ERGONOMICS, 1998, VOL. 41, NO. 4, 485 – 500

Measures of operator performance in complex, dynamic microworlds: advancing the state of the art

DIANNE E. HOWIE and KIM J. VICENTE†*

Centre for Applied Cognitive Science, Ontario Institute for Studies in Education, University of Toronto, Toronto, Ontario, Canada

†Cognitive Engineering Laboratory, Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, Ontario, Canada M5S 3G8

Keywords: Process control; Cognitive engineering; Performance measurement: Microworlds; Expertise.

Microworld research provides a useful complement to field studies and highly controlled laboratory studies, aiming to strike a balance between representativeness and experimental control. Yet microworld research has associated methodological difficulties, particularly the problem of performance measurement. Researchers generally adopt a variety of measures to provide converging evidence concerning questions of interest. To confront problems with existing measures, this paper examines a series of objective measures used to characterize the performance of human operators in process control. These measures include novel, quantitative extensions to existing graphical analyses and new graphical representations. The measures are applied in the context of a 6-month longitudinal study using an interactive, thermal-hydraulic process control microworld (DURESS II). The following measures are discussed: steady-state time, action transition graph complexity, the path length in state space diagrams, the area under distance-to-goals graphs, divergence from the temperature goal line in mass inventory versus energy inventory graphs, and the proportion of control actions near the beginning of the trials represented by timelines. Two case studies emphasize the performance and strategy differences of individual operators across the battery of measures.

1. Introduction

> When you can measure what you are speaking about . . . you know something about it, but when you cannot measure it . . . your knowledge is of a meagre and unsatisfactory kind.
>
> Lord Kelvin, 1883 (cited in Weiner 1994: 203)

Conducting cognitive engineering research is a methodologically challenging endeavour (Sheridan and Hennessy 1984). On the one hand, there is a strong need for research results to generalize to operational settings, otherwise the research is of little practical use. This requirement implies that experiments need to be conducted under conditions that are *representative* (Brunswik 1956) of actual work domains. On the other hand, there is also a strong need for defensible research results that can lead to a principled understanding of the factors that affect human performance in

*Author for correspondence.

0014–0139/98 $12.00

complex systems. This requirement implies that experiments need to be conducted under controlled conditions, where theoretical principles can be rigorously tested. To maximize representativeness, studies could be conducted with full-scope simulators, but experience has shown that it is very difficult to obtain defensible and statistically reliable results under such conditions, primarily due to lack of experimental control (Baker and Marshall 1988). To maximize control, one can greatly simplify the phenomenon of interest, but this can lead to results that are not relevant to industrial systems (Chapanis 1967). Some cognitive engineering researchers have suggested microworlds — small-scale computer simulations of systems that are intended to be representative of industrial-scale complex human machine systems — as a useful alternative to complement field studies and highly controlled laboratory studies (Brehmer and Dörner 1993). Their rationale is that computer-based microworlds are research vehicles that can lead to defensible results that are also potentially generalizable, thereby generating useful and informative findings.

Microworld research has been around for years (Morris *et al.* 1985), although it has only recently been labelled as such. During this time, researchers have commented that there are many methodological difficulties associated with such research, since many traditional laboratory methods cannot be meaningfully applied in such rich contexts (Moray *et al.* 1986). Performance measurement is a particularly thorny problem for several reasons. First, because microworlds generally present participants with many degrees of freedom, there is no 'one right way' to perform the task. Different, equally-acceptable strategies are possible, making comparisons between participants difficult and averaging across participants not very meaningful. Second, because participants are usually given a great deal of practice in microworld tasks (so that the results may generalize to experienced operators), coarse measures of performance (e.g. task completion time, number of errors) rarely reveal differences between treatment groups (Pawlak and Vicente 1996). Third, because so many data are typically generated by microworld experiments, it is very difficult to identify meaningful patterns of findings. Fourth, because the systems usually have slow dynamics, experienced participants tend to act on the system infrequently. As a result, there can be significant periods of time during which there is no overt behaviour to record. Researchers have concluded that the only way to deal with these measurement problems is to adopt a diverse, complementary battery of measures that can lead to converging evidence pertaining to the questions of interest. A number of different measures have been proposed and have been shown to be useful (Moray *et al.* 1986, Moray and Rotenberg 1989, Sanderson *et al.* 1989, Pawlak and Vicente 1996).

Nevertheless, there are significant limitations with the existing set of methods. For instance, *verbal protocol* data (Sanderson *et al.* 1989, Pawlak and Vicente 1996) are difficult to interpret objectively and can lead to a conservative picture of mental activity if participants fail to verbalize regularly (Rasmussen and Jensen 1974). Also, some methods, such as *eye movement* analysis (Moray and Rotenberg 1989) and verbal protocol analysis, are extremely time consuming, thereby discouraging researchers from thoroughly analysing their data. Furthermore, other methods, such as *action transition graphs* and *time history plots* (Moray *et al.* 1986), have been presented qualitatively as graphical overviews of data but have not been objectively measured in a quantitative fashion. Thus, differences between individuals are only identified informally by visual inspection of the graphs. Finally, still other methods, such as the *state-space* technique described by Sanderson *et al.* (1989), are based on

qualitative interpretations of actions and, as a result, can lead to misleading measurements.

This last point is perhaps subtle, so it is worthwhile elaborating on. In the method of Sanderson *et al.* (1989), the current state of the system is compared to the goal state, and the minimum number of actions that are needed to reach the goal state are determined in a qualitative manner. For example, if the volume and temperature of water in a reservoir are both currently too low, then the optimal actions might be to increase the input valve setting (to increase the input flow rate, thereby increasing volume) and to increase the heater setting (to increase the heat transfer rate, thereby increasing temperature). The problem with this technique is that it ignores the magnitude of the participant's actions, thereby leading to occasionally inaccurate measurements. To follow with the same example, if a participant increased the heater setting by only a small amount and increased the valve setting by a large amount, the actions would be classified as being in the same category as the optimal set of actions that would take the participants directly to the goal state. However, because of the magnitude of the changes, the state of the system could actually go *away from the goal state* because the temperature may decrease, rather than increase, because the increase in heater setting can be offset by the large increase in the flow of cold water entering the tank. It is not known how often such pathological results are obtained in practice, but the fact that they can occur is a cause for concern.

