Int. J. Man–Machine Studies (1992) **36**, 861–895

Diagrammatic displays for engineered systems: effects on human performance in interacting with malfunctioning systems

DAVID KIERAS

University of Michigan, TIDAL Building, 2360 Bonisteel Boulevard, Ann Arbor, MI 48109-2108, USA

(Received 26 June 1990 and accepted in revised form 21 May 1991)

Computer graphics displays make it possible to display both the topological structure of a system in the form of a schematic diagram and information about its current state using color-coding and animation. Such displays should be especially valuable as user interfaces for decision support systems and expert systems for managing complex systems. This report describes three experiments on the cognitive aspects of such displays. Two experiments involved both fault diagnosis and system operation using a very simple artificial system; one involved a complex real system in a fault diagnosis task. The major factors of interest concerned the topological content of the display—principally, the extent to which the system structural relationships were visually explicit, and the availability and visual presentation of state information. Displays containing a topologically complete diagram presenting task-relevant state information at the corresponding point on the diagram appear to be superior to displays that violate these principles. A short set of guidelines for the design of such displays is listed.

Introduction

DIAGRAMS OF ENGINEERED SYSTEMS

This research is concerned with displays of diagrams of engineered systems. By engineered system we mean systems such as electronic, hydraulic, mechanical or other such systems that are made up of a set of conventional or standard components that are interconnected in a configuration specific to the system. Examples are typical electronic circuitry and cooling and hydraulic systems. In such systems, that pattern of interconnections, the system *topology*, is the critical aspect of the system design. That is, since the components are conventional, rather than unique, how they are connected constitutes the distinctive character of the system; the behavior of the system depends on the flow and control relationships implied by the topology.

The type of diagram that is of concern in this report represents the system topology with various symbols for the components, and lines that show the connections. Such diagrams are normally *schematic*, showing the logical or functional connections, rather than the actual physical or spatial relations that are sometimes shown in *pictorial* diagrams. Often such diagrams contain the conventional primitive components for a domain, such as resistors and transistors in the electronics domain, and pumps and valves in the mechanical domain, and these standard components are shown as conventional symbols. However, *block* diagrams are also common; subsystems that have no standard conventional symbols are shown

0020-7373/92/060861 + 35$03.00/0

as boxes with alphanumeric labels. The user must know the *rules of interpretation* for the diagram, which are the mappings between the visual features of the diagram and the structure and state of the system. Typically, there are fairly strong conventions for diagrams within a technical domain. For example, electronics diagrams follow a very standardized set of rules for the symbols and how they can be combined. Interpreting diagrams is clearly an acquired skill, specific to a technical domain. Such diagrams are a standard representation of devices in the technological world; however, despite their importance and ubiquity, there has been very little research on how people understand and use such diagrams.

COMPUTER-GENERATED DIAGRAM DISPLAYS

Currently, there is great interest in displaying system diagrams on computer graphics displays, to take advantage of both the information storage and retrieval advantages of computers compared to paper-based documentation, and also the possibilities for animation and color-coding to show the current states or behavior of the system. The major potential applications for such displays are in supervisory control tasks, in which a human operator oversees a process that is normally automated, or in other control tasks in which the operator monitors and controls a system remotely. Examples are chemical process control, steam propulsion systems and electrical power distribution systems.

The human factors literature on supervisory control tasks is quite extensive (see Sheridan, 1987; Woods, O'Brien & Hanes, 1987; Woods & Roth, 1988, for reviews) and often emphasizes the key role of system displays and how they must be relevant to the operator's task (e.g. Woods, 1984). However, essentially all of the empirical literature on displays in this context concerns displays of quantitative information, especially in the advanced form of integral displays (see Woods, O'Brien & Hanes, 1987, Goodstein, 1982). Displays of system structure, i.e. diagrammatic displays, have received very little attention, and when discussed (e.g. Goodstein, 1981; Wise, 1986; Rasmussen & Goodstein, 1988), the emphasis is on the hierarchical arrangement or the level of abstraction of the displays. Apparently it has been assumed that diagram displays are, in fact, valuable to the operator. In other words, there is ample reason from existing literature to believe that dynamic diagram displays would be very useful in tasks involving control and diagnosis of complex systems. But as yet we apparently do not have an empirical argument that this is indeed true, and we also do not have any direct information on the possible cognitive factors in using such diagram displays.

MENTAL MODELS AND DIAGRAMS

Much of the supervisory control literature makes the assumption that mental models of the system are important and are conventionally conveyed by diagrams. Diagram displays are a key part of many intelligent tutoring systems for engineered systems, stemming from a belief that diagrams help people form such mental models (cf. Larkin & Simon, 1987). A good example is the STEAMER project (Hollan, Hutchins & Weitzman, 1984; Hollan *et al.*, 1987) in which high-resolution animated color graphics displays were used as part of a tutoring system to teach the complex principles and operating procedures involved with steam propulsion plants. Other

examples are Woolf *et al.*, (1986) and Govindaraj (1988). Unfortunately, the empirical literature on the role of diagrams in learning about systems is also very limited.

It should be clear that there is a pressing need for results and theory about diagrams and diagram displays, and how they support mental model reasoning. To this end, this work relies on a well-developed empirical paradigm for exploring mental model effects (Kieras & Bovair, 1984) and its theoretical implications (Kieras, 1988*a*). The basic strategy is to focus directly on the role of diagram displays in a simple task setting known to be sensitive to the level of mental model understanding.

This work is based on the claim that *diagrams convey mental models.* As discussed by Kieras (1988*a*) a mental model contains:

(i) *Knowledge of the system structure*: the components and their interconnections (the system *topology*).
(ii) *Knowledge of the principles* that govern the behavior of the system.
(iii) *Strategic knowledge* about how to perform tasks using the structure and principles information

Several experiments (see Kieras, 1988*a*, for a review) have shown that understanding how a system works can improve both the learning of procedures when they are explicitly taught, and also the inferring of procedures for operating the device without explicit instruction. The explanation advanced by Kieras and Bovair (1984) is that the mental model enables the person to *infer* the procedures, thus resulting in an improved ability to reconstruct explicitly taught procedures when they have been forgotten, as well as an ability to construct procedures that were not taught. In the domain of the simple devices used in the first two experiments in this paper, it is possible to construct cognitive simulation models for inferring how to operate the device given the mental model knowledge (see Kieras, 1990). Such models represent explicitly how the system state and operating procedures can be determined by inferences based on knowledge of the system structure and principles.

Clearly, making use of a mental model can require a considerable amount of memory retrieval, reasoning and inference. Diagrams can relieve specific aspects of this processing. That is, a good diagram will show the structure of the system explicitly and in a visually clear fashion; there will be no need for the person to store and retrieve facts about the system structure; the diagram can be used as a sort of external memory to rapidly access such information. However, it seems that conventional diagrams provide no other processing relief because the person must still apply the causal principles to infer the system behavior or state.

But if a computer-generated diagram can change to present such state information directly, it could relieve this part of the person's processing as well. Such *dynamic* diagrams in the form of animated sequences are often used in training films, and even cartoon-like diagrams appear in many basic electronic texts to show the sequence of events involved. Finally, large diagrams equipped with indicator lights or gauges are often used in electric power control facilities and railroad switchyards. Of course, computer-generated displays could be much more powerful. But for any new method of displaying diagrams, the user must learn new rules of interpretation and this overhead must be taken into account.

OVERVIEW OF THE PAPER

This report presents some of the first systematic empirical research on the cognitive effects and processes involved in such diagrams and displays. The presentation here is abbreviated; see Kieras (1988*b*) for a fully detailed report. The basic results support the intuition that such displays can be beneficial; however, it is possible for displays that would seem to be better to produce poorer performance. The research reported here is only a beginning; considerably more work needs to be done on this topic.

This paper presents three experiments that show how displays that are more effective in conveying a mental model produce more efficient performance in identifying system malfunctions. The first two involve a simple device and dynamic displays; the third, a complex device and a static display. The paper concludes with a section of guideline advice based on the results.

Experiment 1

RATIONALE

Purpose

The first experiment had the simple purpose of determining whether diagrams of any sort were useful, and whether a dynamic display is beneficial compared to a static one. Experiment 2 used similar materials, procedures and analyses, so this experiment is presented in detail sufficient for the presentation of both experiments.

Task

This experiment was based on the Kieras and Bovair (1984) studies of the effects of mental models for a simple device using a "Star Trek" cover story. The device is a simple control panel, shown in Figure 1, and is controlled by a computer in such a way that it behaves as if its internal structure were that shown in Figure 2. Subjects learn the structure of the system from a textual description. The subject's task is to route "energy" from the source (SP) to a destination (PB) by setting the controls. The internal components (the boxes shown in Figure 2) might be malfunctioning, so the subject must use the indicator lights (I1 to I4) to infer the site of the malfunction and set the controls to compensate for it if possible, and report what malfunction was present.

The basic principles for this system can be described. Energy flows from the ship's power supply through the switches into the components. If the component is functioning properly, an energy input will cause the component to become energized and it will then output energy. If energy is applied to the phaser bank, and the phaser bank is working properly, then the phaser will "fire", and indicator I4 will flash. The other indicator lights in the system come on if there is energy at the attached point or component in the system.

