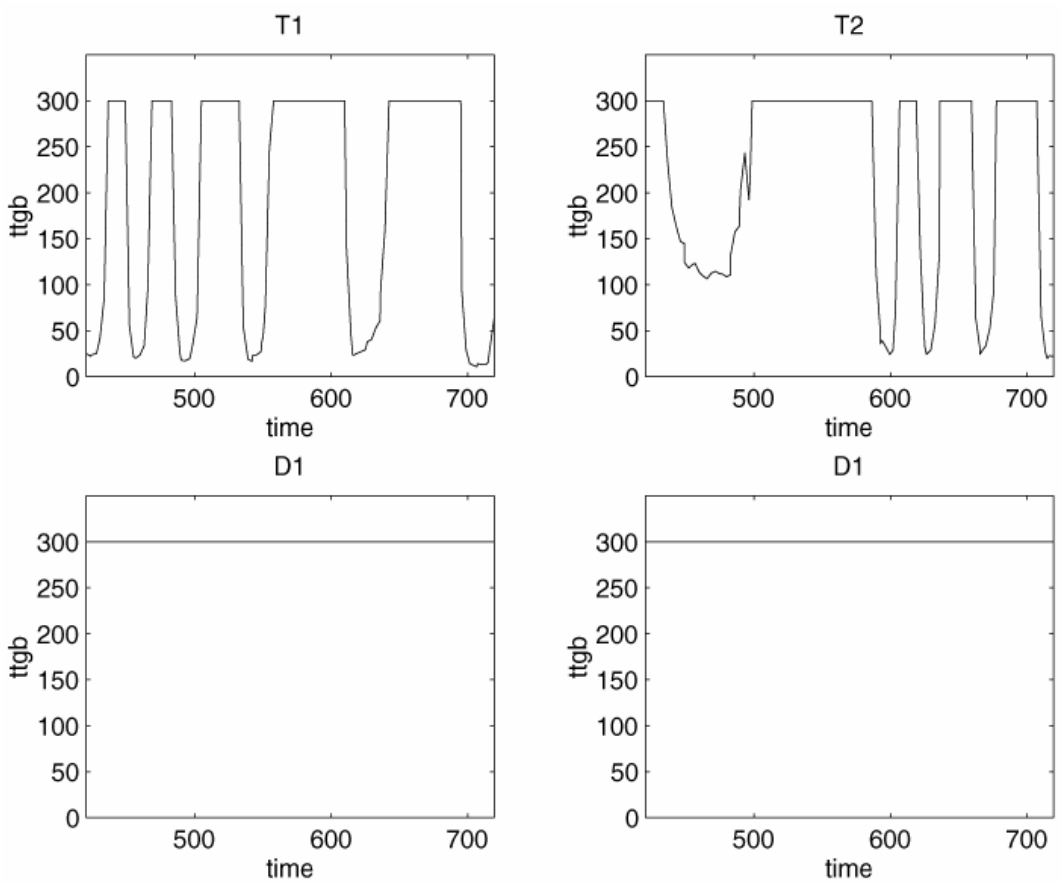


**Figure 63:** Graphs of time to goal boundaries for WL (trial 165)

**Table 22:** Numbers of control actions of IS, ML and WL (trial 165)

| Subjects | IS | ML | WL |
|---|---|---|---|
| Number of actions | 15 | 71 | 39 |

There were no obvious differences in detection times for the two groups in this fault trial. While all of the P+F subjects detected the fault, only two subjects in the P group did so. In addition, the P+F subjects again were able to diagnose the fault more accurately (Christoffersen et al., 1994). Given the fact that only three subjects were successful, it is hard to make any solid conclusion about interface differences from this fault trial. Also, given that this fault occurred during the steady state portion of the tuning trial while the other two similar faults (trials 94 and 113) occurred during the startup task, it is difficult to make any comparison between these faults.

It should be noted that this fault was designed specifically to occur late in the trial when some subjects would have been preparing for shutdown. While this allowed us to investigate the effects of a fault introduced at this unique time, it also reduced our ability to compare across subjects and trials.

**Trial 177**

This was the second fault trial involving a valve blockage. Valve VB2 became blocked gradually starting at 7 minutes into the trial. This fault was designed to occur during the first stage of the tuning task.

The only subject that was not able to successfully compensate for this fault was WL of the P group. By turning off the valves downstream from PB he blew up this pump, which is quite a remarkable error considering the experience that he had with the system by this point.

Since the fault happened during tuning task, we first considered the temperature disturbances caused by the VB2 blockage. The areas under the curves of temperature perturbation are shown in Table 23. As it turns out, there was virtually no temperature perturbation for all five of the successful subjects except AS.

**Table 23:** Area under the curves of perturbation on temperatures (trial 177)

| Subjects | T1 | T2 |
|---|---|---|
| AS | 0.0986 | 0 |
| AV, IS, WL,TL | 0 | 0 |

To investigate the temperature perturbation further, we calculated temperature recovery time (i.e., how long it took to bring the temperature back into the goal boundaries) and largest temperature deviation (see Table 24). The results are consistent with what was learned from the area under the temperature perturbation plot (see Table 23). Since there was no temperature perturbation for AV, IS, TL and WL, their recover time was zero, and their temperature variations were within the tolerance boundaries (2 degrees on either side of the goal). This analysis also confirmed the fact that AS experienced a relatively large temperature deviation that he took some time to recover from.

**Table 24:** Recover time and largest deviation of temperatures

| Subjects | Recover time | | Largest deviation | |
|---|---|---|---|---|
| | Reservoir 1 | Reservoir 2 | Reservoir 1 | Reservoir 2 |
| AS | 39.4466 | 0 | 2.4700 | 1.4500 |
| AV | 0 | 0 | 1.2300 | 0.5600 |
| IS | 0 | 0 | 1.1700 | 1.2600 |
| TL | 0 | 0 | 1.2400 | 0.6000 |
| WL | 0 | 0 | 1.7100 | 1.0500 |

For steady state performance, we plotted the graphs of time to goal boundaries and calculated the areas under these curves. Only partial results of the latter measure are included here (see Table 25). These results reveal that WL had the least stable steady state.

**Table 25:** Area under the curves of time to goal boundaries for temperatures

| Subject | T1 | T2 |
|---|---|---|
| AS | 0.8909 | 0.8202 |
| AV | 0.8477 | 0.9304 |
| IS | 0.9058 | 0.8681 |
| TL | 0.8310 | 0.9169 |
| WL | 0.5282 | 0.8054 |

The number of control actions to compensate used by subjects to compensate for the fault are shown in Table 26 (note that also included in this table is ML, who failed to complete the task). This table reveals that subjects who were proficient performers in their respective groups (IS of the P+F group, TL of the P group) used fewer control actions than the other subjects. Overall, the P+F subjects used fewer control actions to compensate than the P subjects.

**Table 26:** Number of actions (trial 177)

| Subjects | AS | AV | IS | TL | ML | WL |
|---|---|---|---|---|---|---|
| # of actions | 43 | 30 | 13 | 16 | 39 | 52 |

In summary, there is not strong enough evidence to differentiate between individuals and interface groups. The results presented do, however, reveal a small advantage for the P+F subjects as (1) all P+F subjects completed the task, while only two P subjects were able to, and (2) the P+F subjects used less compensatory control actions. Finally, it is difficult to make a comparison across the two valve blockage faults as they occurred at different parts of the trial.

**Trial 183**

Trial 183 was the third instance of a heater fault. At 7.5 minutes after the start of the trial, the heating rate of the lower reservoir heater dropped exponentially to half of its setting. The fault occurred at the period of tuning task for all of the subjects.

In the P+F group, all of the subjects successfully detected, diagnosed, and compensated for this fault. IS was the fastest at detecting the fault, and AS was the slowest. In the P group, only ML detected the fault. TL and WL also completed the task, but both of them did not detect the fault (see Christoffersen et al., 1994).

The graphs of temperature perturbation are shown in Figures 64 and 65 for the P+F and P groups, respectively. Two findings emerge from these graphs. First, there was a slight disturbance of the upper reservoir temperature for AS, but no disturbance for the other two P+F subjects. In the P group, only TL experienced variation in the temperatures of both reservoirs. This may be because TL did not know that a fault had occurred. Second, the time to complete the task was shorter for the P+F subjects.