This is not to say that the methods that have been proposed in the literature do not have their advantages. However, it is clear that there is significant room for improvement. The contribution of this paper is to address the limitations identified above. Below, a set of performance measures that are objective and quantitative will be presented. These measures differ from others that have been proposed in several ways. Some measures are quantitative ways of objectively measuring data that have previously only been presented in a qualitative, informal manner using graphical representations. Other measures are novel quantitative measures that overcome the aforementioned limitations of qualitative measures. All of the measures are objective in the sense that they are obtained solely from log files of participant actions and system state. As a result, the derivation of these measures can be formalized and therefore automated, saving a great deal of time in data analysis.

These contributions to performance measurement are presented in the context of a 6-month longitudinal study of human performance in a process control microworld. The main objective of this paper is to illustrate how the proposed measures can show differences in expertise as a function of the amount of practice in controlling the system. Accordingly, the authors will *not* present a detailed discussion of each participants' performance on each measure. Instead, a comparison of the performance of two participants will be made to illustrate the value of the proposed measures.

2. Method

2.1. *Experimental design*

This study employed a between-participants design with two factors: Interface and Block (early or late) with repeated measures for Trial. The early and late blocks of trials include the first 20 trials and last 20 normal, non-fault trials over a 6-month period. This paper will not discuss the effect of the interface on operator

performance, but will emphasize the changes in operator performance with experience, and individual differences in performance. The role of interface design is discussed in more detail in Christoffersen *et al.* (1996a, b, 1997) and Howie (1996).

2.2. *Participants*

The participants were six males ranging in age from 23 to 32 years. Five of the six participants had a science or engineering background.

2.3. *Apparatus*

The experiment was conducted using the DURESS II (Dual Reservoir System Simulation) thermal-hydraulic process control simulation (figure 1). DURESS II is a simplified, real-time computer model of a process control system, which exhibits some characteristics that are representative of a complex system (Vicente and Rasmussen 1990). For instance, the system variables are coupled and have time lags. In addition, there is a degree of risk involved, since ineffective control may cause the system to fail and the trial to end prematurely (for example, heating an empty reservoir).

Duress II contains two redundant feedwater streams, each containing one pump and three valves (e.g. pump PA, and valves VA, VA1, and VA2 for feedwater stream A). Participants may use any or all of these pumps and valves to meet the external

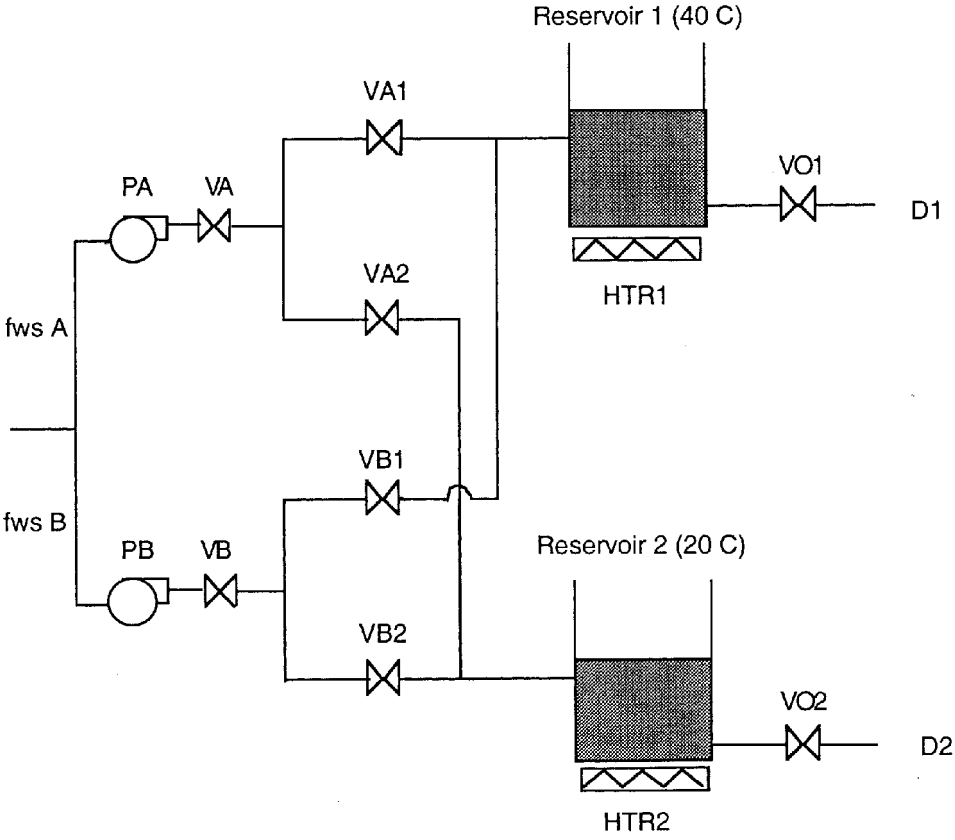

Figure 1. Mimic diagram for DURESS II.

demands for water. There is also one valve that controls the output flow from each reservoir (for example, VO1 for reservoir 1). Water enters the system at a temperature of 10°C, and is heated to the required temperatures using the heater in each reservoir (for example, HTR1 for reservoir 1).

During a trial, users may control the system by adjusting the settings of the heaters, pumps, and valves. The goal is to satisfy the required output water flow rates (demands) and temperatures for each of two reservoirs. The temperature goals for each reservoir remain the same for each trial (40°C and 20°C), while the demand pairs vary between trials.