On some occasions, one of the internal components might be malfunctioning. Notice that the behavior of a malfunctioning component is completely all-or-none in this system. If energy comes into the box but no energy comes out, then it is malfunctioning. The subject must use the indicator lights to infer the malfunction status of the system, and set the controls to compensate for it if possible. After

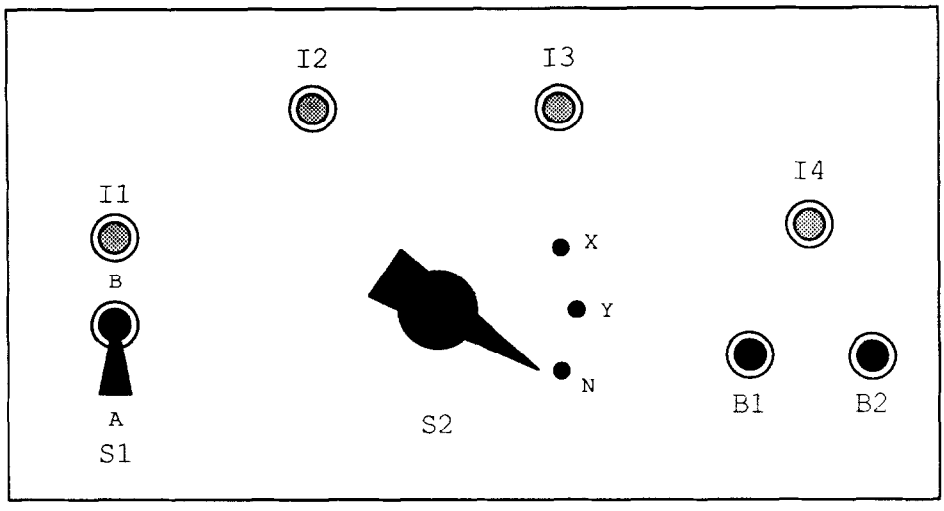

FIGURE 1. Sketch of the control panel device. I1 through I4 are indicator lights; S1 is a toggle switch; B1 and B2 are pushbuttons; and S2 is a rotary selector switch.

attempting to fire the phasers, the subjects report whether they were successful or not, and they also report any malfunctions that were present.

The work reported in Kieras and Bovair (1984) shows that while the "Star Trek" cover story may have some motivating effects, it does not appear to influence in any substantial way the nature of the critical mental model information. That is, the important content of the mental model is not the fantasy of the fictitious physics underlying how phaser banks operate, but rather the information contained in the diagram shown in Figure 2, namely the system topology, and also the principles by which the components behave. An additional result, which will be apparent below, is that despite the extreme simplicity of this system, it presents substantial difficulty for typical subjects.

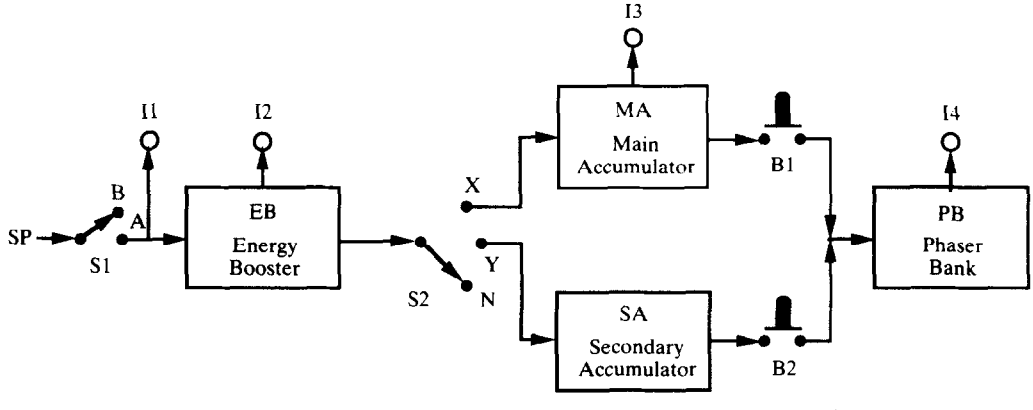

FIGURE 2. The diagram of the fictitious internal structure of the control panel device, showing the components, controls and indicators. This diagram was used in the Static and Dynamic conditions in Experiment 1 and the Topological conditions of Experiment 2.

MANIPULATIONS

The first experiment compared performance with different diagram displays while people performed the task of operating the control panel device in normal and malfunction situations after studying a textually presented mental model for the device. One condition group had *No Diagram* available, and so had to operate the device based only on their memory for the textually presented description of the device. A second group had the same textual description to study, but also had available a *Static Diagram,* essentially the one shown in Figure 2, which was displayed on a computer graphics display throughout the experiment. A third group had a *Dynamic Diagram* displayed on the computer graphics display. This display was identical to the static diagram in Figure 2 when the device was in the initial state, but the controls and indicators changed as the control panel was manipulated; energized connections and components were color-coded, and malfunctioning components had a special color-code. The Dynamic Diagram used the following rules: The normal color of the diagram is blue; if there is energy present along a connection, control or component, it is color-coded as red. A malfunctioning component is indicated by the box being yellow rather than red or blue.

The textual materials studied by all the groups included some general background on the device, such as why phaser systems have energy boosters, a complete verbal description of the device topology and the principles of operation of the device, such as the definition of a component malfunction. Finally, for the groups that had diagram displays, the training material included the rules of interpretation for the display. To be sure that subjects had made a serious effort to acquire this information, they were required to pass a quiz on the material before they could proceed to the problem-solving portion of the experiment.

The subjects solved a series of problems of six different types corresponding to various malfunction states of the system. One of these situations is the *Normal* state; an additional four situations are obtained by a single one of the four components (the boxes shown in Figure 2) being defective. These are referred to by the abbreviation for the malfunctioning component (EB, MA, SA, PB). An additional malfunction state is obtained by a double fault in both accumulators (referred to as the BA situation). The other possible fault states were not used because they are not distinguishable in the behavior of the device.

METHOD

Materials

The training materials were based on those used in Kieras and Bovair (1984). The first section of the material presented the names and functions of each component in the system, along with stating that the components could malfunction, but the connections, controls and indicators could not. The second section dealt with the connections and controls. The No Diagram group was given all of the connections in textual form, with care being taken to insure that this was clear and intelligible. The Diagram conditions had the diagram present throughout the experiment, and the materials stated that the lines with arrows in the diagram indicated the connections between components, and reminders appeared throughout the materials to look at

the diagram for the part of the system under discussion. For all groups, rules were provided for how the system behaved in terms of the energy flow. For example, a connection means that energy can flow from one point to the next, whereas controls would allow energy to flow when they establish a pathway between two terminals, and indicators show the presence of energy. The lack of an indicator for the secondary accumulator (SA) was explicitly pointed out. A malfunction was defined in terms of a component receiving energy but neither becoming activated nor putting out energy.

The Dynamic Diagram group received additional training on the rules of interpretation for the display. They were told that the controls would change their positions on the display according to operations performed on the control panel. Also, where energy was present, the corresponding connection, control, or component would be shown in a red color on the diagram. A malfunctioning component would show as a yellow box.

After reading the second set of materials, the subjects had to correctly answer all of the eight questions in a second quiz on the connections, controls, and energy rules. The dynamic diagram group received an extra three questions on the color coding and animation rules.

Design

The three display conditions, No Diagram, Static Diagram and Dynamic Diagram, were a between-subjects factor. The different problem situations made up a within-subjects factor; each subject saw each of the six situations a total of five times each; the situations were grouped together into five blocks, such that each situation appeared once in each block. The order of presentations of the situation was fixed in the first block and was in the order: *Normal,* defective energy booster (*EB*), defective main accumulator (*MA*), defective secondary accumulator (*SA*), defective phaser bank (*PB*), both accumulators defective (*BA*). In the second to the fifth block, the situations appeared in each block in a random order determined for each individual subject. Subjects were assigned at random to the three conditions, subject to the constraint of an equal proportion of males and females in each condition.

Subjects

The subjects were students at the University of Michigan, recruited through campus newspaper advertisements and posters. They were paid five dollars for participation. About 1 h was required to complete the experiment. After a set of four test subjects, 49 subjects were run, of which four were dropped, two for failing to complete in the time available, one due to accidental loss of data, and one who had been inadvertently scheduled for the experiment after having been in a previous experiment on the same control panel device. This yielded a total of 15 subjects per group.

Apparatus

The control panel device was a slope-front box, whose controls and indicators were interfaced to a DEC VAX 730 via a digital I/O port. The color graphic displays were generated on an Apollo DN3000 with a 1024 by 800 color graphics display. The VAX controlled the experiment, presenting the training materials and instructions

on a standard video terminal, and signalling the Apollo to modify its display. The video terminal was positioned to the left of the control panel device, with the Apollo display to the right. The system box and keyboard of the Apollo were placed out of reach of the subject behind a partition. The laboratory computer recorded and time-stamped all of the subject's actions in progressing through the experiment and interacting with the device.

Procedure

The subjects first studied the mental model material on the video terminal, and answered two sets of quiz questions. If they made an error in answering one of the first quiz questions, they had to go back and read all of the material. Likewise, an error on the second quiz required them to reread all of the material, but they were returned only to the second quiz. After completing study of the device, the subjects read the set of instructions on the problems. They were told that their goal was to get the phaser bank to fire in spite of the different malfunctions, and that they were to report their outcome and a malfunction diagnosis. To encourage subjects to use mental model reasoning instead of trial and error, they were told to use the fewest possible actions to fire the phaser. They were also told that it would not help to keep "banging" on the device; that if it did not work the first time in the situation, it would not work again unless the settings were changed. Subjects were told to plan what they were doing because they would be asked to explain it. The subjects then went on to the first block of the six situations. They were given no feedback on whether their response was correct or not.

Each problem started with a check of whether the device was in the initial state shown in Figure 1. Subjects were told to "shut down" the device at the end of each trial to put it into this state. Then appeared a screen with the command "Fire the phaser", along with an *outcome response menu* that had four alternatives, which they used to report the outcome of their attempt. The first alternative was that they had fired the phaser and that there were no malfunctions to report. The second was that they had fired the phaser but there was a malfunction present. After entering this response, the subjects were prompted to enter a description of the malfunction with the maximum length of one line on the video terminal. The third alternative was that the phaser could not be fired due to a malfunction, which they were similarly prompted to describe. The fourth alternative was simply to "give up"; this was supplied to give subjects some option other than random guessing.

RESULTS

For brevity in this report, details of the statistical analysis are omitted; they can be found in Kieras (1988*b*). Since the analysis methodologies were similar across experiments for the different measures, they will be summarized here to save space below. Where the data consisted of contingency tables, they were analysed using either a simple chi-square test or a log-linear model analysis (Bishop, Fienberg & Holland, 1975; Reynolds, 1977). For the remaining measures, the subset of the data from trials where the outcome response and diagnosis was correct was analysed using multiple regression to deal with the unequal sample size. These analyses were done using a stepwise multiple regression with the within-subjects factors and interactions analysed as suggested by Pedhazur (1982). As a check, the data from

both correct and incorrect trials were also subjected to an ordinary analysis of variance. The analyses of the overall data agreed with analyses of the correct trial data, differing only in showing slightly weaker significance levels. To save space, these check analyses will not be reported. For brevity, the effects of block and situation will not be presented unless they are particularly noteworthy. Except where noted, all effects cited are significant at the 0.05 level.