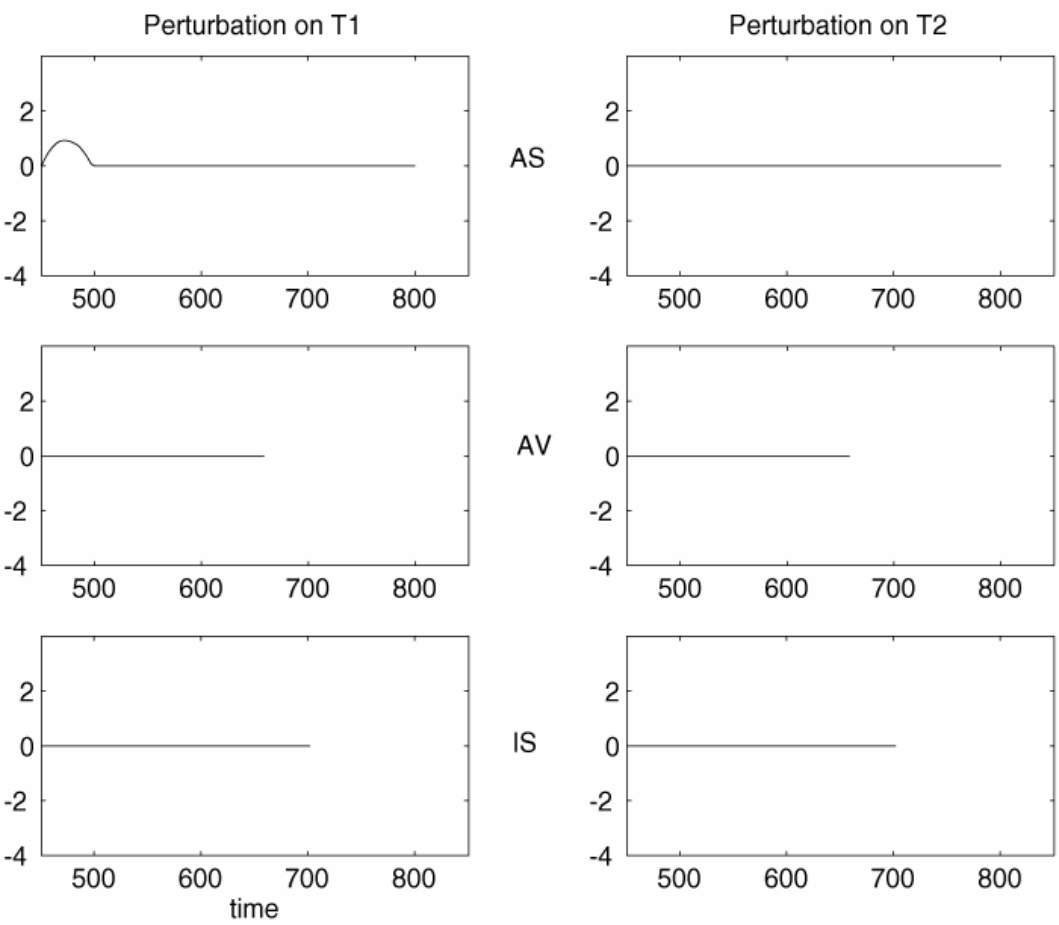


**Figure 64:** Graphs of perturbation on temperatures for P+F group (trial 183)

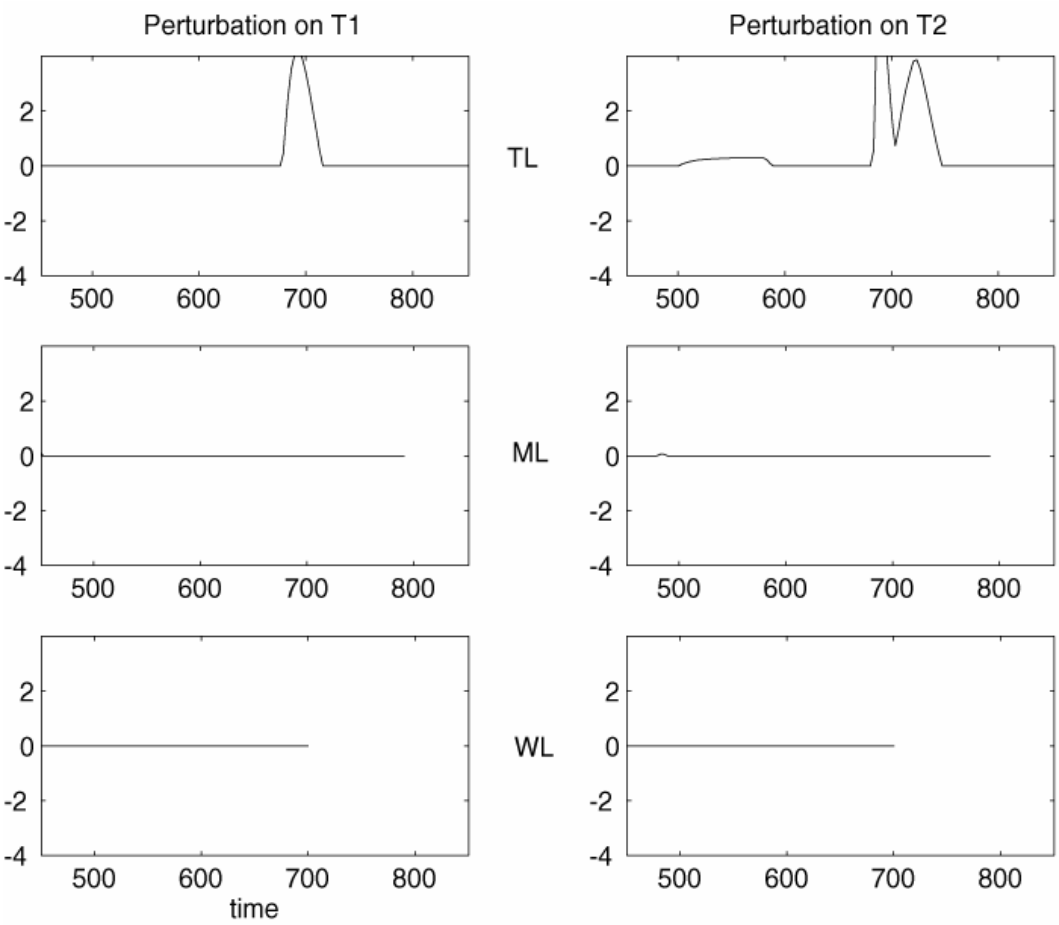


**Figure 65:** Graphs of perturbation on temperatures for P group (trial 183)

Recover time and largest deviation of temperatures are shown in Table 27. AS and TL had long recover times and deviation of temperatures, making them the worst in their interfaces groups on these measures in this fault. TL's performance was the worst of all the subjects. On average, the P+F subjects had shorter recover times, and smaller deviations.

**Table 27:** Recover time and largest deviation (trial 183)

| Subjects | Recover time | | Largest deviation | |
|---|---|---|---|---|
| | Reservoir 1 | Reservoir 2 | Reservoir 1 | Reservoir 2 |
| AS | 46.8366 | 0 | 2.9300 | 0.5100 |
| AV | 0 | 0 | 0.7100 | 1.2300 |
| IS | 0 | 0 | 1.4300 | 1.2400 |
| TL | 260.4602 | 291.9051 | 6.1900 | 9.9400 |
| ML | 0 | 35.1636 | 2.1000 | 2.0799 |
| WL | 0 | 0 | 1.8700 | 1.3300 |

The area under the curves of time to goal boundaries was calculated for all subjects. Since this fault involved only the lower heater, only the results for T2 are presented in Table 28. Two interesting findings stand out. First, the P+F subjects achieved more stable steady states than the P subjects. Second, WL's performance was the most unstable of all subjects.

**Table 28:** Area under the curves of TTGB for T2 (trial 183)

| Subjects | AS | AV | IS | TL | ML | WL |
|---|---|---|---|---|---|---|
| T2 | 0.9629 | 0.9724 | 0.8521 | 0.8302 | 0.8202 | 0.7458 |

The number of control actions used by subjects to compensate is shown in Table 29. TL's used the most control actions of all the subjects. This is not surprising considering that he did not understand that a fault had occurred. In the absence of some form of diagnosis, he just tweaked the system to make it work, and this involved many control actions. On average, the P+F subjects used fewer control actions than the P subjects.

**Table 29:** Number of actions (trial 183)

| Subjects | AS | AV | IS | TL | ML | WL |
|---|---|---|---|---|---|---|
| # of actions | 23 | 15 | 15 | 47 | 19 | 27 |

Some interesting findings were made in the analysis of this fault trial. First, the P+F subjects more easily detected, diagnosed, and compensated for this fault than did the P subjects (Christoffersen et al., 1994). This more detailed study suggested that the P+F subjects relied on the information provided by the interface to handle the fault, while the P subjects lacked this information and instead had to rely on the rules they had developed to formulate a trial and error strategy to diagnose the fault. The performance analysis demonstrated that the P+F subjects exhibited slightly more stable control in terms of temperature deviation, recover time, and number of control actions.