2.4. *Procedure*

The experiment included a total of 224 trials per participant on the DURESS II simulation, spanning a period of 6 months. The participants were gradually introduced to the experimental tasks during the first month. First, participants learned to take the system from a shutdown state to steady state in the start-up task. Steady state was achieved when the participant met the temperature and demand goals for both reservoirs for 5 consecutive minutes. Shutdown and tuning tasks were subsequently added, but will not be discussed in this paper.

For the next 5 months, each trial consisted of start-up, tuning, and shut-down tasks performed contiguously. Various faults were distributed randomly, and infrequently, throughout the trials. The participants were not informed of the results of each trial, although they could receive feedback by observing the elapsed time at any point in the trial, and any system failure messages displayed on the screen. These messages simply stated which component had failed (for example, 'Reservoir 1 heated empty'). The trial was terminated immediately after a system failure message was displayed.

2.5. *Performance measures*

Whenever a participant made a control action, the component being manipulated, the time, and the state of the system variables were all recorded. For each trial, these values were stored in a log file, making them available for further analysis. All of the measures discussed were derived from these log files.

3. Results

The first 20 trials and the last 20 non-fault trials were examined in detail to investigate differences in the participants' performance with experience. Only the start-up tasks were analysed. This portion of the data provides a consistent base for comparison of the participants' strategies. (For other analyses of data from this experiment, see Christoffersen *et al.* 1996a,b, 1997).

3.1. *Steady-state times*

Steady-state time is the time required for a participant to bring the system from a shut-down state to the goal state and to keep it within tolerance for five consecutive minutes. This is a common measure of operator performance. Generally, as operators gain experience, they are able to reach the goals more quickly, and to maintain the system in the goal state. Steady-state time has the advantage of being an objective, quantitative measure that is easy to obtain. However, steady-state times do not give any insight into the strategies that the participant used to reach the goal state. Steady-state time is a *product* measure,

whereas the other measures discussed in this paper emphasize the *process* that the participant used to reach the goals.

Over the 6-month experiment, the participants' mean steady-state times decreased substantially from 645.7 to 394.2 s ($F(1,5) = 58.44$, $p < 0.001$). The mean steady-state times for each operator are summarized in table 1. There was also a significant effect for Trial ($F(19,94) = 3.80$, $p < 0.0001$), and a Block $\times$ Trial interaction ($F(19,66) = 4.33$, $p < 0.0001$). The steady-state times decreased rapidly in the first block of trials as the participants learned to operate DURESS II, but remained relatively constant in the last block of trials after the participants became experienced with the system.

As shown in table 1, AV had some of the fastest steady-state times while AS was generally slower, despite using the same interface. Thus, these two participants can be viewed as prototypes of high and low expertise, respectively. In the remainder of the paper, the authors will focus primarily on contrasting the performance of these two prototypical participants to see if the novel measures proposed can help to explain why these participants differed in expertise.

3.2. *Action transition graphs*

Action transition graphs reveal aggregate sequential relationships in behaviour (Moray *et al.* 1986). Each component that can be acted on is represented by a node (e.g. PA, PB, VB, etc.), and those nodes that are accessed in sequence are joined by a line. The thickness of a line joining any two nodes is proportional to the frequency of that transition. The loops above some components show that the participant adjusted the same component repeatedly. Figure 2 shows the action transition graphs for the two prototypical participants, AV and AS, on their first completed trial and last normal trial.

The graphs for each participant generally became less complex with experience, indicating both that the participants made fewer control actions and that their control actions were more sequentially consistent. These findings agreed with the qualitative findings of Moray *et al.* (1986). Further, the action transition graphs illustrate the relative skill level of the participants. The action transition graphs of AS contain a tangled web of connecting lines even after substantial experience with the system. AV's graphs are much simpler, illustrating his greater relative skill. The dominant dark lines connecting heaters 1 and 2 (HTR1 and HTR2), particularly in the first trial of AS, show that AS had more difficulty controlling the heaters than AV.

Table 1. Mean steady state times.

	Steady state time (s)	
Participant	Block 1	Block 2
AS	911.2	437.2
AV	529.0	353.5
IS	647.5	399.0
ML	663.0	390.2
TL	485.2	357.4
WL	639.8	437.2