Once the data was examined, a problem in the design of the experiment was revealed. Previous studies with this device involved giving the subject a command about which setting of S2 to use. Under these conditions, the SA malfunction would be revealed by a failure of the system to work on this commanded setting and the subject was expected to then use the other setting of S2. However, in this experiment, subjects were not commanded which S2 setting to try first and so could use either the X or Y setting initially. This had unintended side-effects on the data. Because there is no indicator on the SA component, to the No Diagram and Static Diagram groups, the SA malfunction situation looks just like the Normal situation unless the subject happended to have tried the Y setting of S2 first. This is fairly rare in the data, apparently due to the fact that X setting is associated with the "main" accumulator; subjects have a strong preference for trying the main accumulator first. Thus, the No Diagram and Static Diagram groups had a far higher rate of missing this malfunction, treating the situation as Normal instead. However, the Dynamic Diagram group had the opportunity to determine the state of SA even if they did not intend to use it, because if they were watching the display while S2 was rotated past the Y setting, they could see the SA box on the display light up in yellow. Thus, the nature of the SA situation is sharply different between the different experimental conditions. In this data, this problem was handled by giving subjects in the No Diagram and Static Diagram conditions "credit" for treating SA as Normal, unless they tried Y first. This is the best compromise with these data; some of the results presented below do not include the SA situation in order to make the effects more clear.

Accuracy of outcomes and diagnoses

As mentioned above, a correct response to the normal situation was treated as a correct diagnosis, even though subjects did not enter a diagnosis statement. The diagnosis responses were printed out on slips of paper, and categorized blind to the experimental condition. These categories were then aggregated into quality categories. The response was considered *Correct* if it was a specific diagnosis that matched the actual situation. Notice that a correct response to a Normal situation is considered correct in this sense. The response was considered *Incorrect* if it was a specific diagnosis that did not match the situation; thus, such responses were well stated, but not correct. The diagnosis was classified as *Poor* if it was superficial, vague, or impossible. It was classified as *Other* if it was uninterpretable or the response was not available due to data-recording problems.

Table 1 presents the proportion of responses in each quality category for the conditions. The Normal Correct column gives the proportion correct for just the Normal situation. The Malfunctions Correct column in the table contains the proportion correct for the malfunction situations not including SA. The log-linear analysis showed significant effects on response quality category of situation,

TABLE 1.
Diagnosis quality for each condition

Condition	Correct	Incorrect	Diagnosis quality: Poor	Other	Normal correct	Malfunctions correct
No Diagram	0·58	0·05	0·28	0·08	0·99	0·43
Static Diagram	0·83	0·06	0·08	0·03	0·99	0·77
Dynamic Diagram	0·87	0·01	0·07	0·04	0·95	0·89
Mean	0·76	0·04	0·15	0·05	0·98	0·69

The Malfunctions Correct column contains proportion correct for the malfunction situations EB, MA, PB, BA only; normal and SA not included.

condition, and the situation and condition interaction. As shown in Table 1, most of the diagnoses were correct, but the No Diagram condition produced fewer correct responses and more poor responses. The two Diagram conditions were roughly equivalent.

The effect of situation is summarized in the two right-hand columns of Table 1, showing proportion correct in the normal and all malfunction situations. The accuracy in the normal situation was quite high for all groups, but simple chi-square tests performed just on the Malfunctions Correct aggregation shows that the three conditions are significantly different from each other.

In summary, there was a substantial improvement in the quality of diagnosis responses as the amount of information available to the subject was increased. This effect was mainly on the malfunction situations; the normal situation was relatively easy and so may have shown no effect. A major effect of providing the diagram information was to eliminate poor responses, suggesting that the information improved the subjects' understanding of the system.

Time to complete the task

The time required to complete the task was measured from the appearance of the screen containing the "fire the phasers" command until subjects made their outcome response. Subjects were free to perform the "shutdown" steps either before or after they made their menu response. Table 2 shows the mean completion times for correct trials for all blocks. There was a significant effect of condition; the Dynamic group was fastest, but most of this effect appears to be due to the Static group actually being slowest. A significant condition × situation interaction took the form that the hard malfunctions, EB, PB and BA, were improved by the Dynamic Diagram much more than the other situations, and the Static Diagram group was especially bad in these situation. The SA situation is anomalous due to the special treatment mentioned above. There was a condition × block interaction (not shown in Table 2) in which during Block 1 the Dynamic group is much better in the hard situations, and by Block 5 the difference between conditions is considerably reduced, but it was still present and significant in a separate analysis of Block 5.

In summary, the Dynamic Diagram condition is indeed fastest, with the Static being somewhat slower, but not significantly different from, the No Diagram

TABLE 2.
Mean execution time (s) in each condition and situation

Condition	Normal	EB	MA	SA	PB	BA	Mean
No Diagram	14·9	11·8	12·5	9·0	14·8	14·6	12·9
Static Diagram	12·8	14·6	13·1	11·2	16·8	15·1	13·9
Dynamic Diagram	14·0	8·0	11·0	11·1	11·3	8·6	10·7
Mean	13·9	11·5	12·2	10·4	14·3	12·8	12·5

condition. The Dynamic Diagram group was especially better at the hard problems, and remained so.

Response sequence quality

The sequences of actions produced by subjects were tabulated up to but not including the shutdown or outcome response. These sequences were classified into several categories, which were assigned blind to the experimental conditions associated with each sequence. The categories were then grouped into two quality categories, *good* and *poor*. There were three categories of good response sequences. *Pure optimum sequences* (POS) were sequences that for the situation contained no wasted moves. *Near-miss* (NM) sequences were defined as an optimum sequence followed by a single additional well-defined action or sequence, such as firing the phaser a second time. Finally, *Inspection* sequences were good, but S2 was placed at both settings. There were two general categories of poor responses. *Try-both* sequences represented simply blindly using the other accumulator when the first one failed. *Erroneous* responses reflected an erroneous understanding of the device and how to operate it. The final category was *Other*, consisting of all sequences that did not fit into one of the above categories.

A log-linear analysis yielded significant effects for response type, and effects on response type due to situation, condition and the situation × condition interaction. The Dynamic group produced more of the optimum, near-miss, and inspection sequences, whereas the other groups produced more of the try-both and erroneous sequences. The effects are shown in abbreviated form in Table 3, in which the proportion of response sequences categorized as good are shown for Normal situations, malfunction situations and all situations. The malfunction situations shown in this table did not include the SA situation, which averaged almost as high a proportion of good sequences as the Normal situation (0·88). The proportion of good response sequences was much higher for the Dynamic Diagram condition than the other two, which were about the same, as shown by a chi-square test on the frequencies underlying the *All* column in this table. However, quite a lot of this effect is due to the malfunction situations being very much better in the Dynamic Diagram condition than for the other groups, whereas for the Normal situation, both the No Diagram and Dynamic Diagram conditions produced much better sequences than the Static Diagram.

DISCUSSION

The Dynamic Diagram produced substantial performance improvements compared to the Static Diagram, and very large improvements compared to No Diagram. An

TABLE 3.
Proportion of response sequences categorized as good

Condition	Normal	Malfunctions	All
No Diagram	0·92	0·37	0·54
Static Diagram	0·77	0·44	0·56
Dynamic Diagram	0·97	0·75	0·84
Mean	0·89	0·52	0·65

Malfunctions do not include the SA situation, which averaged 0·88.

interesting result is that, compared to No Diagram, the Static Diagram was a burden in terms of execution time but better in accuracy and quality of responding. Apparently, the Static Diagram information is useful, but it takes time to use it, and it is not as useful as the Dynamic Diagram information. As noted in the results, the execution time effects tend to diminish with practice, but note that extensive practice may be lacking in the real situation in which such displays might be used.

Experiment 2

RATIONALE

Purpose

The first experiment showed that the Dynamic display was beneficial, but notice that the dynamic display contained three kinds of information: the topological information, the energy state information and the malfunction state information. The results do not show which of these confounded factors actually contributed to performance.

Topological information, showing how the components and controls were related to each other, does not seem to be of great value because the Static Diagram group was only a little better (and sometimes worse) than the No Diagram group in performance. However, since providing this topological information would be quite expensive in practical terms, it is important to determine definitely whether it plays any important role.

The state information about where energy was present consisted not only of which components were energized, but also showed which of the connections between components were energized. For example, when the power switch S1 is turned on, not only does the EB box change to red, but also it is obvious that energy is applied to the arm of the selector switch S2, because the interconnecting line and the arm of the switch change to a red color. A possible problem is that the dynamic display shows the state of the secondary accumulator; this may have simplified the task in ways that are not directly related to the use of a dynamic display itself. The No Diagram and Static Diagram groups had to deal with the *hidden state* of the secondary accumulator, which in earlier work done with this device was a major source of difficulty for subjects.

Finally, the dynamic display presented a direct indication of which components were malfunctioning. If energy was applied to a malfunctioning component, that component lit up in a bright yellow color. It could certainly be argued that this directly supplied state information, and not other properties of the dynamic display,

allowed the subjects in the Dynamic Diagram condition to do a more efficient job of operating the device than in the other two conditions.

The second experiment attempted to separate the contribution of the topological information from the contributions of the two kinds of state information. Due to the large number of possible partitionings of the information in the Experiment 1 Dynamic Diagram condition, only a subset of the combinations of the possible factors and their levels were tested.

Manipulations

The same device and task were used as in Experiment 1. There were two display condition factors. The first factor was whether or not the display made the topological connections between the components visually explicit. The *Topological* display was the same as that used in Experiment 1; in the *Non-topological* display the connection information was present, but was not represented as visually clear lines between the components.