Compared with the previous trial of the same type, trial 144 (see Table 30), it is interesting to note that the performance of the P+F group improved in terms of almost every measure considered, while the P subjects did not exhibit this improvement.

**Table 30:** Comparison between trial 183 and trial 144

| | Area under Perturbation on Temperature* | | Recover Time* | | |
|---|---|---|---|---|---|
| Subjects | Res 1 | Res 2 | Res 1 | Res 2 | Number of actions* |
| AS | +50% | 0 | +11% | 0 | +26% |
| AV | 0 | 0 | 0 | 0 | +21% |
| IS | 0 | 0 | 0 | 0 | -36% |
| TL | -inf | -inf | -inf | -inf | -262% |
| ML | -inf | -inf | 0 | -inf | -46% |
| WL | 100% | 0 | +inf | 0 | +31% |

*(Positive percent number indicates the improvement in performance, and negative number indicates the decrease in performance.)

**Trial 187**

This was the third fault trial involving a valve blockage. Eight minutes after the beginning of the trial, VB became blocked. Compared with the previous fault trials with the same type of fault (trials 64 and 177), this trial was more challenging as it involved a primary, rather than a secondary, valve. The fault was designed to occur during the tuning stage for all subjects.

In the P+F group, both AS and IS detected the fault very quickly. AV did not even experience the fault because he was using a one-feedwater stream strategy when the fault occurred that did not use VB. IS completed the task while AS damaged PB and terminated his trial prematurely. In the P group, ML detected the fault but failed to complete the task, while WL neither detected the fault nor completed the task. Just like AV, TL did not experience the fault as he was using a one-feedwater stream strategy. Thus we could not consider AV and TL in this analysis.

Among the rest of the four subjects, only IS completed the task successfully. While we did do a performance analysis for IS, there is little sense in presenting the results as there is little to compare them against.

**Discussion**

These fault trials can all be considered as perturbations of system structure. The fault management tasks presented to the subjects were challenging, and gave subjects opportunities to learn about the system constraints and functions. In this section, we have presented results from a detailed analysis of these routine fault trials. A number of these findings are notable. First of all, in agreement with what was reported by Christoffersen, et al. (1994), there were considerable differences between subjects and interface groups in terms of detection, diagnosis, and compensation times:

- IS in the P+F group was the fastest at detecting the faults. On the average, P+F subjects were faster at detecting faults than the P subjects.
- P+F subjects were able to make better and faster diagnoses than the P subjects.
- The compensation time of the P+F subjects was generally shorter than that of the P subjects.

The state-space performance analyses presented in this report confirmed and fleshed out the findings listed above:

- IS was the best in performance among all subjects. AS, and ML were usually the worst in their respective groups.

- In terms of the state-space measures, there was a slight performance advantage in favour of the P+F subjects.
- There is evidence that the performance of the P+F subjects improved over time, while there is no such support for the P subjects. While this cannot easily be seen in all cases due to difficulties in comparison, in one case (trial 183) the evidence for this hypothesis was considerable.

A notable finding here, then, is that the P+F interface did assist subjects in routine fault management better the P interface. It is likely that this can be attributed to the richer and more systematic display of information provided by the P+F interface.

## 5. STATE SPACE ANALYSIS: NON-ROUTINE FAULT MANAGEMENT

Non-routine faults were designed to be representative of unfamiliar, unanticipated events that are rarely, if ever, experienced by operators. Three types of non-routine faults were defined in this experiment.

- A reservoir leakage followed by an increase in the temperature of incoming water.
- A valve blockage followed by the introduction of an external heat source.
- A heater failure followed by an increase in the temperature of incoming water.

Three non-routine faults were presented to the subjects late in the experiment (one of each type). In this section, we will present the analysis results of these three trials. We will first give a qualitative description of the overall fault management performance for all subjects based on the research of Christoffersen et al. (1994), and will then present the results of a state space analysis.

**Trial 201**

This was the first non-routine fault trial. During this trial, a 4 unit/second leak appeared in the upper reservoir at 4 minutes into the trial, followed by a 6 °C increase in the input water temperature.

The two faults occurred during the startup task for all subjects except AV. He was still in the startup phase when the first fault appeared, but was in the tuning phase at the onset of the second fault. Since we are interested in the impact of non-routine faults on the subjects' performance and their adaptation, we will focus on the analysis of the data after the second fault to the end of the task.

A process tracing analysis (Christoffersen et al., 1994) described the fault management activities of each of the subjects. AV was the best performer in the P+F group, as he was fastest

at detecting and diagnosing the faults and also completed the task the most quickly. IS detected the two faults very quickly. Although he was not able to diagnose the faults, he completed the trial after a relatively long time. AS detected and diagnosed the first fault, but not the second one. It took him longer than IS to compensate and he eventually completed the task as well. In the P group, TL was the only one who clearly indicated that he had detected the two faults, as he found that one of the rules he had developed was violated. ML blew up the system as a result of the first fault. WL completed the task having not even noticed that the two faults had occurred.

The lengths of trajectories in goal space were calculated for AS, IS, TL and WL. Note that this measure was not applied to AV as he met the second trial under different conditions than the other subjects (tuning vs. startup), or to ML whose trial was prematurely terminated. The results are shown in Table 31. AS was quite clearly the worst performer of the four subjects as he did not detect the second fault and had a difficult time compensating for it, and IS was the best.

**Table 31:** Lengths of trajectories in goal space (trial 201)

| Subjects | Reservoir 1 | Reservoir 2 | Reservoir 1 & 2 |
|---|---|---|---|
| AS | 1.9206 | 1.4867 | 2.8294 |
| AV | | Not applied | |
| IS | 0.2137 | 0.2701 | 0.3840 |
| TL | 0.3105 | 0.5005 | 0.6377 |
| ML | | Not applied | |
| WL | 0.2795 | 0.3410 | 0.4962 |

The results of an analysis of the areas under the curves of distance to the goals are quite consistent with the above findings. They are shown in Table 32. Again, AS was the worst of the four subjects.

**Table 32:** Area under the curves of distance to the goals (trial 201)

| Subjects | Reservoir 1 | Reservoir 2 | Reservoir 1 & 2 |
|---|---|---|---|
| AS | 86.3712 | 2.9686 | 87.5394 |
| AV | | Not applied | |
| IS | 2.2946 | 0.1841 | 2.4786 |
| TL | 0.5861 | 1.7322 | 2.0687 |
| ML | | Not applied | |
| WL | 0.4085 | 0 | 0.4085 |

AS was also the worst in the steady state performance. We will focus on the steady state performance of temperature since the second fault was a perturbation of the temperature of the input water. The results are shown in Table 33. AS had greater potential instabilities in both reservoirs than all of the other subjects, and was followed by WL. An interface effect cannot be seen from these results.

**Table 33:** Area under the curves of time to goal boundaries (trial 201)

| Subjects | T1 | T2 |
|---|---|---|
| AS | 0.5922 | 0.6066 |
| IS | 0.7393 | 0.9685 |
| TL | 0.8239 | 0.8592 |
| WL | 0.7311 | 0.7748 |

In summary, in terms of the fault detection accuracy and speed, P+F subjects were better than P subjects in this trial. Individually, AS and ML were the worst in their respective groups. The performance analysis based on the state space approach revealed that AS was the worst in both dynamic and steady state performance.

**Trial 209**

This was the second of the non-routine fault trials. After 7 minutes, VA gradually became blocked, and one minute later, an external heat source added heat to the lower reservoir.

This may have been the most difficult fault trial. Four of the subjects (AS, IS, TL, and WL) prematurely ended the trial due to the first fault, and so never even experienced the second fault. AV did not experience the first fault as he had closed VA in an anticipation of shutdown before the fault occurred. As ML was still in the startup phase when both of the faults occurred, the data from his trial is not comparable to AV's. Thus, it is difficult to make any comparisons between subjects or groups for this trial. Unfortunately, this trial was much too difficult.

**Trial 218**

This was the last non-routine fault trial. Three minutes into this trial, the output of the upper heater began to increase exponentially until it output 150% of its setting. Two minutes later, the temperature of the input water was increased by 11 °C.