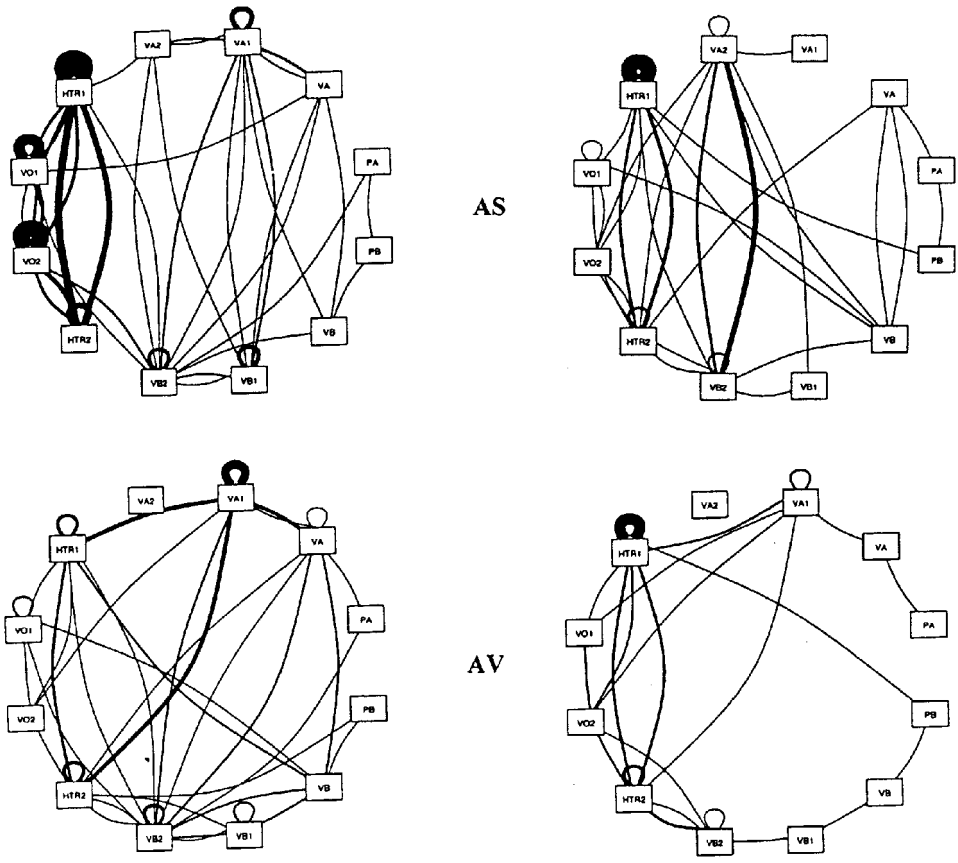

Figure 2. Comparison of action transition graphs for the first (left) and last (right) trials for participants AS and AV.

These graphs also give some insight into the participants' strategies for controlling the valves. In the action transition graph for AV's first trial the node representing VA2 is not connected by any lines, thus AV did not use this valve during the trial. In contrast, AS consistently used all of the valves to control the system. Using fewer valves simplifies the operation of DURESS II by minimizing the interactions between the two feedwater streams. Thus, the action transition graphs reveal that differences in strategies may explain why AV was a more proficient performer than AS.

Although much can be learned through an informal inspection of the action transition graphs, these qualitative assessments should be augmented by a quantitative measure of 'complexity'. In a novel extension to the approach of Moray *et al.* (1986), this study operationalized complexity as the number of lines in an action transition graph. This measure of action transition graph complexity has the advantage of being both objective and precise, as compared to the visual judgements on which previous researchers relied. Complexity is also easily derived from the log files. The mean complexities for each participant are summarized in table 2.

Table 2. Action transition graph complexity.

	Complexity	
Participant	Block 1	Block 2
AS	48.1	31.5
AV	24.8	23.4
IS	25.7	22.0
ML	30.7	28.3
TL	21.0	21.6
WL	28.5	22.1

There was a marginally significant decrease in the action transition graph complexity from a mean of 29.7 to 24.7 between the first and last blocks of trials ($F(1,5) = 5.31$, $p = 0.07$). There was also a significant effect for Trial ($F(19,94) = 2.34$, $p < 0.005$), and a marginally significant Block $\times$ Trial interaction ($F(19,66) = 1.69$, $p = 0.06$). Thus, the complexity generally decreased from the first to the last trial within each block. Further, the rate of change of complexity differed within each block, with the more rapid decrease in action transition graph complexity occurring in the first block of trials. However, this particular measure has limited sensitivity, at least with the low sample size of this experiment.

3.3. *State space diagrams*

State space diagrams portray the system state with respect to the goal state (Sanderson *et al.* 1989). In DURESS II, the goals are two-fold: to meet the required temperature and demand for each reservoir. Thus, the state space diagrams for DURESS II consist of a plot of temperature vs. demand for each reservoir over the course of a trial. The water outflow demands varied from one trial to the next and between reservoirs (within the range of 1 to 20 units/s). Although the temperature goals were constant between trials, they were different for each reservoir (40°C for reservoir 1 and 20°C for reservoir 2). To allow the state space diagrams to be compared between reservoirs and across trials, both the temperatures and demands were normalized with respect to the goals. After normalization, the point where the demand and temperature are both equal to 1 corresponds to the goal state on each diagram. Figure 3 shows the state space diagrams for the two prototypical participants, AV and AS, on their first completed trial and last normal trial.

The path that these two participants took between the initial state and the goal state became more direct with experience. AS's early state space diagrams contained considerable oscillation around the goal state, as illustrated in figure 3. This demonstrates that AS had difficulty keeping the temperature constant once he had reached the goal state. Overall, AV's state space diagrams show a more direct route toward the goal than the corresponding diagrams for AS. Thus, AV was a more proficient performer than AS because he exhibited more stable control.

The state space diagrams also provide some insight regarding the participants' strategies for reaching the goal state. Even after considerable experience, none of the participants took the most direct path through the state space towards the goal (a diagonal). The best participants, such as AV, first reached the temperature goal and only then began to approach the demand goal. By keeping the output valves closed