The Topological version is simply the same diagram as that used in Experiment 1. The Non-topological version of the display, shown in Figure 3, requires some explanation. First, notice that the components, controls and indicators are neatly grouped into a rectangular array. The indicator lights are ordered left to right simply in numerical order of their labels, and the controls are ordered left to right as they appear on the control panel. Notice that for each one of these items there are arrows shown entering and leaving the item. Each arrow is labelled with letters that show corresponding points that are connected together. Thus, energy can flow out of switch S1 to point *a*, which is the same point as the arrow entering the energy booster. Energy can leave the energy booster through connection *b*, which is the input to selector switch S2, and so forth. Thus, by starting at the Ship's Power point and following the labels through, one can trace out the entire connections for the system.

The second display condition factor was the amount of state information present in the display. At the *Low* level, the presence of energy was shown for the controls, indicators and components. That is, the controls, indicators and components were normally shown as blue; if there was energy present at that point, they were shown as red. Note that at this lowest level of state information, the energy state of each component is shown; this has the effect of rendering the state of the secondary accumulator visible just as if it had its own indicator light; the possible problem in Experiment 1 with the hidden state of the secondary accumulator is eliminated in all of the conditions used in this experiment.

At the *Medium* level of state information, in addition to the color-coding used at the low level, the red color-code was also used to show which connections had energy present. Thus, while at the Low level the energy booster box would be lit up in red when the energy booster was energized, at the Medium level the lines leading into and out of the energy booster box would also be shown in red.

The *High* level of state information included not only the Low and Medium color-code schemes, but also a malfunction color-code in which a malfunctioning component was shown as a yellow box. Thus, the Topological and High information level condition was the same as the Experiment 1 Dynamic Diagram condition.

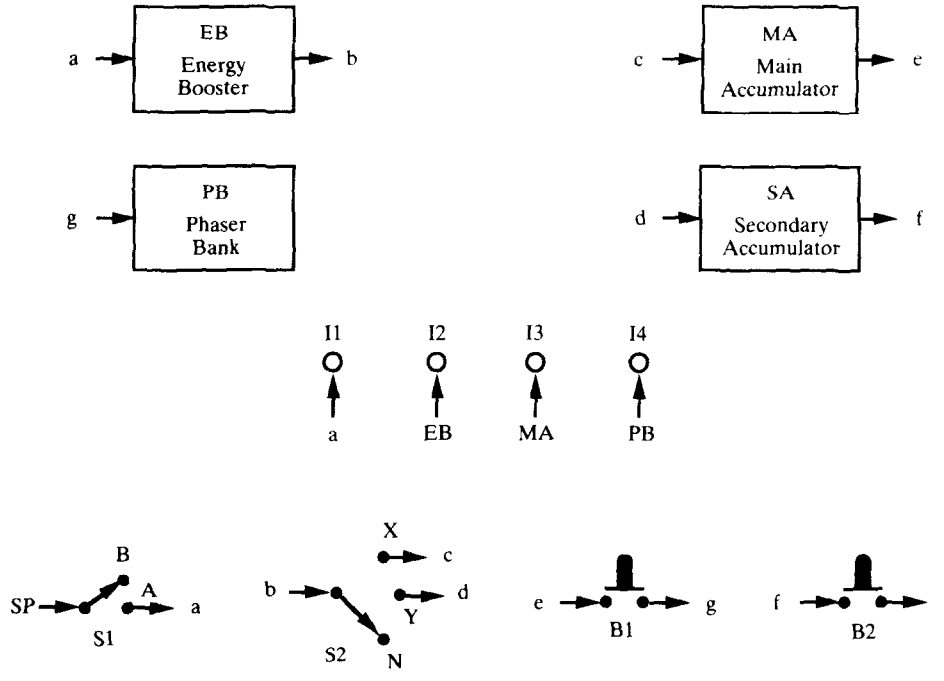

FIGURE 3. The Non-topological diagram of the control panel device used in Experiment 2. Notice how the labelled terminals specify all of the connections shown by lines in the Topological version shown in Figure 1.

Expected results

The expected results were that the Topological displays should be better than the Non-topological displays, because there would be less processing required to determine how the components are connected. There should be an improvement in going from Low, to Medium, to High state information levels because successively less inference is required to deduce the component states. In more detail, the logic behind the choice of these conditions can be seen by considering what information processing would be relieved by the displays in each condition.

With regard to the static information in the display, both the Topological and Non-topological displays relieve the subject of having memorize and retrieve the names of the components and the connections between them. But notice that in the Non-topological display, determining the connections requires searching the display for matching labels, while in the Topological display, the connections can be determined by visually following the lines from one component to the next. Thus, although both displays present the same static information, the Topological display makes it much more visually available.

With regard to the dynamic information, the Low information level display alleviates the need to import the states of the controls and indicators from the actual physical control plane into the diagram. That is, if the subject uses a static diagram as a kind of external memory, one of the required cognitive processes is keeping in short-term memory the positions of the controls and states of the indicator lights

while looking at the diagram. One can imagine this information as being mentally "pasted" on to the diagram, and updated as required. The Low information display makes this importation into the diagram unnecessary, because the display shows this information directly by means of animation of the control symbols and color-coding of the indicator light symbols.

In addition, whether a component is energized is also directly shown on the Low information display. For example, if S2 is on the Y setting, and the energy booster indicator I2 is on, the SA box will have the red color-code for energy; the subject does not have to infer that the secondary accumulator is energized. If the SA does not become energized, then the SA box will still be blue, and the subject can deduce that the SA is malfunctioning from the blue color together with the fact that the positions of the controls establish a pathway between a red component and the blue component. Thus, the inference that a component is operating normally is relieved by the Low information display, and a portion of the reasoning that a component is malfunctioning is also relieved.

The Medium information level display relieves the processing necessary to trace where energy is present along the connections in the system. Pursuing the above example, in the Medium information display the energized pathway between the energy booster and the secondary accumulator would be shown in red. Thus it is unnecessary to deduce that because the energy booster is red, and that there is a blue pathway connecting the energy booster to the secondary accumulator, that there should be energy present at the secondary accumulator. Rather, in the Medium information display, there will be a solid red pathway going from the energy booster to the secondary accumulator. Thus, the inference that the secondary accumulator is malfunctioning should be easier in the Medium than in the Low display, because a simpler rule could be used: If there is a red arrow coming into a box, but the box is blue, then the box is malfunctioning.

Finally, the High information display goes one step further; a malfunctioning component is lit up in yellow. Thus the rule for inferring a malfunctioning component at the High level becomes even simpler: if a box is yellow, then it is malfunctioning. Notice that since the identification of malfunctioning components is an important part of this task, having these directly signalled should be very useful to the subject; perhaps the other levels of information and the system topology is of very little value.

Thus, the pattern of expected results is quite straightforward. The Topological displays should be better than the Non-topological displays, because the processing required to determine how the components are connected would involve less difficulties. There should be an improvement in going from Low, to Medium and to High state information levels because successively less inference is required to deduce the component states. However, as will be seen in the results, this simple pattern did not appear.

METHOD

Materials

The materials were substantially like the Experiment 1 dynamic diagram condition, except that the "rules of interpretation" section of the training was modified to suit

the different displays. The section on the functions of components and the connections had nine quiz questions; the section on principles had a five-question quiz and the rules for interpretation of the display contained a set of five questions answered by all groups, with an additional two for the Medium state information level, and an additional three questions for the High level. The section on rules of interpretation explained the labelled arrows in the Non-topological display with some care, both by providing the general rule for how the labels worked and also with some examples that "walked" the subject through portions of the diagram. The color-code conventions were explained the same way as in Experiment 1.

Design

The between-subjects factors were the Topological *vs* Non-topological factor and the three levels of state information, giving a 2×3 design with six groups. Each subject saw each of the six situations grouped into five blocks, for a total of 30 problems. As in Experiment 1, the first block appeared in a fixed order, which was the same as in Experiment 1; the second through fifth blocks appeared in an order determined at random for each subject. Subjects were assigned at random to the six groups under the constraint that there was an equal proportion of males and females across conditions.

Subjects

The subjects were recruited and paid in the same way as Experiment 1; this experiment also took about 1 h. A total of 105 subjects were run to produce 15 subjects in each group. The first three were test subjects; four subjects were lost due to equipment malfunctions; four subjects were dropped for failure to complete the experiment in the time available; two subjects were dropped for poor performance in that they never reported a malfunction; and two surplus subjects were dropped.

Apparatus and procedure

The equipment was the same as in Experiment 1, but due to the greater complexity of the display program, the display updating took close to 1 s to complete, so there was a noticeable "ripple" of the changes being made on the display. The basic procedure was the same as Experiment 1, but an additional quiz was used, so that the subject answered a quiz after reading each of the three sections of the material. If they missed one of the questions, they reread only that section.

RESULTS

It should be kept in mind that all of the conditions in this experiment should allow subjects to perform better than the No Diagram or Static Diagram conditions in Experiment 1, so the effects in this experiment should not be as dramatic as in the first. The measures of performance and the presentation parallel those of Experiment 1.

Quality of diagnoses

Diagnosis responses were scored and tabulated as in Experiment 1, where a correct outcome response to a normal situation was counted as a correct diagnosis. Table 4 shows the proportion of correct diagnosis responses in each condition. The Topological display conditions were better overall than the Non-topological, but by only seven percentage points. However, this 11% improvement is highly significant.

TABLE 4.
Proportion of correct diagnoses in each condition

Information level	Diagram type		Mean
	Non-topological	Topological	
Low	0·69	0·72	0·70
Medium	0·60	0·60	0·60
High	0·66	0·84	0·75
Mean	0·65	0·72	0·69

The effect of information level was rather different from what was expected. The High information condition produced the highest proportion of correct responses, 0·75, but the Medium condition produced the worst, only 0·60, while the Low information level was 0·70. This main effect was also highly significant. This pattern in which the Medium level is worse than the other two appears with all of the performance measures.

The interaction of topology condition and information level was also significant; the Topological High level condition is much better in producing correct diagnoses than the others; this condition gets 0·84 of the responses correct compared to the grand average of only 0·69. It is interesting that the Non-topological High and Low conditions are similar; providing more information in the Non-topological condition did not help.