Before presenting the analysis results of the state space approach, we will review the findings based on the process tracing approach (Christoffersen et al., 1994). All of the subjects successfully completed the task. In the P+F group, IS quickly detected and (partially) diagnosed both of the faults, and compensated efficiently. AS took longer to detect the faults, was only able to diagnose the first one, and compensated slowly. While AV detected the faults quickly, he was not able to diagnose the root causes. He was also best in the P+F group at compensation. In the P group, TL detected both faults very quickly using the rules he developed, but was only able to diagnose the second fault. WL did not detect the first fault but fully diagnosed the second one. He compensated much faster than the other two subjects in the P group. ML was slow at detecting both of the faults, was only able to diagnose the first one, and was the slowest at compensation.

The trajectories in goal space for the P+F and P groups are shown in Figures 66 and 67, respectively. The graphs start from the time the second fault occurred. On these graphs, there appears to be a difference in the curves in favour of the P+F subjects. The lengths of these curves are reported in Table 34. AS seemed to have some trouble to get the lower reservoir under control. Excepting AS, the P+F subjects had shorter lengths of trajectories in goal space than the P subjects.

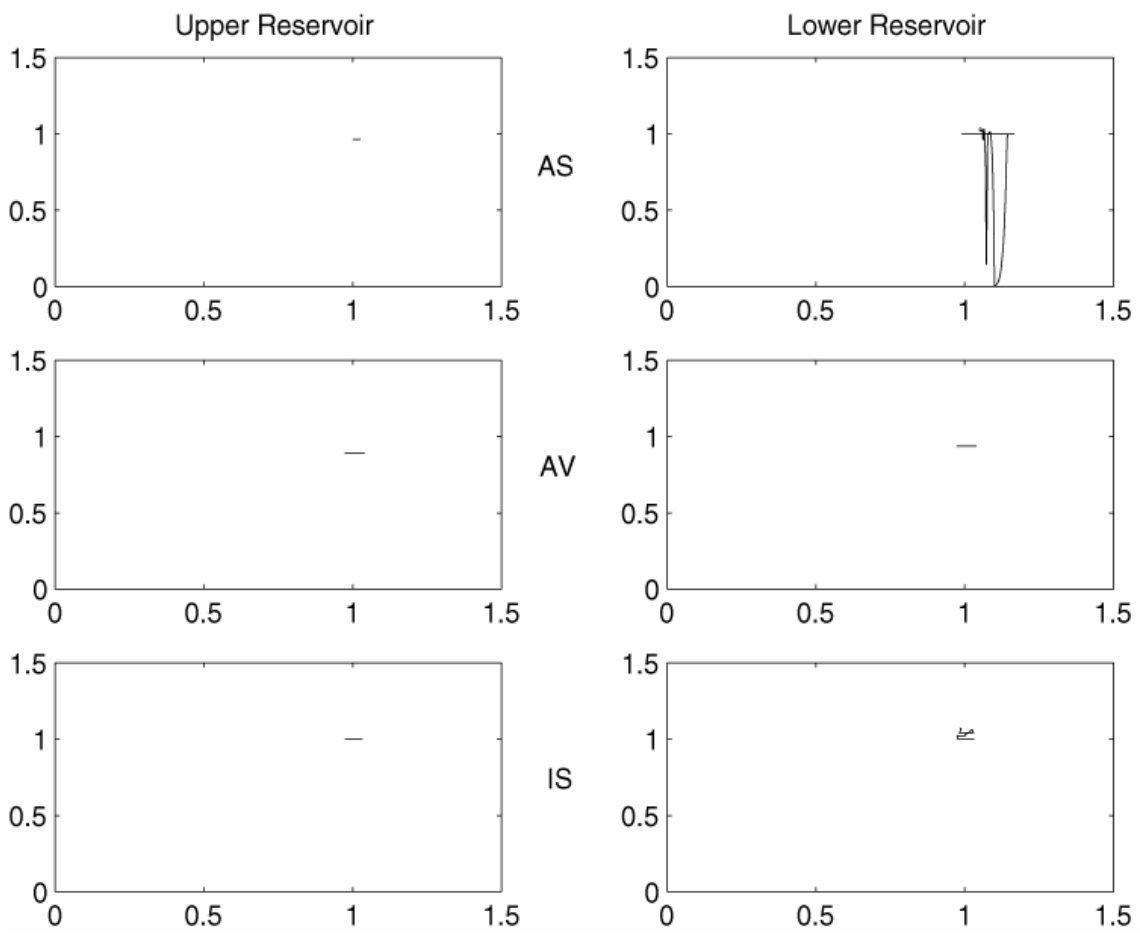


**Figure 66:** Trajectories in goal space for P+F subjects (trial 218)

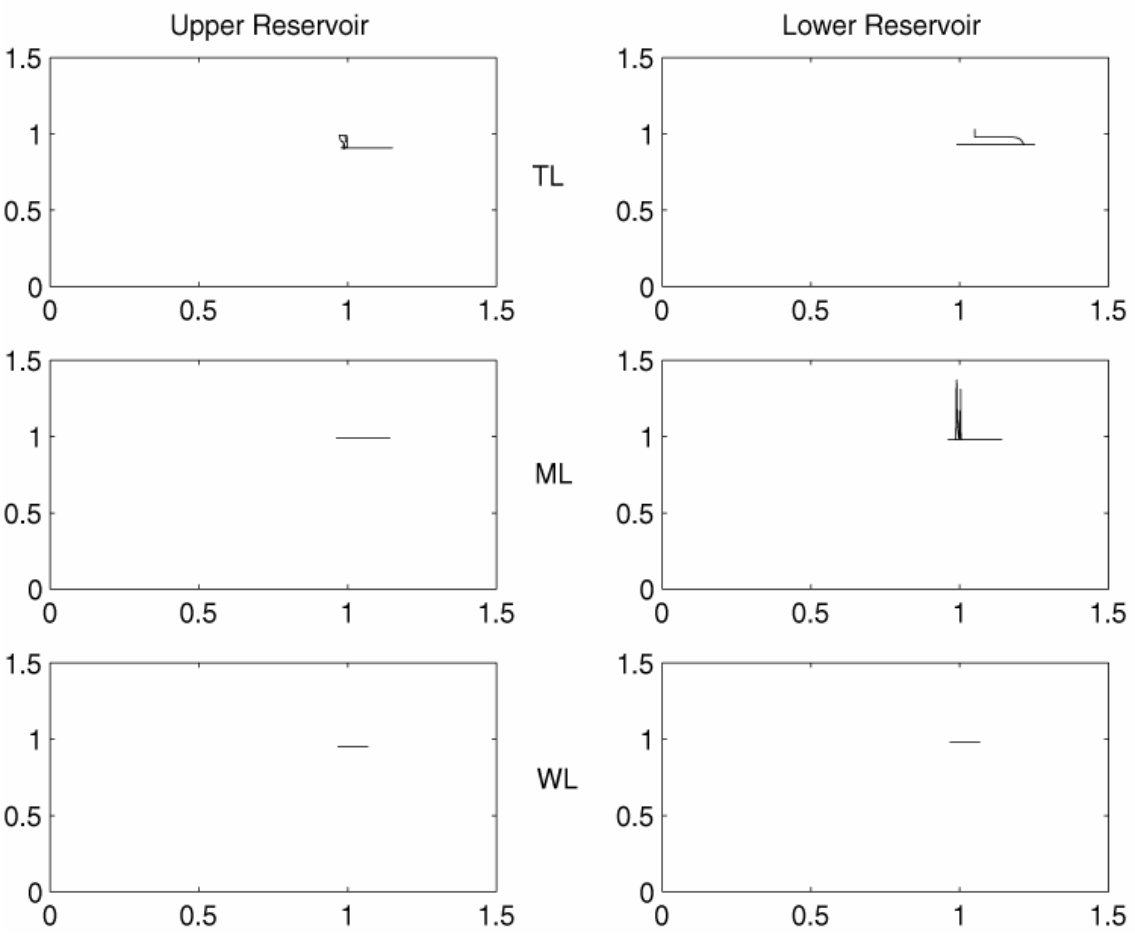


**Figure 67:** Trajectories in goal space for P subjects (trial 218)

**Table 34:** Lengths of trajectories in goal space (trial 218)

| Subjects | Reservoir 1 | Reservoir 2 | Reservoir 1 & 2 |
|---|---|---|---|
| AS | 0.0557 | 4.1529 | 4.1605 |
| AV | 0.2475 | 0.1250 | 0.3195 |
| IS | 0.1968 | 0.3228 | 0.4403 |
| TL | 0.6831 | 0.5419 | 1.0298 |
| ML | 0.5332 | 2.4646 | 2.7033 |
| WL | 0.5675 | 0.1180 | 0.6235 |

In what follows, the dynamic performance of each subject will be reported. The graphs of distance to the goals are shown in Figures 68 and 69, for the P+F and P groups, respectively. AS had the largest deviation from goals in the lower reservoir. AV and IS in the P+F group, and WL in the P group, only experienced a slight deviation of the goals after the second fault.