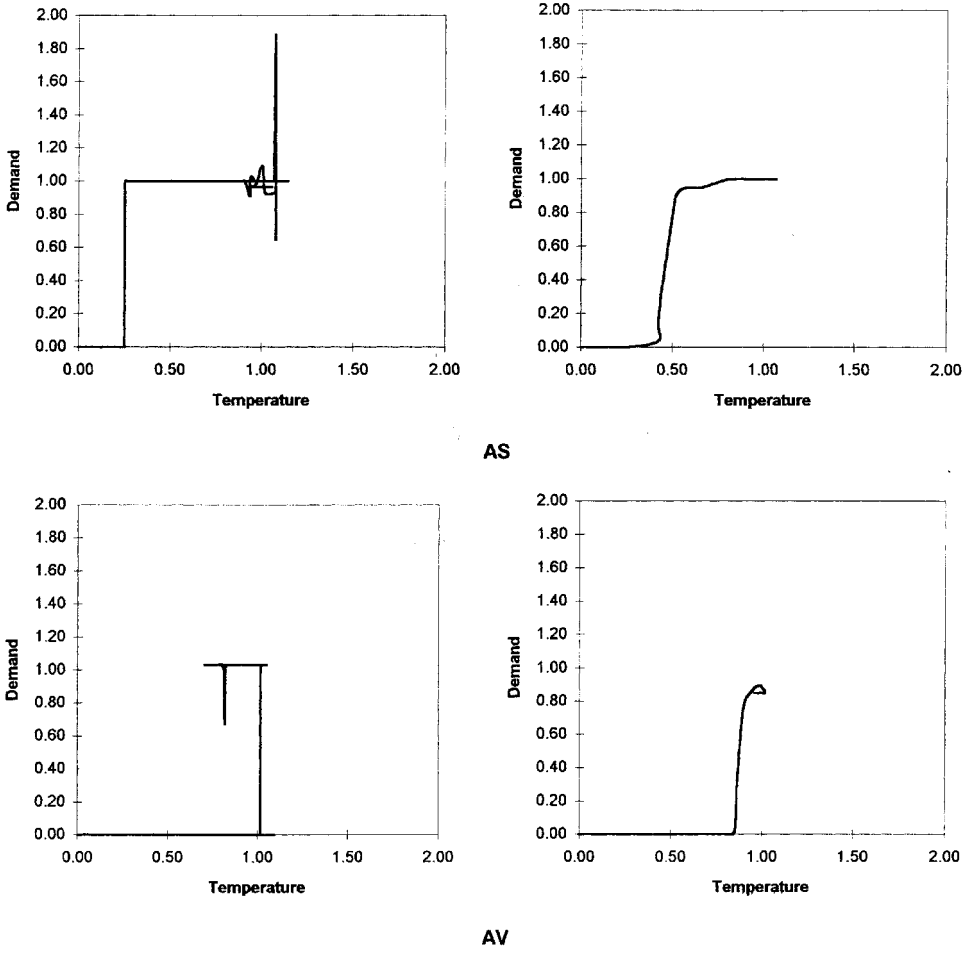

Figure 3. Comparison of state space diagrams (Reservoir 1) for the first (left) and last (right) trials for participants AS and AV.

Table 3. Mean path length in state space.

	Path length in state space			
	Reservoir 1		Reservoir 2	
Participant	Block 1	Block 2	Block 1	Block 2
AS	8.26	2.13	9.08	2.32
AV	2.59	2.50	2.60	2.23
IS	3.41	1.94	2.97	2.09
ML	2.48	2.99	2.53	4.33
TL	3.00	1.99	2.28	2.01
WL	4.66	2.42	2.39	3.38

until the water in the corresponding reservoir was at the goal temperature, these participants ensured that only appropriately heated water left the system to meet the demands. In an industrial setting, this strategy would translate into energy savings, and possible downstream implications for temperature-sensitive operations.

As discussed in Sanderson *et al.* (1989), a substantial amount of information may be derived from a qualitative examination of state space diagrams. However, this analysis is lacking in precision. These subjective observations can be extended by adopting a quantitative measure of the directness of a path to the goal state — the path length in a state space diagram. The mean path lengths for each participant are summarized in table 3.

The path lengths decreased from a mean 4.0 to 2.3 units for reservoir 1, and from a mean of 3.8 to 2.8 units for reservoir 2. For reservoir 1, this Block effect was marginally significant ($F(1,5) = 5.95$, $p = 0.06$), and the Trial effect was significant ($F(19,94) = 1.91$, $p < 0.05$). For reservoir 2, there was a significant effect for Block $\times$ Trial ($F(19,66) = 2.03$, $p < 0.05$), reflecting the different rates of change of path length in each block of trials. AV's mean path length is shorter than that of AS for the first block of trials, demonstrating his greater initial skill level. However, after 6 months of experience with the system, the path lengths of the participants became comparable, perhaps suggesting that this measure loses its power to distinguish between participants after substantial experience. This possibility needs to be investigated further.

Shaw *et al.* (1992) suggest that the distance-to-goal is also a useful measure of performance. Distance-to-goal graphs can be derived from state space diagrams,

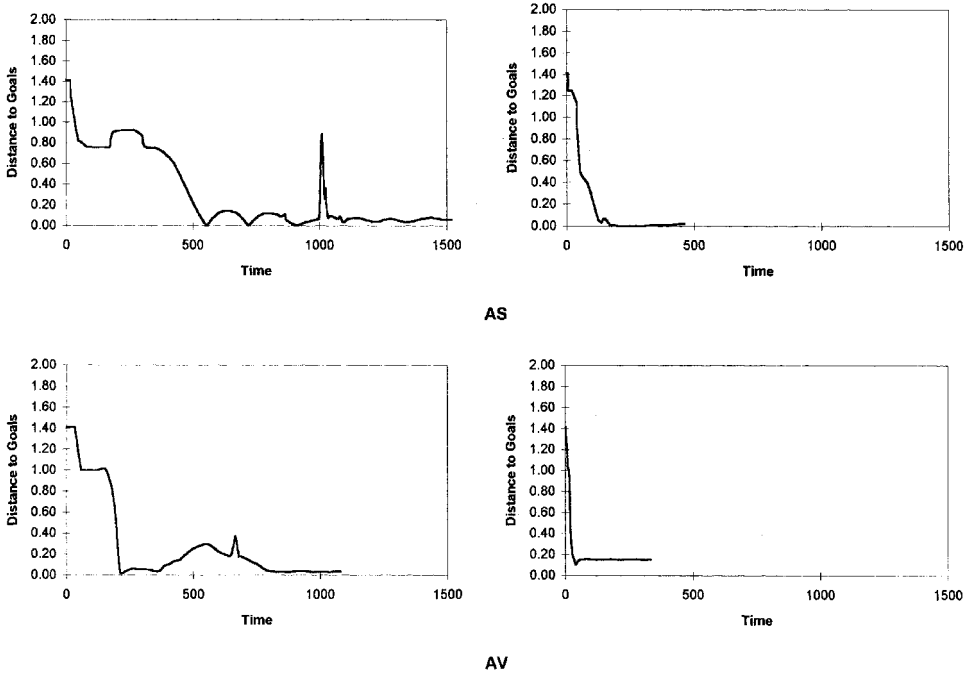

Figure 4. Comparison of distance-to-goal versus time graphs (Reservoir 1) for the first (left) and last (right) trials for participants AS and AV.

providing the representation of time that is lacking in state space diagrams. In this study, the shortest distance from the system state to the goal state was taken as the measure of the distance to the goals. Figure 4 shows the Euclidean distance-to-goal versus time graphs for the two prototypical participants, AV and AS, on their first completed trial and last normal trial.