A log-linear analysis showed a strong effect of situation, with the Normal situation producing an average of 0·98 correct, EB, PB and BA averaging about 0·69, but with the MA and SA situation averaging only about 0·53. The MA and SA situations have a special role in these data, because they are the only malfunction situations in which it is possible that the subject might miss that there is a malfunction present, but still get the system to operate apparently properly. This will be discussed in more detail below. The interaction of topology, information level and situation was also significant. The Topological High level condition was very much better than the other conditions on the malfunction situations, especially EB, PB and BA. The Medium level was almost always worse than the Low level of information across the situations. It is noteworthy that the Medium level was especially bad on the SA problem, only 0·34 correct, which will be discussed more below.

In summary, the High information level and Topological factors both led to an increase in accuracy of diagnosis, but the improvement from these two factors appeared mostly only when they were both present—the Topological High Level display was clearly superior to all other conditions. Unexpectedly, the Medium level was the worst.

Execution time

The mean execution times for correct trials in each condition are shown in Table 5. The regression analysis showed that the Topological condition was faster by about 1·5 s than the Non-topological. Information level was significant, again with the Medium level being the worst. The topology × information level interaction was also significant; the Topological High level combination is fastest, while the

TABLE 5.
Mean execution times for correct trials in each condition

Information level	Non-topological	Topological	Mean
Low	13·0	14·0	13·5
Medium	17·9	13·9	15·9
High	13·8	12·4	13·1
Mean	14·9	13·4	14·2

Non-topological Low level is very slow; the remaining conditions are about the same. The information level × situation interaction was also significant (not shown in Table 5). In some situations the Medium level of information was especially slow, such as the PB and BA situations. Other situations were especially fast for the High information level, such as SA and BA. In most situations, High and Low level had similar execution times, suggesting that these conditions were actually rather similar in processing effort required.

In summary, the High information level and the Topological display produced faster times to complete the task, but again, these factors interacted heavily; the benefits of the Topological display appeared primarily at the High level of state information. Again, the Medium level of state information was by far the worst, especially with the Non-topological display.

Response sequence quality

As in Experiment 1, each sequence of actions performed by the subject was classified as a pure optimum sequence, a near miss to an optimum sequence, a try-both sequence, erroneous operation, inspection or a new category of *miss*. A miss was defined as executing a response sequence that would lead to missing the presence of a malfunction that was present. For example, in the MA situation, if the subject got the phasers to fire by using the secondary accumulator, and used an optimal sequence of operations to do so, then the subject would miss the fact that the MA malfunction was present. In Experiment 1, in the SA and MA situations, such misses were allowed as pure optimum strategies or near misses, but here they are separated because it appears that malfunction misses are important to understanding why the Medium level produced inferior performance to the Low.

Table 6 shows the results of aggregating these sequence categories into *Good* and *Poor*, showing the proportion of patterns classified as good (pure optimum, near miss or inspection sequence) in each condition. These data were subjected to a log-linear analysis. There were slightly more good sequences (0·56) in the Topological condition compared to the Non-topological (0·50). Again, the Medium level is the worst information level with only 0·48 good response sequences, whereas the High and Low level were similar with 0·57 and 0·54 proportions respectively. The topology × information level interaction was significant. As shown in Table 6, the Topological display improves the Low and High information levels substantially, but the Medium level stays about the same, or is perhaps even worse. The highest level of performance is in the Topological High level condition. The interaction of topology, information level and situation was significant; the Medium level is very

TABLE 6.
Proportion of good response sequences in each condition

Information level	Diagram type		Mean
	Non-topological	Topological	
Low	0·50	0·59	0·54
Medium	0·50	0·46	0·48
High	0·51	0·63	0·57
Mean	0·50	0·56	0·53

poor on malfunctions, especially in the Non-topological case. Not shown in Table 6 is that there was a very low proportion of good sequences in most of the malfunction situations, especially the SA situation, which averaged only 0·35.

Some insight into why the Medium level was worse obtained by analysing those situations in which malfunctions were missed. As mentioned above, a missed malfunction error was defined as a response that the phaser system was normal when it was not in the Normal situation. There was a higher rate of missed malfunction errors in the Medium information level. This suggests strongly that the Medium subjects had less understanding about the state of the device.

To investigate this further, the missed malfunction reponses were tabulated and analysed in detail. The Medium level conditions had more misses than the others, and there were more misses in the SA situation than in the MA situation. As noted above, these were the only situations where subjects could miss the presence of a malfunction and still get the phaser to fire. The others are "fatal" malfunctions in that the phaser cannot be made to fire no matter what the control settings, and so the subject will sooner or later come to the conclusion that there is a malfunction. However, in the MA and SA situations, the subject could miss the presence of a malfunction in two ways. One is to use the functioning accumulator without trying the defective one. For example, in the MA situation, setting S2 to Y and pressing B2 will fire the phaser; since S2 is never set to X, the state of the main accumulator would never be shown on either the control panel or the display. The second way to miss the malfunction is a *pure miss*—the subject had set S2 to the malfunctioning accumulator, but did not notice on the display that something was wrong with it. Notice that the secondary accumulator setting (the Y setting of S2) is "on the way" to using the main accumulator, so subjects have a natural opportunity to interrogate the state of the secondary accumulator. But, the secondary accumulator does not have an indicator light, so the only state information available is the display. Thus the SA situation is the malfunction that is most sensitive to the quality of the display.

The pure misses were tabulated and analysed. The frequency of a pure miss of the MA malfunction was only about 5% of the responses, and occurred slightly more often in the Topological display, but was not significantly affected by information level. But the SA situation was more informative. Pure misses of the SA malfunction made up about 6% of the responses, but occurred significantly more often in the Non-topological displays and in the Medium level, and were especially common in the Non-topological Medium condition. Table 7 shows the frequency of

TABLE 7.
Proportion of missed SA malfunction errors in each condition

Information level	Non-topological	Topological	Mean
Low	0·43	0·20	0·31
Medium	0·60	0·51	0·55
High	0·39	0·23	0·31
Mean	0·47	0·31	0·39

Total $N = 450$.

misses in the different conditions in the SA situation only. A log-linear analysis showed that there were many more misses in the Medium level of information, in both the Topological and Non-topological conditions. The Low and High level of information benefited from the Topological display, cutting the level of misses in half, but there is only a slight benefit for the Medium level. However, this apparent interaction was not significant.

These results are a strong suggestion that the Medium subjects had difficulty seeing state information that was more apparent to subjects in other conditions. The SA situation should be detectable, because subjects have to switch past that setting of S2. It is possible that they could do this faster than the display could update, and so they might miss some of the SA information. But there does not seem to be any reason why the tendency to switch rapidly through the S2-Y setting would differ so radically with experimental conditions, so this explanation is unlikely.

In summary, the response sequence quality conformed to the other measures in that the Topological display and the High information level yielded a better quality of responding, especially when combined. The Medium level again yielded much poorer response sequences, especially in situations where it was possible for subjects to miss the presence of a malfunction.

DISCUSSION

Summary

These results show that the Topological display was superior to the Non-topological display, and this superiority held even at the High level of state information. Thus the superiority of the Dynamic Diagram display in Experiment 1 cannot be attributed solely to the fact the explicit state and malfunction information supplied subjects with everything that they needed to carry out the task efficiently. Rather, the visually explicit connections in the Topological display seem to be useful even when the state information is highly informative. In fact, the best condition is the Topological High level; the topology information, in fact, makes the biggest difference at the High information level. The High level of state information was best, but an unexpected result was that the Medium level was the worst and the worst condition was the Non-topological Medium level.

Why is the medium level worst?

It is important to try to explain the poor performance in the medium level even if only a post-hoc explanation can be given at this time. A possible explanation is that

the different information levels differ in the availability of gross, easily discriminable visual cues. Perhaps subjects could use these gross cues in a simple and efficient manner to determine whether their control settings were leading to a solution. In the Low information level, manipulating a control associated with a *functioning* component would cause some *additional red features* to appear on the screen. If the component was *malfunctioning*, there was *no* color change in the display. For example, if S2 was set to Y, and there was a malfunction, there would be no gross change in the display; the SA box would remain blue. If there was no malfunction, there would be a gross change; some additional red would appear as the SA box turned red. Thus the rule for determining whether a component was malfunctioning in the Low information condition could be written simply as, *if red appears anywhere when a control is manipulated, the associated component is functioning correctly*; *otherwise the component is malfunctioning.* Thus subjects could tell if they were on the right track simply by watching for additional red to appear on the screen.

In the High information level, yellow is used to indicate a component that is defective, and so there is also a gross change in the display that indicates a malfunction state. Thus the rule for processing this display could be stated as, *if yellow appears anywhere when a control is manipulated, the associated component is bad*; *otherwise the component is good.* Notice that the red color-code can be simply ignored, because the appearance of yellow is adequate information for the subject to decide which control settings are usable.

However, in the Medium level of information, the red color-code is ambiguous; it shows both where energy is present in the system, and also whether a component is operating properly. Thus a rule based simply on a gross change of the amount of red on the display will not be adequate to determine whether a component is functioning or not, as in the Low level display. On the other hand, the appearance of red on the display cannot simply be ignored as in the High level display, because the red color for a box is the only information concerning the malfunction state of a box. Thus, the rule for processing the Medium level display is more complicated, *if red appears on the display when a control is manipulated, then examine the associated component to see if it is red. If it is, then the component is good*; *otherwise the component is bad.*

Thus, the Medium level display is actually less effective in its use of color-coding than the Low level, because the ambiguous coding requires more processing. Apparently, the extra effort required is enough to cause this supposedly more informative display to be less usable. This effect may have been aggravated in the Non-topological display condition because the red-colored items are scattered about the display and harder to find. Now that it is clear that more informative state displays are not necessarily more usable displays, future research could clarify these effects in more detail.

Experiment 3

RATIONALE

Purpose

Experiments 1 and 2 suggest some important principles, but in the context of a very simple artificial system using a relatively elaborate "full feature" display. The

research problem at this time is to get an early indication of whether these principles would appear at the other extremes of the task domain, using a more realistic combination of system and display. Thus, this experiment used a complex system that was also a *real* system, and which required a heavy use of background knowledge. The task required only malfunction diagnosis, rather than procedure inference, and also did not involve interacting with the device and the display in real time. The animation in the display was very limited, consisting just of changes in gauge readings, which, once set, remained static through the course of a problem, and the display was monochrome, making no use of color-coding.