The areas under the curves of distance to goals are shown in Table 35. AS had the worst performance with largest area. Again, excepting AS, it seems that the performance is in favour of the P+F subjects in terms of this measure as well.

**Table 35:** Area under distance to the goals (trial 218)

| Subjects | Reservoir 1 | Reservoir 2 | Reservoir 1 & 2 |
|---|---|---|---|
| AS | 0 | 33.9293 | 33.9293 |
| AV | 0.0192 | 0 | 0.0192 |
| IS | 0 | 0.1238 | 0.1238 |
| TL | 7.2085 | 12.9075 | 19.3535 |
| ML | 4.3233 | 16.0020 | 18.7200 |
| WL | 0.2042 | 0 | 0.2042 |

The area under the curves of time to goal boundaries is shown in Table 36. There is no clear interface difference between the two groups.

It should be noted that the two faults were both designed to affect the temperature of the water in both of the reservoirs. Because of the peculiar interaction between the two faults, it is not a surprise that some of the subjects were only able to detect one of the faults. In terms of fault detection, diagnosis and compensation time, the P+F subjects outperformed the P subjects. Finally, while AS was the worst subject in terms of dynamic performance, the other two P+F subjects clearly outperformed the P subjects.

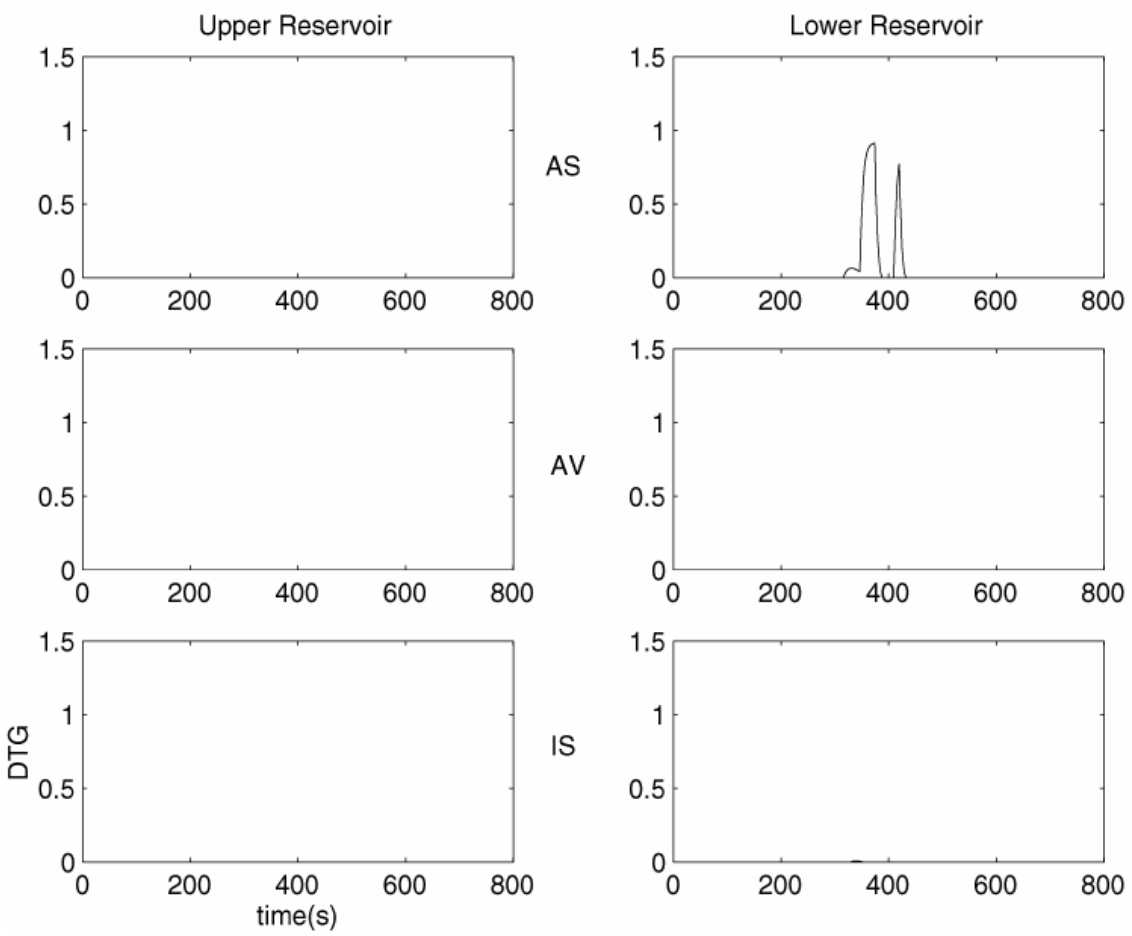


**Figure 68**: The graphs of distance to the goals for P+F subjects (trial 218)

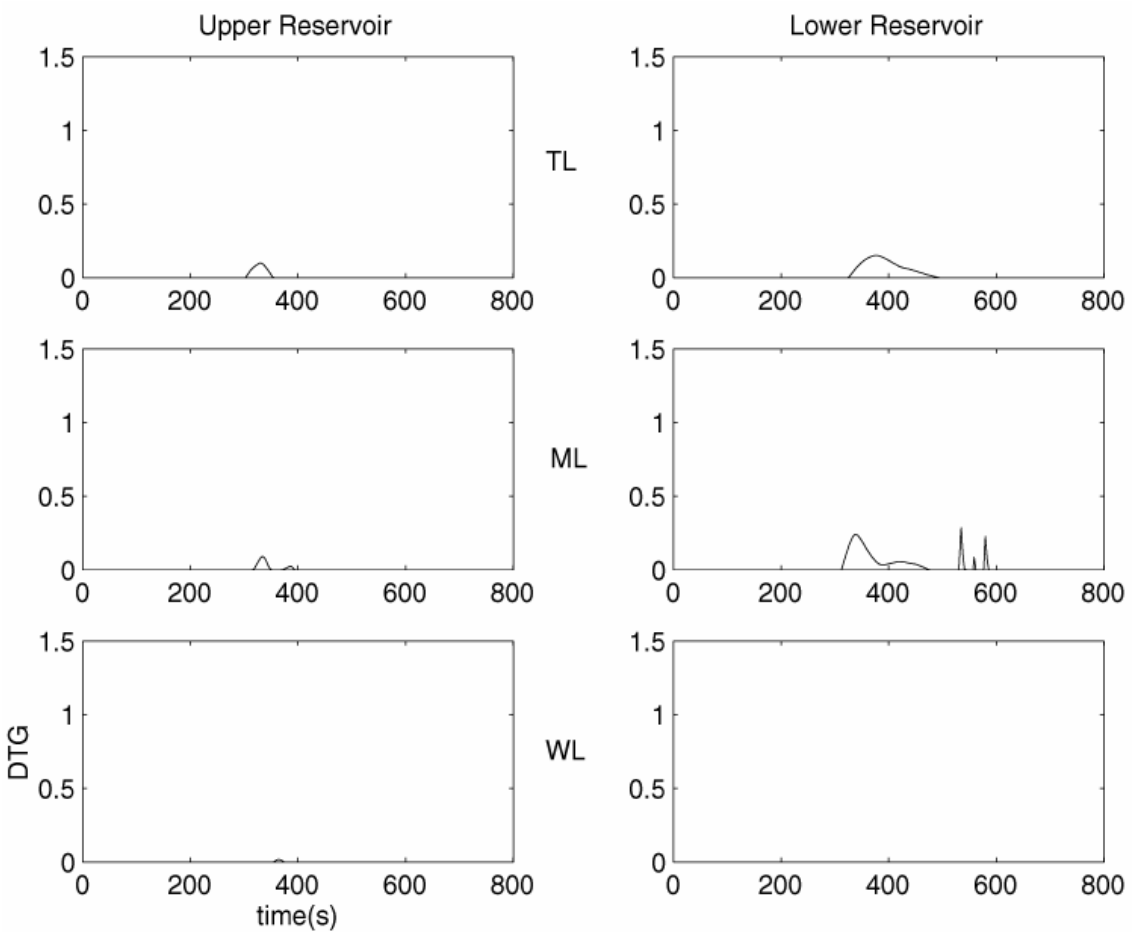


**Figure 69:** The graphs of distance to the goals for P subjects (trial 218)

**Table 36:** Area under the curves of time to goal boundaries (trial 218)

| Subjects | T1 | T2 |
|---|---|---|
| AS | 0.8505 | 1.0000 |
| AV | 0.6505 | 0.8486 |
| IS | 0.7558 | 1.0000 |
| TL | 0.8788 | 1.0000 |
| ML | 0.8517 | 1.0000 |
| WL | 0.6553 | 0.8286 |

**Discussion**

The non-routine fault trials in the longitudinal experiment were designed to be representative of rare and unanticipated faults that can occur in industrial systems. These faults were more difficult to detect, diagnose, and compensate for than the routine faults subjects also experienced, and so required that subjects dig deep into their understanding of the system structure and functioning. The new analysis methods developed during last year's contract have provided some new results for these faults.