It is immediately obvious from these graphs that both AV and AS completed their trials more quickly with experience, since their distance-to-goal graphs terminate sooner on the time axis. With experience, both participants also managed to reach the goal state (distance to goals = 0) more quickly and to maintain the system state within tolerance once they had reached the goals. The stability differences between AV and AS identified earlier can also be seen here. Particularly in his early trials, the distance goal graphs of AS include many spikes demonstrating rapid divergence from the goal state and then his subsequent efforts to bring the system back under control.

Ideally, the participants would like to be within tolerance of the goals at the start of the trial (time = 0) and simply maintain this state for the required 5 min. This is not physically possible, since the system requires some time to get water into the reservoirs and heat this water to the goal temperatures. A useful way to quantify the participants' performance is to measure how far the participants deviate from this ideal by calculating the area under the distance-to-goal graphs. The closer the area is to zero, the better the participants' performance according to this criterion. Table 4 gives the area under the distance-to-goals graphs.

Both the means and standard deviations decreased between the first and last blocks of trials. Further, AV's greater relative skill level is confirmed by the lower average area under his distance-to-goals graphs, as compared with AS. This quantitative measure allows the performance of the participants to be aggregated over a range of trials. Further, it allows finer distinctions to be made between the abilities of the participants, particularly in the final block of trials.

3.4. *Mass and energy inventories*

Mass inventory versus energy inventory graphs are similar to state space diagrams in that they provide a view of the system state with respect to the goal state. Mass inventory versus energy inventory graphs are novel since they show a different aspect of the participants' route to the goals. The temperature goals appear as straight diagonal lines on these graphs since temperature is proportional to the ratio of

Table 4. Area of deviation from temperature goal line.

	Area in state space			
	Reservoir 1		Reservoir 2	
Participant	Block 1	Block 2	Block 1	Block 2
AS	361.0	134.7	89.4	33.6
AV	130.2	78.2	23.3	16.3
IS	216.4	80.7	55.6	19.7
ML	189.6	61.9	47.3	21.5
TL	74.2	29.5	42.1	16.6
WL	257.0	105.3	66.6	27.3

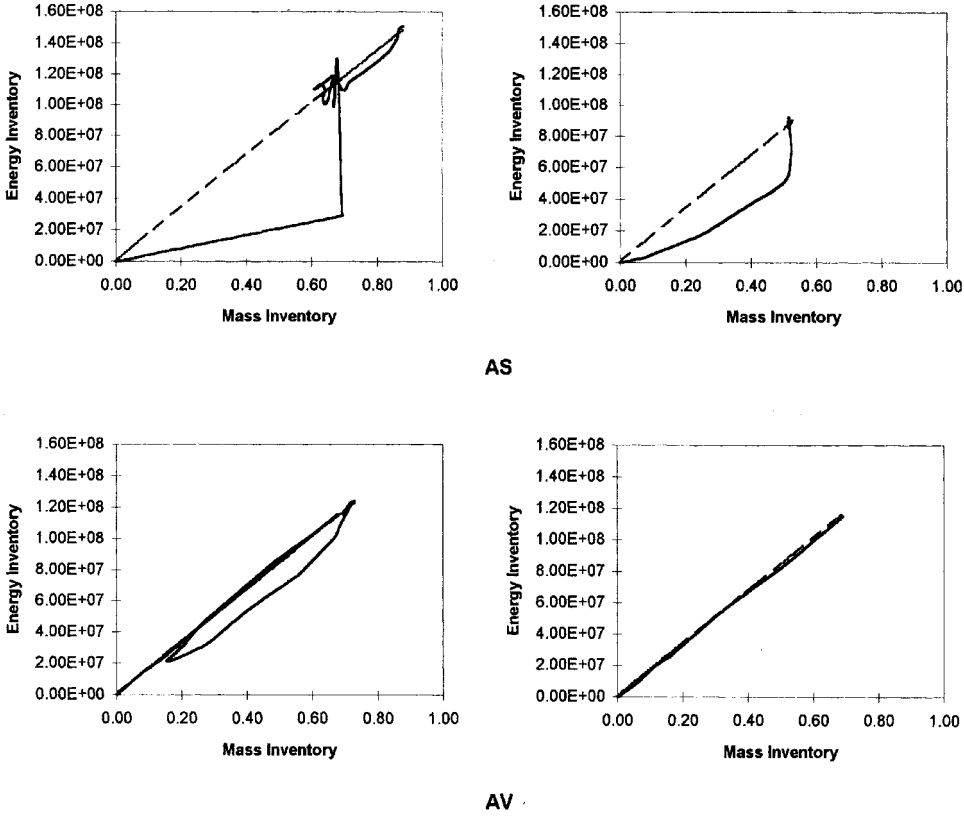

Figure 5. Comparison of mass inventory versus energy inventory graphs (Reservoir 1) for the first (left) and last (right) trials for participants AS and AV. The dotted line shows the temperature goal line.

energy inventory to mass inventory. The participants can satisfy the temperature goal at any point along this line, the exact location depending upon their strategies. The participants will finish the trial further to the right on the temperature goal line, depending on the volume of water they maintain in the reservoirs.

Figure 5 shows the graphs for the prototypical participants, AV and AS, on their first completed trial and last normal trial. The divergence of the participants' paths (solid line) from the temperature goal line (dotted) became less with experience for all of the participants. The difference in the performance of AV and AS is particularly dramatic in this representation. Even after 6 months of experience with the system, AS's path through mass inventory versus energy inventory space shows marked deviation from the temperature goal line. He converges with the temperature goal line at only one point. In contrast, after the same experience with the system, AV's path closely hugs the temperature goal line along a large range of mass inventories. Thus, AV was more proficient than AS because he was able to coordinate changes in mass and energy in a way that was almost perfectly tailored to the goal setpoints.