The basic structure of the experiment was to compare a "good" and a "bad" version of a diagram display for the system; the good display had a relatively complete topological representation and state information was associated with the topology; the gauges and indicators were distributed about the diagram. The bad version had an incomplete device topology, and the gauges and indicators showing the state information were visually dissociated from the diagram, being neatly grouped at the side. Thus, as in Experiment 1, different aspects of the display were confounded with each other as a simple way to see if performance with a complex system could be facilitated by a "better" display.

Manipulations

The choice of the system and display manipulations was clearly critical. The real system must be both complex and also learnable in a reasonable length of time. For practical relevance, the bad version of the display must be *realistically* bad, representing the quality of display that would be developed in the absence of research such as this. Both of these requirements were met by using work previously done by NASA that led to these experiments. Malin and Lance (1987) developed an expert system that was able to diagnose faults and take corrective actions on a space-vehicle life-support subsystem, the CS-1, that removes carbon dioxide from cabin air using an electrochemical fuel cell process. The CS-1 system is fairly complex, involving electrical, gas and fluid systems. The expert system, called FIXER, used a dynamic diagram display in its user interface, based on a diagram similar to the engineering diagram from the technical report for the CS-1 system (Heppner, Dahlhausen & Schubert, 1983). The FIXER display is a realistic display in that it represents a reasonable product of an engineering effort that did not include systematic human factors testing. Likewise, the engineering diagram in the technical report was prepared according to customary engineering practice to document the system, not to support human troubleshooting. Thus, the engineering diagram and the FIXER display served as the basis for the realistically "bad" display used in this experiment.

Figures 4 and 5 show the "bad" and "good" versions of the displays used in the experiment, with the gauges shown for typical malfunctions used in the experiment. The bad version in Figure 4 was based on the FIXER display, with certain gauges added and others dropped to better suit the malfunction diagnosis task used in the experiment. But the bad version preserves the spirit of the FIXER display, in that there are large, clear and neatly arranged gauges that are separate from the diagram, and the diagram is topologically incomplete; notice, for example, how there is no information on the internal topology of the EDCM component.

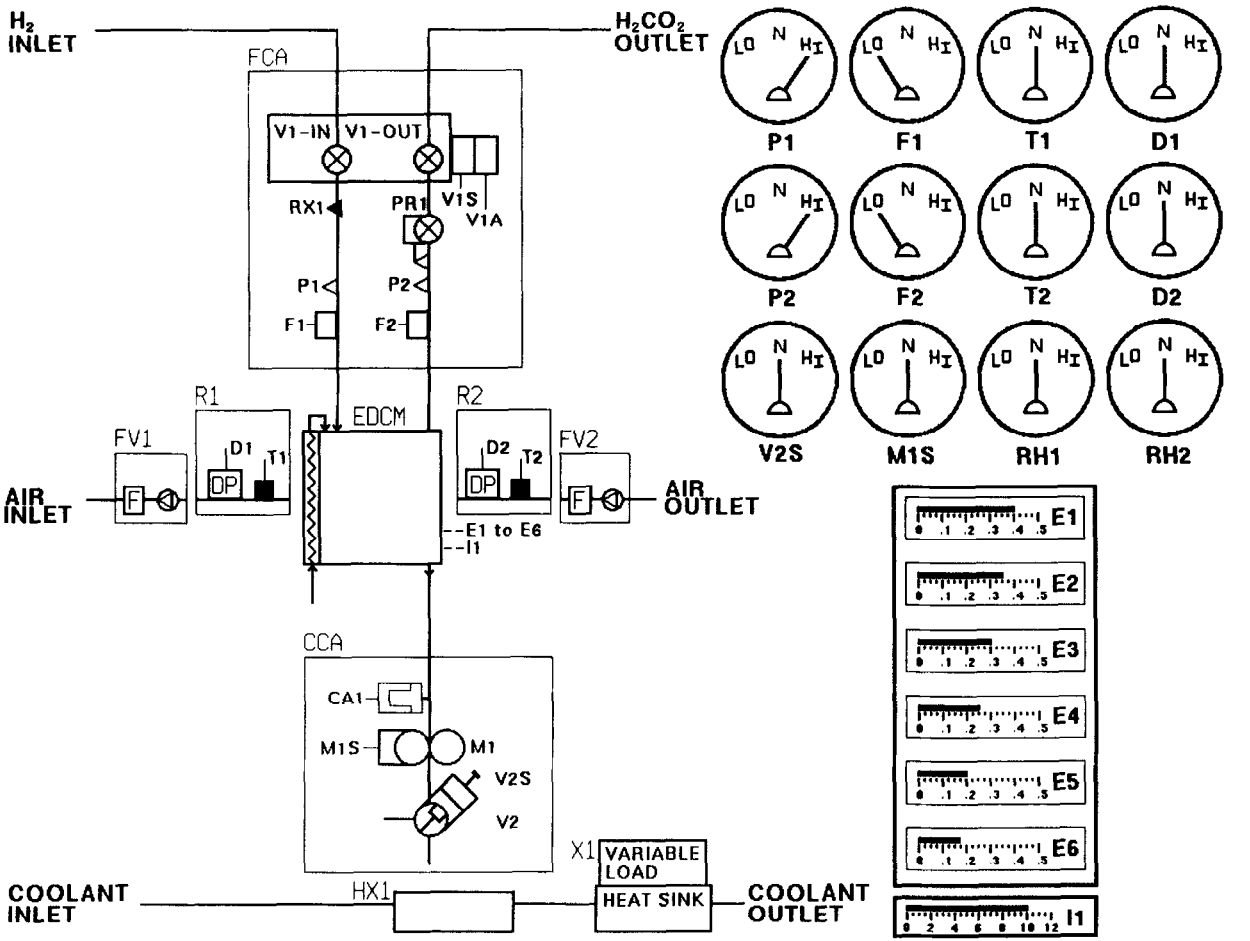

FIGURE 4. The bad version display used in Experiment 3, based on the FIXER display. Notice the lack of internal structure of the major EDCM component and the many missing connections between components, sensors and gauges. The gauge readings indicate a malfunction; the hydrogen flow restrictor, RX1, is obstructed.

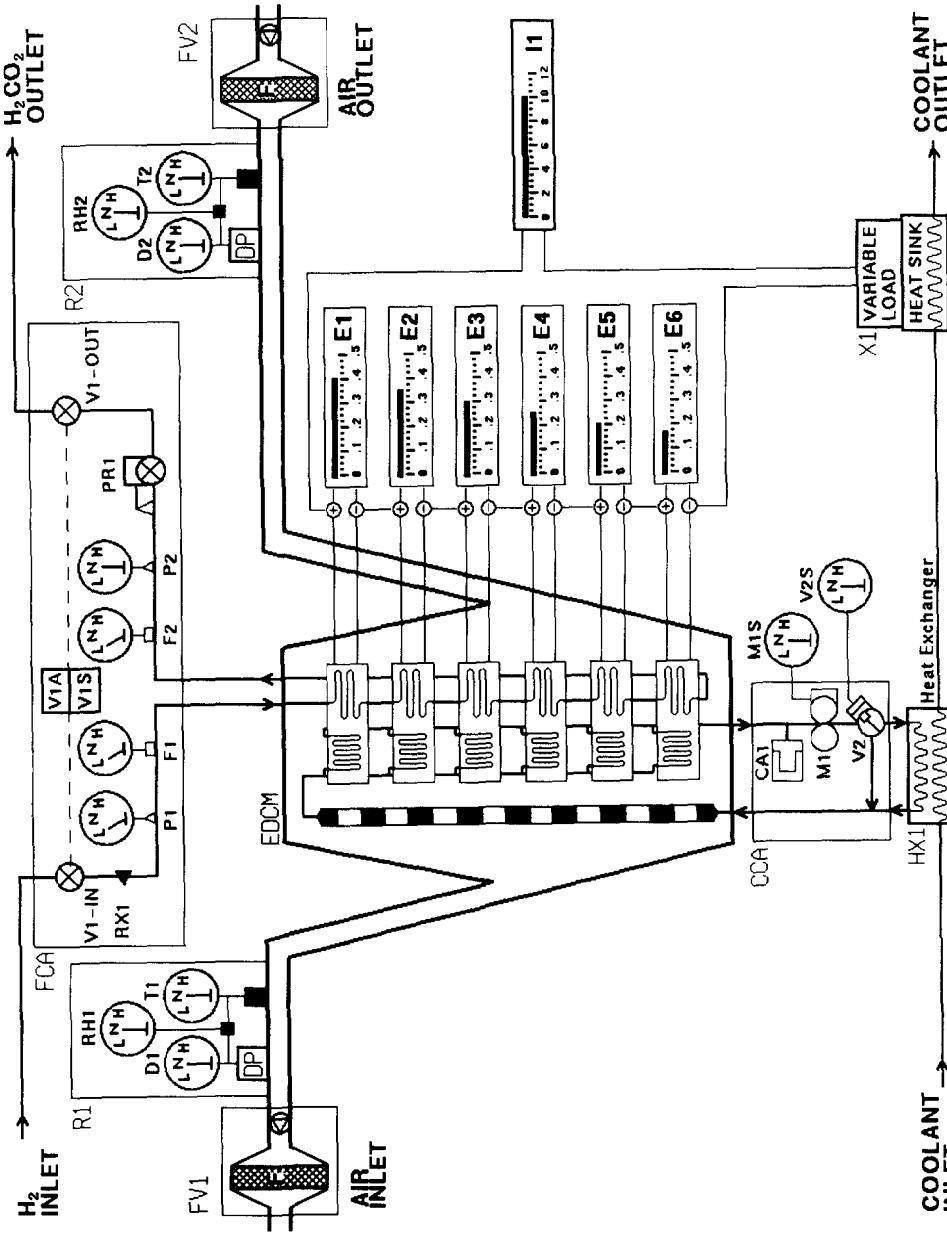

FIGURE 5. The good version display in Experiment 3. The gauges have been placed on the diagram at the sensor locations. The diagram is more topologically complete, showing the internal structure of the EDCM module and the connections between components and sensors. The malfunction indicated by the gauges is an obstructed hydrogen flow restrictor, RX1.