First, we were able to obtain the same results as Christoffersen, et al. (1994), and found differences between subjects and interface groups on the measures of fault management performance. A performance analysis based on the state space approach revealed a number of new results, which can be summarised as two general findings:

- AS was the worst performer in the P+F group. AV and IS were roughly at the same level of performance.
- Excepting AS, the P+F subjects had better dynamic and steady state performance than the P subjects.

It is difficult to compare across the trials, as this efforts lacks a common basis for comparison. Generally speaking, performance for non-routine fault management was better for P+F subjects, but it should be noted that this finding was strongly moderated by individual differences.

## 6. INTEGRATION OF FINDINGS ON LONG-TERM ADAPTATION

In this section, we will integrate the findings from this year's and last year's research projects on the characteristics of long-term operator adaptation. The section will be divided into two parts, the first dealing with methodological issues and the second dealing with theoretical issues.

### 6.1 Methodology

The methodological approach we have adopted during these past two years represents a relatively unusual way of investigating operator cognitive activity in NPPs. We have focused exclusively on the data obtained in log files and have adopted a dynamical systems theory (DST) perspective on operator adaptation (see the literature review in Yu et al., 1997a). Typically, when researchers are interested in evaluating how adaptive operators' mental models are, they have instead relied upon verbal protocol analysis. The work that we conducted earlier for JAERI (Christoffersen et al., 1994) and that JAERI itself has conducted (Yamaguchi & Tanabe, 1995;

Yamaguchi, Furukawa, & Tanabe, 1997) are illustrative examples. Which type of methods provide the greatest insight into long-term adaptation?

Based on our experience with each type of methodology, we have come to the conclusion that each has a complementary set of strengths and weaknesses, and thus, can best serve different purposes. Verbal protocol methods seem to be best suited to generating hypotheses, whereas DST measures seem to be best suited to testing hypotheses. In the remainder of this subsection, we will provide a justification for this conclusion.

Verbal protocol analysis. Verbal protocols provide a very rich source of insight into operator's cognitive activities because they tap operators' thought processes relatively directly. However, as is well known, they suffer from a number of limitations. First, protocols usually paint an incomplete picture of cognitive activity. Not all thoughts are verbalised, and some operators provide less verbalisations than others. Second, the protocols are also ambiguous in the sense that they do not provide a direct indication of operators' mental models. Instead, inferences about mental models must be based on the verbal protocols. This inferential step is ambiguous because the relationship between verbalisation and knowledge cannot be determined with certainty. Third, the interpretation of verbal protocol data is subjective because it requires researchers to determine the significance of operators' verbalisations. The meaning is not immediately apparent from the words verbalised. Other limitations associated with verbal protocol analysis have been identified but we will concern ourselves with just these three here.

Researchers have used several techniques to overcome these limitations. The research program conducted by JAERI provides an excellent example (Yamaguchi & Tanabe, 1995; Yamaguchi et al., 1997). The incompleteness problem can be addressed by taking into account the context of cognitive activity. By examining operators' actions, communication patterns, and the state of the work domain, researchers can narrow down the possible interpretations that can be meaningfully attributed to a particular verbalisation. While this technique is very useful and powerful, it is limited because it cannot completely overcome the incompleteness problem. Data interpretation is still a subjective, and thus uncertain, process. Consider, for example, the act of building a diagram of functional relationships based on verbal reports. If variables X and Z are functionally connected indirectly via variable Y, and an operator utters a statement connecting X and Z without mention of Y, we do not know if the operator knows about relationship Y. It is possible that he does not have a complete, deep model of the work domain, and thus only knows

that X and Z are related. On the other hand, it is also possible that the operator does have a complete, deep model of the work domain, and that he merely omitted mention of Y because of a mental shortcut.

Another technique that researchers have used to overcome the limitations of verbal protocol analysis is to formalise the coding of the protocol data, thereby improving the rigour of the analysis. This process makes the data analysis more objective and defensible. Nevertheless, it does not address all of the limitations of verbal protocol analysis. For example, it does nothing to address the fact that protocol data provide an incomplete picture of mental activity. Also, there is still an element of subjectivism in that the same data can be coded in many different ways. Thus, the choice of coding scheme remains an open issue. Furthermore, the increase in rigour associated with formal coding comes at a cost, namely a loss in semantics. The more formal the analysis, the less rich the insights.

The bottom line is that, even if additional precautions are taken, protocol analysis is still a subjective activity that is subject to incompleteness and a less than ideal level of rigour.

DST. Analysis of information in log files using the constructs of DST provides an objective, unambiguous basis for understanding operator adaptation. The data collection is completely automated and the derivation of higher-order variables is precisely specified mathematically. Nevertheless, there are some obvious drawbacks. First, the DST measures may be noisy. For example, they may confound a construct of interest (e.g., context-conditioned variability) with other constructs that are not of interest (e.g., trial and error behaviour) in the same empirical measure. Second, it is a relatively well accepted fact that cognitive processes are underspecified by behaviour. That is, the very same set of actions (i.e., the same log file) could have been generated by different cognitive processes or mental models, some being deep and others being more shallow.

As with verbal protocol analysis, several techniques have been developed to address the limitations of this methodology. For example, the confounding problem may be addressed in some cases by developing more diagnostic DST measures. The set of state space measures we developed based on the abstraction hierarchy provide a good illustration (Yu et al., 1997a). By examining behaviour across all levels of abstraction, we can overcome some of the problem associated with confounding. As another example, the underspecificity problem can be addressed by the introduction of perturbations. The fault trial results discussed earlier show that

perturbations can help distinguish between different classes of cognitive processes. Subjects that seem to behave in the same way under normal trials may be distinguished when they are required to cope with faults. Perturbations of other factors that we have not yet investigated (e.g., work domain parameters, interface graphics) can further distinguish between different types of adaptation. For instance, if a perturbation of a particular graphical element affects performance, then we can infer that that information plays an important role in that operator's mental model. Similarly, if a perturbation of the component time constants does not affect performance, then we can infer that that operator's mental model is deep in that it is robust enough to take into account variations in a structural property of the work domain. Thus, by examining behaviour under a broad and varied set of conditions represented by different classes of perturbations, we can differentiate between the long-term adaptations of different operators.

While these refinements can improve the data obtained from DST analyses, they do not provide a panacea. For example, the questions of what higher-order measures are in fact diagnostic and what perturbations will help differentiate between different levels of adaptation are left unaddressed. Thus, like verbal protocol analysis, the DST methodology does not solve all of our problems.

Reconciliation. Given the comparative advantages and disadvantages just described, we can propose a hybrid approach to the analysis of long-term adaptation that capitalises on the strengths of each approach. The strengths of verbal protocol analysis are its richness and its direct relationship to cognitive processes. We can learn a great deal about how people think about a work domain, what strategies they use, and what information they use in particular situations. Although protocol analysis suffers from problems of incompleteness and subjectivity, these limitations are not of particularly great concern if we are the stage of generating, as opposed to testing, hypotheses. After all, the key goal at that stage would be to get rich and meaningful ideas that could subsequently be tested using methods that are more rigorous and complete.

The strengths of the DST methodology are complementary to those of verbal protocol analysis. Once we have hypotheses about operators' mental models, we are in a much better position to develop diagnostic measures of adaptation based on log files and specific perturbations that can distinguish between different types of adaptation. DST measures can then

provide an objective and unambiguous basis for testing these hypotheses. Thus, DST seems best suited to testing, rather than generating, hypotheses.

Collectively, verbal protocol analysis and DST seem to provide a balanced and complementary approach to the analysis of operator adaptation. Multiple measures are needed to provide converging evidence in dealing with this challenging problem.

6.2 Theory

The analyses of the tuning and fault trials that we conducted for this year's project were not sensitive or controlled enough to provide substantially new theoretical insights. Thus, the bulk of the theoretical contribution of this two-year research program was made in last year's project. Because those insights have already been documented in detail in last year's final contract report (Yu et al., 1997a), we will not repeat them here.