Note that both of these participants had relatively high volumes of water in their reservoirs at the end of their final trials—from one-half to over two-thirds of capacity. Although a large volume of water heats more slowly than a small volume, a

large volume of water is useful since it keeps the temperature in the reservoirs more stable. Further, the goal temperature tolerance on the energy inventory is larger with higher volume, making it easier to keep the temperature in the goal region (Pawlak and Vicente 1996).

It is more difficult to quantify the relationships in these graphs since, unlike the state space diagrams, the goal on the mass inventory versus energy inventory graphs is not a point, but a line. Thus, there is no specific minimum distance between the starting point and the goal. The temperature goal may be met with differing reservoir volumes (mass inventories), and this is proportionately reflected in the energy inventory. When comparing the deviations from lines of unequal length, the area is confounded by the length. Therefore, measures of expertise are limited to those derived from a qualitative inspection of the graphs. Despite the lack of quantitative measures, mass inventory versus energy inventory graphs provide a useful alternative perspective of the participants' strategies and relative skill levels.

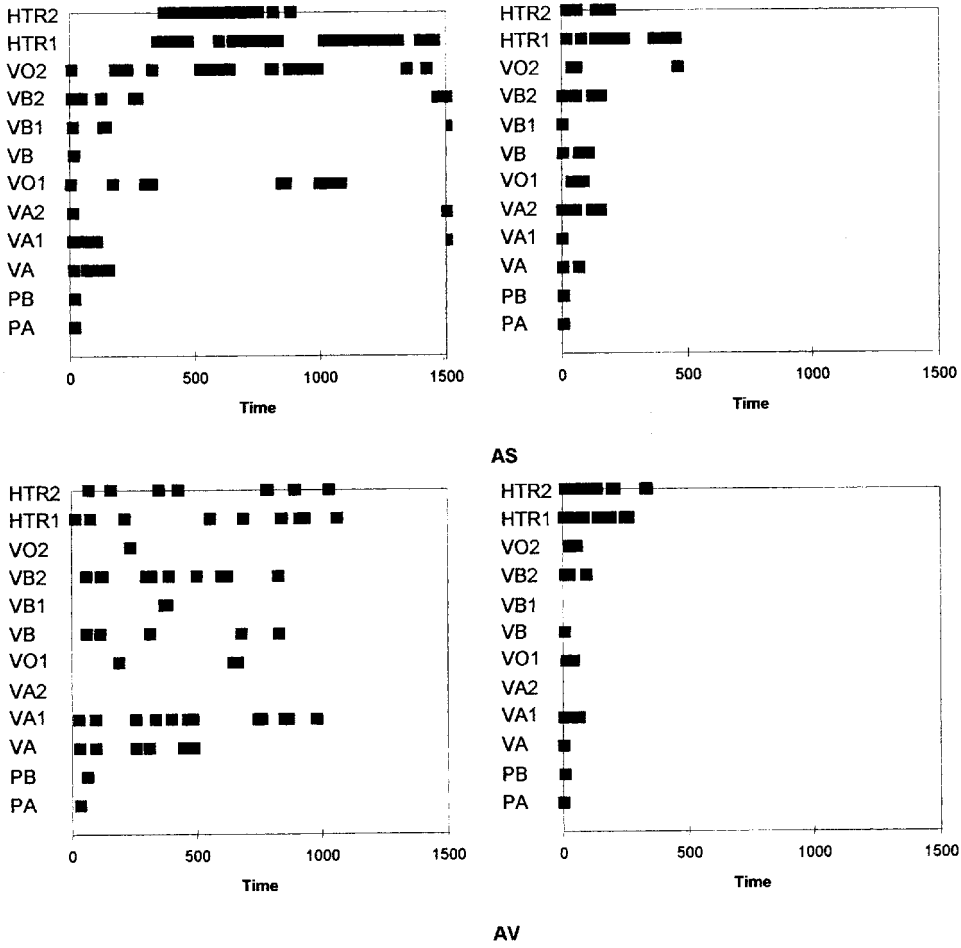

Figure 6. Comparison of time lines for the first (left) and last (right) trials for participants AS and AV.

3.5. *Timelines*

Timelines provide a way of visualizing a sequence of actions over time (Moray *et al.* 1986). Control actions are plotted against time on the horizontal axis, and are grouped according to the component acted upon on the vertical axis. Each time a component setting is changed, this control action is indicated by a square on the graph. Timelines do not indicate the magnitude of a control action, but they preserve the order.

Figure 6 shows the timelines for the two prototypical subjects, AV and AS, on their first completed trial and last normal trial. Immediately, one can see that the trial times become shorter with experience, causing the control actions to be clustered closer to the *y*-axis in the later trials for both participants. Also, both AV and AS performed fewer control actions with experience, replicating the results of Moray *et al.* (1986). The type of actions made by AV and AS illustrate their different levels of expertise in controlling DURESS II. Throughout the experiment AS made more frequent heater control actions than AV, reflecting his difficulty in operating the heaters.

The distribution of control actions within a trial also changed with experience. In the final trials, the participants seem to have a larger proportion of their control actions near the beginning of the trial, even when the length of the trial is disregarded. In order to empirically evaluate the changing distribution of control actions, the percentage of control actions in the first quarter of each trial was calculated. Table 5 summarizes these results. Despite large individual differences, there was a significant difference in the proportion of control actions in the first quarter of the trial (χ^2 $(1, N = 210) = 10.84$, $p < 0.01$). The actual percentage of control actions in the first quarter of the trial seems most related to the participants' heater control strategies rather than directly to their skill in operating DURESS II (Howie 1996).