The good version of the display in Figure 5 is very different from the bad one. It is visually considerably more complex than the bad version, but the claim is that it should actually be easier to use because of the more complete and more useful representation. The good version diagram is topologically more complete in three ways: first, the connections between the components are all present; second, the relevant internal structure of the EDCM component is shown with emphasis on the series and parallel relationships of the air, cooling and hydrogen circuits; third, the good version shows the electrical topology, making explicit the series electrical connections between the cells, the individual voltage reading connections, and also the connection between the cells and the load X1. The good version also associates state information with the system structure. The gauges in the good version are distributed on the topological diagram representation. Each is positioned near and visually connected to, the sensor or the device that the gauge reports on. The size and arrangement of the gauges were sacrificed to make this distribution possible. Thus, the gauges are considerably smaller and less regular in arrangement than in the bad version of the display.

In both the good and bad versions, the only portion of the display that was animated were the gauge readings; each malfunction situation corresponded to a different set of gauge readings. However, during the time the subject was solving a particular problem, the display was completely static; the gauge readings did not change in real time.

Thus, the manipulation in this experiment was to augment the bad version display to make it more topologically complete, and re-arrange the gauges and indicators to associate them visually with the diagram. The good version is visually more complex than the Bad version, but is topologically more complete and the gauges are more informative and corrects state to structure better. The prediction was that better performance should result.

Task

The experimental task was similar to that in Experiments 1 and 2. The subjects first learned about the system from an on-line manual, and then solved problems that consisted of diagnosing malfunctions using a display of the system. Unlike Experiments 1 and 2, the subjects did not have to pass a quiz on the contents of the manual; rather they could go back and read the manual at any time during the experiment. The major difference from the previous experiments was that the display was static, and the subjects did not manipulate the system in any way.

METHOD

Training materials

In general, the training material was intended to give subjects a complete mental model of the CS-1 system. Since this was a complex and real system, the subjects were selected to have a background in the appropriate branches of science and engineering so that they knew the relevant concepts from physics, chemistry and technology. Included in the training materials was a set of pictures and diagrams from the engineering manual; it was felt that it was important to ensure that subjects understood the CS-1 as a concrete, existing system. The materials also included the

rules for the interpretation of the display, and the rules and principles governing the possible malfunctions and the diagnosis task that the subjects were to perform.

The materials were presented in the form of an on-line manual presented on a standard video terminal. The overall length of the manual was approximately 40 standard video terminal screens of information. The subject also had available a loose leaf binder with 10 pictures and diagrams from the CS-1 technical report. The pictures were photographs of the overall appearance of the CS-1 system and each of its major sub-components. The diagrams showed the physical shape and arrangements of some components, such as the fuel-cell structure, and some of the overall system relationships. One was a schematic diagram showing the chemical reactions within a single cell. However, it is important to note that subjects did not see the engineering diagram although it was part of the CS-1 technical report, nor did they see any diagram that showed the system component and topology information conveyed by Figures 4 or 5.

The on-line manual explained several principles of the CS-1 system, such as how the fuel-cell reaction worked, and the reasons why cooling and control of the humidity of the airstream are required. It also included the structure, appearance and function of each major component, and described the series and parallel arrangements of the air flow, coolant flow and hydrogen pathways through the EDCM module. The section on the explanation of the display was carefully prepared so that the same explanation text would work well for both the good and bad versions of the display, which was visible while this section was being read. For example, there was no elaboration of the objects inside of the EDCM module, since these were not shown in the bad version display. The explanation consisted of walking the subject through each of the system pathways, such as the route for the hydrogen flow, with an explanation about each object along the path.

The training materials included specific information about malfunctions because the types of possible malfunctions and their effects in the CS-1 were often not very obvious even to one with an extensive background in science and technology. For example, one effect of a malfunction in the cooling system is to produce lower voltages in the fuel cells. Why this happens is easy to understand, but it is not at all obvious from first principles. Thus, the training materials mentioned each possible type of malfunction and the general effects that would result. The object was to include all facts that would lie on a pathway of inference from the symptoms to malfunctions. For example, it was stated that one possible malfunction is for a cell to develop an internal leak so that hydrogen could get into the airstream and present the risk of explosion. A side effect of this leak would be that more hydrogen would be going into the EDCM than coming out, which would be indicated by the pressure and flow meters on the hydrogen inlet and outlets of the EDCM.

The object was to ensure that subjects knew all of the concepts underlying the system, how components might malfunction, and what the effects of each kind of malfunction would be. However, even though the training materials included all relevant facts, care was taken to avoid presenting simply a list of specific symptom–cause patterns. A prime example is that a problem with the hydrogen flow will often be manifested by a declining pattern of voltages, in which voltage E1 is higher than E2 which, in turn, is higher than E3, and so forth. The materials did not simply say that a declining voltage pattern meant that one of a specified set of faults

was present. Rather, the materials stated that if not enough hydrogen was available, the series connections of the cells meant that each cell would get less than the previous one, and a declining voltage pattern would result. The materials also included information about the assumptions governing the malfunction diagnosis task, the most important of which was that there could be only one fault present at a time.

Problems and response menus

A total of 20 malfunction problems were selected for use in the experiment, which can be found in Kieras (1986*b*). These malfunctions were based on a set of faults described by Malin and Lance (1987) for the FIXER system, with some distortions and simplifications introduced to simplify the materials and tasks in the experiment. Subjects specified their malfunction diagnosis by responding to a two-level menu. The first level was a menu of *ORUs* (Orbit Replaceable Units). This was simply a list of the boxed items shown in Figures 4 and 5, such as the EDCM and the Inlet Air Humidity Sensing Cluster, R1. There was an additional item in the ORU menu, which read *Malfunction is not in the CS-1 system,* which was the correct choice for malfunctions such as *Cabin air humid* or *Dry hydrogen* whose ultimate causes were located outside the CS-1 system. Each choice in the *ORU Menu* was associated with a *Specific Malfunction Menu,* which listed the specific faults that could occur within that ORU.

Design

There was one between-subjects factor consisting of the Good *vs* Bad display condition. There were 20 problems making up a within-subjects factor; each subject saw each problem in a random order. The subjects were assigned at random to the two conditions by assigning alternate subjects to the conditions in the order that they were run in the experiment. Care was taken to insure that the same proportion of males and females appeared in each condition.

Subjects

Because of the complex and realistic nature of the system, subjects were recruited who had a substantial background in science and technology. Thus, engineering students were recruited at the University of Michigan by means of advertisements posted around the Engineering College and by visiting engineering classes. It was found that only students in certain areas of engineering had adequate background; the best were students from aeronautical engineering classes, who also had considerable interest in an experiment involving actual space-vehicle equipment. Thus, the results may be generalizable only to similarly technically sophisticated populations, but such people, not the general public, are the potential audience for such displays. There were only two female subjects, one in each group. Each subject was paid \$10·00 for participating, and the experiment took about 2·5–3 h for each subject to complete. A total of 51 subjects were run, of which 11 were dropped. These consisted of one test subject, two subjects that quit during the course of the experiment, apparently as a result of its difficulty and eight subjects who were

dropped and replaced because they got less than 50% of the problems correct. The final sample size was 20 subjects per group.

Apparatus

The equipment used was the same as that from Experiments 1 and 2, except that no control panel device was involved. The subject viewed all of the instructional materials and response menus on a standard video terminal which was placed next to the Apollo color graphics display. The program running on the VAX signalled the Apollo to put up the display with the gauges showing the readings for the problem, and handled all sequencing and data recording.

Procedure

During the entire experiment the notebook of pictures and figures described above was available at all times. The subject first read all the way through the on-line manual. The subjects could back up to the previous screens within each section. The display of the CS-1 system was kept blank until the subject arrived at the section of the materials that explained the display, whereupon the display showing the normal readings appeared and remained on for the rest of the reading. After completing the training materials, the subject saw a menu with two options of either rereading the manual or beginning the problems. If the subject chose to reread the manual, they got the *Table of Contents Menu,* which contained an item to begin the problems. After reading a section of their choice, they could choose another section or begin the problems. If they chose to begin the problems, they next read the instructions for the malfunction-diagnosis task, and then began the problems. These instructions attemped to motivate the subjects to be as accurate as possible.

For each problem, the CS-1 display for that problem would appear on the Apollo screen, and a *Problem Menu* would appear on the video terminal with three options. The subject could choose to *Reread the manual*; they would go to a Table of Contents Menu that had an option to return to the Problem Menu. During this rereading the CS-1 display would stay on. If the subjects chose to *Report malfunction,* the CS-1 display immediately went blank and the ORU Menu would appear on the video terminal. The display was blanked to encourage subjects to do all of their reasoning before choosing to respond, and to discourage them from using the fault menus as a guide to their problem-solving process. The subject would choose an ORU for the site of the malfunction, and then would view the Specific Malfunction Menu for that ORU. The third alternative in the Problem Menu was a "Can't figure it out" choice which was made available to discourage subjects from guessing at random. This choice also appeared in the ORU Menu and the specific malfunction menu as well.

Once the subject chose a malfunction, they were told whether the choice was correct or incorrect, but they were not told the correct answer or why their choice was incorrect. After responding to all 20 problems, the problems that were answered incorrectly were presented again in a random order, and subjects had a second chance to get the problem right. The data recorded was the time spent on each portion of the on-line manual, each menu, and the responses made to the menus. From this information, several measures were calculated that will be described below.

RESULTS

Accuracy

A response was counted as correct if both the response to the ORU menu and to the specific diagnosis menu were correct. Table 8 shows the accuracy of subjects' diagnoses from the two conditions. There were significant, but weak, effects of the diagram display condition on accuracy. Considering accuracy in terms of a correct response on either the first or the second trial, the good version is somewhat better by about 6·5 percentage points. This difference is significant ($p = 0{\cdot}044$). There were strong effects of problem which are not shown in the table; there was a low of 5% correct and a high of 100% correct; this effect was highly significant. However, the interaction between display condition and problem was not significant using a log-linear analysis of the responses ($p > 0{\cdot}1$). In terms of the accuracy on the first attempt only, the good version condition was only about 4·5 percentage points better than the bad version; this is not significant ($p > 0{\cdot}1$). Again there is a strong problem effect, and an insignificant problem by condition interaction. The accuracy on the second trial is not shown; there were no significant effects of condition observed, due probably to substantial carry-over effects from the first try, as suggested by the times on the second try being much faster than the first.