Nevertheless, it is worthwhile pointing out that the results obtained during this year's project are consistent with the theoretical approach that was identified during last year's project. Operator adaptation in complex systems seems to be a special case of the generic systems problem of co-ordination via co-ordinative structures. Proficient operators use the information available in the environment to adapt to deep structural properties of the work domain, as evidenced by low variability at high levels of abstraction and superior performance under perturbations caused by faults. In contrast, less proficient operators, who may not have access to the same information because of poor interface design, must resort to developing ad hoc heuristic rules to control the work domain, as evidenced by low variability at low levels of abstraction. These rules break down under fault conditions, and thereby lead to worse performance under these conditions.

Proficient operators also adapt to their very own information processing limitations, as evidenced by the results for the feedwater configuration task in both startup and tuning trials. The most proficient operators find ways of dealing with work domain and task demands so that the cognitive demands they experience are minimised. In contrast, less proficient performers make the task more difficult than it needs to be. Thus, there seems to be a close relationship between adaptation to information processing constraints and performance.

The most promising theoretical insight obtained in this two-year research program is that it may be possible to develop a generic theory of the systems problem of co-ordination that may generalise across many different physical instantiations. We already have evidence of parallels

between co-ordination in motor control and co-ordination in a human-machine system (Yu et al., 1997a). The possibility of extending these insights to account for co-ordination in social systems and distributed systems composed of human and machine agents is very promising, and thus seems worthwhile to pursue.

## 7. CONCLUSIONS

This was the second year of the project investigating the impact of computer interface design on operators' long-term adaptation. The measures developed for the startup tasks were modified to fit the analysis of tuning tasks and fault management trials, the primarily interest of this year's project. The tuning tasks were designed to be representative of goal changes in systems, and fault management trials were designed to be representative of changes in system structure. In order to respond to these changes, human operators are required to adjust their control strategies. The computer interface design should support operators in handling such situations.

The aim of this research was to understand operators' adaptation to changing system goals and structures. These two changes were designed into the DURESS II system in an attempt to make their management challenging, especially that of changing system structure. It was hoped that these challenges would provide the data necessary to gain insight into the operators' knowledge of the system functions. Many of the measures possible from the toolset developed in last year's contract were applied to the data from the six-month longitudinal experiment to see if more could be learned about subjects' adaptation to changing goals and system structures. Some measures were not as informative as expected, several confirmed what was found in earlier studies, while others revealed new findings as well. In this concluding section, we will describe the contributions, implications, and limitations of this research, and will finish by making some suggestions for future experiments, as well as some promising ideas for future research.

### 7.1 Contributions

Since this was a continuation of last year's project, this research re-confirmed many of the findings made last year. These findings were here expanded to a larger data set that included both tuning and fault management trials.

Research on adaptation has long suffered from a lack of quantitative measurement techniques that are objective and require only modest computational resources. Thus, the first significant contribution of this research was the broadening of the scope of application of the

measures designed during last year's project. This work reconfirmed that measures can be based on a dynamic description of a system at different levels as defined by the abstraction hierarchy.

Following what we did last year, we again have extended Payne et al.'s (1993) work on adaptive strategy selection to a more complex setting. Our research shows that their work can be generalised to predict adaptive steady state behaviour in a dynamic, closed loop system. Furthermore, we have also again extended Payne et al.'s concepts to the study of expertise. Our findings show that subjects who adapt their behaviour to their information processing limitations are usually the subjects with the best performance in their group.

7.2 Implications

In many industrial control systems, the control interface is the only information source and control platform for handling the system. The design of this human-machine interface is critical. This investigation of long-term operator adaptation has important implications for interface design in process control systems, particularly complex sociotechnical systems like NPPs. Because of the limitations of this research (see below), these implications should be considered as tentative hypotheses to be tested more rigorously rather than definitive conclusions.

- It is possible for operators to achieve a very high level of performance under the tuning operating conditions, while using different control strategies. This finding is consistent to what we have found with the normal startup tasks, and also agrees with the contribution made by Edwards & Lees (1974).
- There is a clear difference in performance consistency between the interface groups. The P+F subjects had less variable performance than the P subjects. If we consider that this might be due to the rich information provided by the P+F interface, then subjects using the P interface could not rely fully on the interface information. Instead, they had to resort to their internally-driven rules, and were not as able to adapt to the changes of demand pairs.
- It is also possible to exhibit a very high level of adaptation to system structure and to information processing limitations in the tuning period, while using qualitatively different strategies.
- Subjects who had developed rules for the startup tasks also seem to have used these rules in the tuning tasks. There was an implicit understanding that the rules are task independent.
- There seems to be some connection between performance and the feedwater control strategies the subjects used. Subjects who showed evidence of adaptation to information

processing limitations were generally more proficient than those who did not show such evidence.

- The P+F interface supports the improvement of subjects' performance over the course of experiment.
- The P+F interface also supports fault detection. Subjects who detected faults the fastest were always in the P+F group. The average fault detection time for the P+F subjects was also lower than it was for the P subjects.
- Information that is displayed systematically to support operators in understanding the system structure is critical in fault management.

In general, the P+F interface seemed to support subjects' adaptation to goal changes, as well as to changes in system structures. These advantages are a function of both the information displayed on this interface and the way that it is displayed.

7.3 Limitations

The scope of last year's investigation was deliberately limited to the investigation of normal startup trials as they provided a useful data set for the development and testing of a toolset of novel measures. As a result, last year's findings are valid for startup trials only. In this year's project, we explored the extent to which these measures could be generalised to the tuning and fault management tasks. While this research has been worthwhile, the findings are somewhat disappointing overall. This lack of clarity can probably be attributed to a number of methodological limitations. First, there is the issue of low sample size, especially with respect to fault management trials. Given the duration of investigation necessary to investigate long-term adaptation and the number of fault modes that were investigated, large sample sizes were not practical. Therefore, it is difficult to make any strong claim about the generalisability of the results we have presented. Second, the small sample size also makes it difficult to do statistical tests on these results. All of our inferences are based on visual inspection of graphs and informal comparisons of quantitative data. We do not know whether the patterns we have detected are statistically reliable. Third, some of the experiments were not very well designed for the questions in which we are now interested in asking. There was not enough experimental control for us to meaningfully apply these more sensitive and detailed measures of operator adaptation. For example, in the tuning tasks, the changes in the goal variables were not significant enough; in the fault management trials, the faults either occurred at an improper time or were not noticed.

Fourth, because analysis for fault and tuning tasks started mid-trial, subjects rarely had identical initial conditions, making it difficult to compare results across subjects and conditions.

7.4 Suggestions for Future Experiments

The six-month experiment has been of great help in this program of research investigating the impact of EID on human adaptation. Given the experience gained in this analysis, we make the following suggestions for future experimental design.

- For the tuning trials, the change of demand goals should be larger. To make the tuning task more difficult, tuning should also involve a change in the temperature goal.
- In order to compare across tuning trials, it is necessary that these trials have the same amount of change in the demands and temperatures goals.
- In order to eliminate the effect of different initial conditions for different subjects in the tuning tasks, the experiment could be designed to start the tuning tasks from a pre-run startup trial.
- The tolerance of output demands could be smaller than what they were in the experiment. Narrower tolerances could make the change of demand goals more significant, and could also make it more challenging to stabilise the system.
- As much as possible, faults should be programmed to occur during one control task as opposed to spanning multiple tasks so that performance can be better compared across subjects.
- The time constant for each component could be made larger to make the system more dynamic. In addition, the time constant for different components could be altered.

7.5 Future Research

EID seems to be very effective for computer interface design in process control systems. However, there is still room for improvement from a research perspective. First, in order to use the measures based on the abstraction hierarchy and dynamical system theory to study the fault management trials, more experiments on fault management need to be done, and the design of the experiments should consider the suggestions proposed above. That is, those experiments need to more systematically controlled to address very specific questions using detailed measures of adaptation. Second, we should further study the similarities between motor control and control of complex, dynamical processes. The application of the abstraction hierarchy as a systematic technique for the measurement of adaptation could be a significant contribution to the

existing motor control research. Further, we should explore a general theory of human adaptation in process control systems. This might seem somewhat ambitious but we already have a theoretical foundation for computer interface design, i.e., EID, and we have had a number of qualitative measures that are objective and are proved to be effective.

Last but not least, a more rigid and convincing case study needs to be established to further test the effectiveness of EID. The example should not use a simulator, but rather should involve the physical control of a process with a computer control interface. In the initiation of research, a complex system has to be simplified so that we can focus on important factors that have influence on human operators. Now that we have confidence about the advantages of our methods, we should move on to real systems. It is important to study how complex features of practical systems, such as delay, non-linearity, coupling, and environmental disturbances, have an effect on the validity of our interface design methods.