4. Discussion

There is no one right way for a participant to complete a trial in the DURESS II microworld. A variety of strategies are possible and appropriate, making individual analyses of participant data more meaningful than group averages. While traditional measures, such as steady-state time, may show differences in expertise, they do not offer much insight into the origin of those differences. Strategy differences among the participants can be better understood with new types of performance measures of the type that have been proposed in this paper. Existing qualitative graphical analyses were supplemented by novel, quantitative measures. In particular, measures used for

Table 5. Percentage of control actions in first quarter of trial.

	Control actions in first quarter (%)	
Participant	Block 1	Block 2
AS	30	48
AV	45	70
IS	36	65
ML	47	83
TL	39	58
WL	38	41

comparing action transition graphs, state space diagrams, distance-to-goals graphs, and timelines (or time history plots) were introduced. These performance measures address methodological limitations in existing microworld research (Moray *et al.* 1986, Sanderson *et al.* 1989). A new graphical measure — mass inventory versus energy inventory graphs — was also introduced.

Not surprisingly, almost all of the measures showed that participants developed greater expertise over a 6-month period. More importantly, however, these measures were also useful in examining the strategy differences and relative skill levels of individual participants. These insights were revealed by comparing the performance of two prototypical participants, AS and AV, on various measures. Although some measures were not as sensitive as we would have hoped, as a whole they showed why AV was a more proficient performer than AS. For instance, AV exhibited more stable control of the system than did AS. One of the sources of difficulties for AS seemed to be with stabilizing the reservoir temperatures, as evidenced by his frequent heater actions. AS repeatedly adjusted the heater settings because he was not able to stabilize temperature in the goal regions quickly and efficiently.

AV also exhibited a more refined level of adaptation to system constraints than did AS. For instance, AV used fewer valves to control the system whenever possible. This strategy allowed him to minimize the interactions between the two feedwater streams, thereby reducing the demands he experienced to perform the task. In contrast, AS performed the task in a way that made it more difficult than it needed to be. AV was also able to coordinate changes in mass and energy in a way that was exquisitely sensitive to the task and system constraints. He was able to track the temperature goal line almost perfectly throughout the trial. In contrast, the behaviours of AS were not nearly as well adapted, even after 6 months of quasi-daily practice. Collectively, these results show that the measures that have been described in this paper advance the state of the art in performance measurement in complex, dynamic microworlds such as DURESS II.

Although DURESS II was designed to be representative of industrial-scale complex human-machine systems, it is still far simpler than industry-scale complex, dynamic systems. Thus, an important next step in this research will be to apply these methods to more complex environments, such as simulators and field settings. For example, the path length, distance to goal, and area under distance-to-goal measures can be applied to any system whose goals can be quantitatively measured over time. Also, the mass vs. energy inventory graphs can be meaningfully used in any physical system that has temperature as a goal variable. It may even be possible to generalize this measure to any goal variable that has two contributors that can be varied independently or quasi-independently. In addition, the action transition graph, and percentage of control actions measures can be applied to any system that involves manual control. Note that many, if not all, of these measures can be applied to both discrete and continuous systems. To conclude, now that this battery of measures has proven useful in a microworld, these measures should be put to use in more general and challenging applications. There is no better test of the value of cognitive engineering research than its applicability to complex operational settings.

Acknowledgements

This research was sponsored by a research contract from the Japan Atomic Energy Research Institute (Dr Fumiya Tanabe, contract monitor), and by grants from the Natural Sciences and Engineering Research Council of Canada. The authors would

like to thank Klaus Christoffersen and Chris Hunter, who collected the data, and the six participants who completed this experiment, for their Herculean efforts.

References

Baker, S. and Marshall, E. 1988, Evaluating the man-machine interface—The search for data, in J. Patrick and K. D. Duncan (eds), *Training, Human Decision Making and Control* (Amsterdam: Elsevier), 79–92.

Brehmer, B. and Dörner, D. 1993, Experiments with computer-simulated microworlds: escaping both the narrow straits of the laboratory and the deep blue sea of the field study, *Computers in Human Behaviour*, **9,** 171–184.

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

Chapanis, A. 1967, The relevance of laboratory studies to practical situations, *Ergonomics*, **10,** 557–577.

Christoffersen, K., Hunter, C. N. and Vicente, K. J. 1996a, A longitudinal study of the effects of ecological interface design on skill acquisition, *Human Factors*, **38,** 523–541.

Christoffersen, K., Hunter, C. N. and Vicente, K. J. 1996b, A longitudinal study of the impact of ecological interface design on deep knowledge. Manuscript submitted for publication.

Christoffersen, K., Hunter, C. N. and Vicente, K. J. 1997, A longitudinal study of the effects of ecological interface design on fault management performance, *International Journal of Cognitive Ergonomics*, **1,** 1–24.

Howie, D. E. 1996, Shaping expertise through ecological interface design: strategies, metacognition, and individual differences, Report CEL 96-01, Cognitive Engineering Laboratory, University of Toronto, Toronto.

Moray, N. and Rotenberg, I. 1989, Fault management in process control: Eye movements and action, *Ergonomics*, **32,** 1319–1342.

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

Morris, N. M., Rouse, W. B. and Fath, J. L. 1985, PLANT: An experimental task for the study of human problem solving in process control, *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-15**, 792–798.

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

Rasmussen, J. and Jensen, A. 1974, Mental procedures in real-life tasks: a case study of electronic trouble-shooting, *Ergonomics*, **17,** 203–207.

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

Shaw, R. E., Kadar, E., Sim, M. and Repperger, D. W. 1992, The intentional spring: a strategy for modeling systems that learn to perform intentional acts, *Journal of Motor Behaviour*, **24,** 3–28.

Sheridan, T. B. and Hennessy, R. T. 1984, *Research and Modeling of Supervisory Control Behaviour* (Washington, DC: National Academy Press).

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

Weiner, J. 1994, *The Beak of the Finch: A Story of Evolution in Our Time* (New York: Vintage Books).