**REFERENCES**

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.

Edwards, E., & Lees, F. P. (1974). The human operator in process control. London: Taylor & Francis.

Howie, D. E., & Vicente, K. J. (1998). Measures of operator performance in complex, dynamic microworlds: Advancing the state of the art. Ergonomics, 41, 485-500.

Narenda, K. S., & Annaswamy, A. (1989). Stable adaptive systems. Englewood Cliffs, NJ.: Prentice-Hall.

Pask, G., & Scott, B. C. (1972). Learning styles and individual competence. International Journal of Man-Machine Studies, 4, 217-253.

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

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). The adaptive decision maker. Cambridge: Cambridge University Press.

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

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

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

Yamaguchi, Y., Furukawa, H., & Tanabe, F. (1997). JAERI's activities on simulator experiment -- Evaluation methodologies and its facility. In Proceedings of the 1997 IEEE Sixth Conference on Human Factors and Power Plants. Piscataway, NJ: Erlbaum.

Yamaguchi, Y., & Tanabe, F. (1995). Experimental investigation on cognitive behaviour in coping with simulated abnormal situation in nuclear reactor operation. Paper presented at the Post HCI '95 Seminar on Human-Machine Interaction. Kyoto, Japan.

Yu, X., Jamieson, G. A., Khayat, R., Lau, E., Torenvliet, G. L., Vicente, K. J., & Carter, M. W. (1997a). Research on the characteristics of long-term adaptation (97-04). Toronto: University of Toronto, Cognitive Engineering Laboratory.

Yu, X., Khan, F. S., Lau, E., Vicente, K. J., & Carter, M. W. (1997b). ADAPT user's manual: A data analysis tool for human performance evaluation in dynamic systems (CEL 97-03). Toronto: University of Toronto, Cognitive Engineering Laboratory.

Cognitive Engineering Laboratory

# CEL TECHNICAL REPORT SERIES

CEL 93-01 "Egg-sucking, Mousetraps, and the Tower of Babel: Making Human Factors Guidance More Accessible to Designers"
• Kim J. Vicente, Catherine M. Burns, & William S. Pawlak

CEL 93-02 "Effects of Expertise on Reasoning Trajectories in an Abstraction Hierarchy: Fault Diagnosis in a Process Control System"
• Klaus Christoffersen, Alex Pereklita, & Kim J. Vicente

CEL 94-01 "Cognitive 'Dipsticks': Knowledge Elicitation Techniques for Cognitive Engineering Research"
• Klaus Christoffersen, Christopher N. Hunter, & Kim J. Vicente

CEL 94-02 "Muddling Through Wicked Problems: Exploring the Role of Human Factors Information in Design"
• Catherine M. Burns

CEL 94-03 "Cognitive Work Analysis for the DURESS II System"
• Kim J. Vicente & William S. Pawlak

CEL 94-04 "Inducing Effective Control Strategies Through Ecological Interface Design"
• William S. Pawlak

CEL 94-05 "Research on Factors Influencing Human Cognitive Behaviour (I)"
• Klaus Christoffersen, Christopher N. Hunter, & Kim J. Vicente

CEL 94-06 "Ecological Interfaces for Complex Industrial Plants"
• Nick Dinadis & Kim J. Vicente

CEL 94-07 "Evaluation of a Display Design Space: Transparent Layered User Interfaces"
• Beverly L. Harrison, Hiroshi Ishii, Kim J. Vicente, & Bill Buxton

CEL 94-08 "Designing and Evaluating Semi-Transparent 'Silk' User Interface Objects: Supporting Focused and Divided Attention"
• Beverly L. Harrison, Shumin Zhai, Kim J. Vicente, & Bill Buxton

CEL 95-01 "An Ecological Theory of Expertise Effects in Memory Recall"
• Kim J. Vicente & JoAnne H. Wang

CEL-SP "Strategic Plan"
• Cognitive Engineering Laboratory

CEL-LP "Cognitive Engineering Laboratory Profile"
• Cognitive Engineering Laboratory

CEL 95-04 "A Field Study of Operator Cognitive Monitoring at Pickering Nuclear Generating Station-B"
• Kim J. Vicente & Catherine M. Burns

CEL 95-05 "An Empirical Investigation of the Effects of Training and Interface Design on the Control of Complex Systems"
• Christopher N. Hunter

CEL 95-06 "Applying Human Factors to the Design of Medical Equipment: Patient-Controlled Analgesia"
• Laura Lin, Racquel Isla, Karine Doniz, Heather Harkness, Kim J. Vicente, & D. John Doyle

CEL 95-07 "An Experimental Evaluation of Transparent Menu Usage"
• Beverly L. Harrison & Kim J. Vicente

CEL 95-08 "Research on Factors Influencing Human Cognitive Behaviour (II)"
• Christopher N. Hunter, Michael E. Janzen, & Kim J. Vicente

CEL 95-09 "To the Beat of a Different Drummer: The Role of Individual Differences in Ecological Interface Design"
• Dianne Howie

CEL 95-10 "Emergent Features and Temporal Information: Shall the Twain Ever Meet?"
• JoAnne H. Wang

CEL 95-11 "Physical and Functional Displays in Process Supervision and Control"
• Catherine M. Burns & Kim J. Vicente

CEL 96-01 "Shaping Expertise Through Ecological Interface Design: Strategies, Metacognition, and Individual Differences"
• Dianne E. Howie

CEL 96-02 "Skill, Participation, and Competence: Implications of Ecological Interface Design for Working Life"
• Peter Benda, Giuseppe Cioffi, & Kim J. Vicente

CEL 96-03 "Practical Problem Solving in a Design Microworld: An Exploratory Study'
• Klaus Christoffersen

CEL 96-04 "Review of Alarm Systems for Nuclear Power Plants"
• Kim J. Vicente

CEL 96-05 "DURESS II User's Manual: A Thermal-hydraulic Process Simulator for Research and Teaching"
• Lisa C. Orchanian, Thomas P. Smahel, Dianne E. Howie, & Kim J. Vicente

CEL 96-06 "Research on Factors Influencing Human Cognitive Behaviour (III)"
• Dianne E. Howie, Michael E. Janzen, & Kim J. Vicente

CEL 96-07 "Application of Ecological Interface Design to Aviation"
• Nick Dinadis & Kim J. Vicente

CEL 96-08 "Distributed Cognition Demands a Second Metaphor for Cognitive Science"
• Kim J. Vicente

CEL 96-09 "An Experimental Evaluation of Functional Displays in Process Supervision and Control"
• Catherine M. Burns and Kim J. Vicente

CEL 96-10 "The Design and Evaluation of Transparent User Interfaces: From Theory to Practice"
• Beverly L. Harrison

CEL 97-01 "Cognitive Functioning of Control Room Operators: Final Phase"
• Kim J. Vicente, Randall J. Mumaw, & Emilie M. Roth

CEL 97-02 "Applying Human Factors Engineering to Medical Device Design: An Empirical Evaluation of Two Patient-Controlled Analgesia Machine Interfaces"
• Laura Lin

CEL 97-03 "ADAPT User's Manual: A Data Analysis Tool for Human Performance Evaluation in Dynamic Systems"
• Xinyao Yu, Farzad S. Khan, Elfreda Lau, Kim J. Vicente, & Michael W. Carter

CEL 97-04 "Research on the Characteristics of Long-Term Adaptation"
• Xinyao Yu, Renée Chow, Greg A. Jamieson, Rasha Khayat, Elfreda Lau, Gerard L. Torenvliet, Kim J. Vicente, & Michael W. Carter

CEL 98-01 "Applying Human Factors Engineering to Medical Device Design: An Empirical Evaluation of Patient-Controlled Analgesia Machine Interfaces"
• Laura Lin

CEL 98-02 "Building an Ecological Foundation for Experimental Psychology: Beyond the Lens Model and Direct Perception"
• Kim J. Vicente

CEL 98-03 "Cognitive Work Analysis: Towards Safe, Productive, and Healthy Computer-based Work"
• Kim J. Vicente

CEL 98-04 "Ecological Interface Design for Petrochemical Processing Applications"
• Greg. A. Jamieson & Kim J. Vicente

CEL 98-05 "The Effects of Spatial and Temporal Proximity of Means-end Information in Ecological Display Design for an Industrial Simulation"
• Catherine M. Burns

CEL 98-06 "Research on Characteristics of Long-term Adaptation (II)"
• Xinyao Yu, Gerard L. Torenvliet, & Kim J. Vicente