The accuracy results suggest that the attempt to get subjects who were highly motivated and knowledgeable was reasonably successful; people were at the accuracy end of the speed–accuracy trade-off, in that they were as accurate as the problems, the task and their knowledge enabled them to be. Thus, the major effects of the diagram display manipulations would show up in the time taken to solve the problems.

Time results

Only the performance times on the first attempt were included, and for simplicity of presentation, the results averaged over both correct and incorrect trials are reported; the analysis of correct trials confirms the overall analysis, generally showing more significant effects. The mean times for each condition are shown in Table 8.

Total task time

This was measured from the first appearance of the Problem Menu to the last reponse in a Specific Malfunction Menu. The good version was roughly 30 s faster

TABLE 8.
Mean accuracy and time (s) for each condition

Measure	Condition	
	Bad version	Good version
Correct on either try	0·670	0·735
Correct on first try	0·500	0·545
Total task time	196·9	169·3
Response choice time	32·8	24·5
Total reading time	94·8	86·4
Text reading time	69·1	63·2
Observation time	58·0	45·5

on each problem than the bad version; the difference barely missed conventional significance in the analysis of variance ($p = 0{\cdot}0537$), but was strongly significant in the multiple regression analysis of only the correct trial times. There are strong differences due to the problem, but the problem × condition interaction was non-significant.

Response choice time

This was the time the subject took to choose a diagnosis response from the menus, given that they had already made a reponse to the Problem Menu that they were ready to report the malfunction. The subject could be doing additional reasoning during this time, but the display was blank, and the menus are identical for the two conditions. The good version produced somewhat faster response choice times, and this difference was significant. There were also strong effects of the problem, but again the problem × condition interaction was non-significant.

Reading times

The *total reading time* was the time that subjects spent reading measured from the point when they had chosen the *Reread Manual* response from the problem menu until they had returned to the problem menu. Thus, this time includes the subject choosing which section of the manual to read and might also include examining the display or thinking about the problem. The good version was somewhat better, as can be seen in Table 8, but this difference was not at all significant ($p > 0{\cdot}1$). Likewise the *text reading time* was the time spent only on reading the text frames and this did not include any time spent choosing the reading section, but again the subject could have been looking at the display or thinking about the problem. Again, the good version was slightly better in this regard, but not significantly.

Observation time

This time is the total time that the problem menu was left on the screen, and thus the subject is presumably looking at the display and thinking about the problem, and, apparently, not doing anything else, because neither the on-line manual nor the response menus are present. Thus, this is the purest measure of the time involved in using the display.

As shown in Table 8, the good version produced faster observation times, by roughly 12·5 s, which was significant. There are also strong differences in the problem and a strong problem × condition interaction as well. Some problems took substantially longer than other problems; this effect depended on the condition, in that some problems took similar times in the two conditions, while in others the bad version was substantially longer.

In order to analyze the difference between problems, a simple measure of their relative difficulty was needed. This was simply the number of gauges showing abnormal values: the number of gauges that read either high or low for the low-normal-high gauges, or gauge I1 being different from 10, or gauges E1 through E6 showing differences from 0·4. The set of gauges E1 through E6 was counted as one gauge, since there are only a few patterns of readings possible, and it seems that an abnormal set of readings can be recognized as a single pattern. This method of counting the gauges gave a low number of abnormal gauges of 1 and a high of 6,

with an average of 3·5. The regression lines between the number of abnormal gauges and the mean observation time on the 20 problems has a slope of about 10 s per gauge, and the R^2 is 0·637. This is a surprisingly strong relationship between a simple objective measure of the problem difficulty and the amount of time people spent observing the display and thinking about the problem. Of course, more is involved in solving these problems than simply looking at abnormal gauges, but to a first approximation the time required is apparently a function of how much information has to be processed, and this is closely related to the number of abnormal readings. A more detailed theoretical account would also consider which gauges had normal readings and were involved in making the inferences required to solve the problem, and also how much background knowledge reasoning was required in the inference. However, for present purposes this simple measure of problem difficulty seems to suffice.

When the regression line for the mean observation time spent on the problems is calculated separately for the two display conditions, the regression line for the good display has a higher intercept and a substantially smaller slope than for the bad version of the display. Figure 6 shows scatter plots of observation times and the number of gauges, along with the simple regression times for each. To test this hypothesis in more detail, a multiple regression analysis was done on the full set of

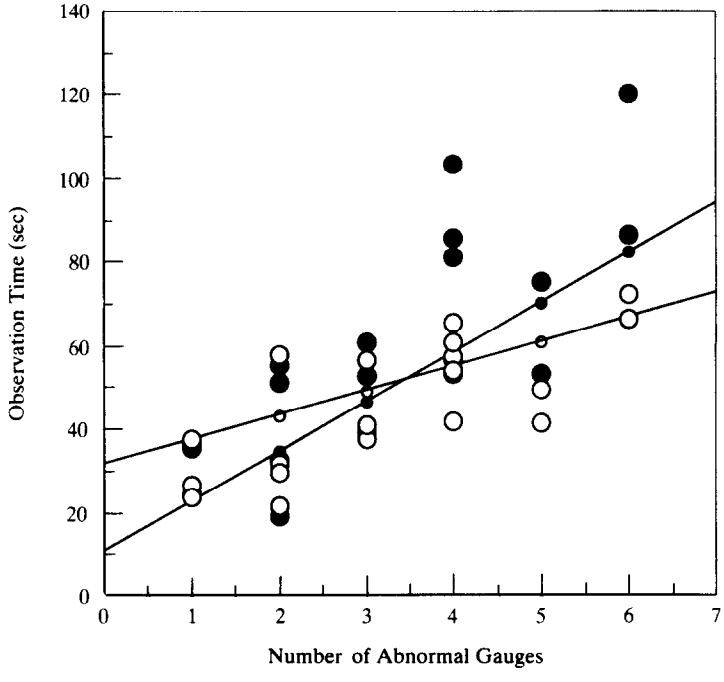

FIGURE 6. The relationship between observation time and the number of gauges showing abnormal readings for the two display versions. The regression lines are plotted through the mean predicted and observed values for the two conditions. The good version has about half the slope but a considerably higher intercept than the bad version, with a cross-over at about four gauges. • = bad predicted, ● = bad observed, ◦ = good predicted, ○ = good observed.

TABLE 9.
Multiple regression analysis on observation time

Predictor	Final coefficient	Final standard coefficient	*F*-to-remove
Intercept	−41·03		
Condition	20·75	0·233	14·16
Subject Mean	1·00	0·442	213·42
No. Gauges	11·90	0·490	145·52
Condition × No. Gauges	−6·04	−0·292	18·60

$N = 800$, $R^2 = 0{\cdot}34$.

first-try observation times ($N = 800$). The predictors were the individual subject mean observation time, the number of abnormal gauges and a dummy-coded condition factor that was 0 in the bad version condition and 1 in the good version condition, and an interaction variable which was the product of the number of gauges and the condition. The results of this analysis are summarized in Table 9. The final R^2 of 0·34 could be raised by including other significant factors such as the number of previous trials and effects of individual problems. However, there are highly significant effects of each of the predictors currently under discussion.

As can be seen in Table 9, each gauge produces an average increase in observation time of about 12 s, but the interaction between the display condition and the number of gauges means that in the good version display, only about 6 s per gauge is required. On the other hand, the good version has a higher baseline time of an additional 21 s. Thus, the more difficult problems involve more abnormal gauges, and so produce much longer bad version times. It is also clear that some problems take longer than predicted by this simple regression equation; these tended to be those involving cooling and humidity control, where the accuracy also tended to be the poorest. Intuitively, these seem to be the hardest problems in that they involved a rather subtle control loop, which involved the computer-based control module, which was not shown in the diagram. Further work would be required to investigate whether including a more complete topology would simplify these problems. However, it is important to note that this difficulty was more serious in the bad version of the diagram.

As shown in Figure 6, the crossover for when the visually more complex good version becomes superior to the bad version is at about four abnormally reading gauges. Apparently, the generally weak main effects of the display condition are simply a result of the problems being both above and below this crossover.

DISCUSSION

The purpose of this experiment was to determine whether the same principles observed in the first two experiments concerning the value of explicit topological content and visually useful state information would apply in a complex system with a static display in a diagnosis task. The results are that the more topologically complete diagram that has state information visually positioned at the topologically relevant places did produce superior problem-solving performance. The Good

version of the display requires less time per gauge than the Bad version, which is consistent with the theoretical model sketched above, in which the subject must process each piece of state information in conjunction with processing the diagram, and the bad version requires more processing of each such piece than the Good version. The evidence in favor of the topological completeness of the diagram being important is much less direct. The most difficult problems in the bad version took much longer than could be predicted from the number of gauges; these problems involved particularly difficult and subtle reasoning about the relationships between the inlet and outlet air humidity sensors and the cooling pump and control valve. Perhaps the more topologically complete diagram made these relationships easier to understand and to reason about. Some answers might be obtained by a more detailed analysis of the present data, but a better strategy would be future experiments directed more specifically at these questions.

Conclusions

SUMMARY

These studies, though just a beginning, show that topologically detailed dynamic displays can produce considerable benefits in operation and malfunction diagnosis tasks. However, the exact properties of good displays can be subtle. This paper characterized these properties using analysis at the level of specific information and events on the display and how they related to the subject's task. More state information can be a disadvantage if the visual effect is wrong, as in the Experiment 2 Medium level. There is apparently a trade-off between the value of topologically complete, visually associated displays and the overall visual complexity, as suggested by the higher baseline for the good version display in Experiment 3. Clearly these results raise many questions, and further work is needed on this fascinating and practically important form of user interface.

GUIDELINE ADVICE

The following guidelines for diagrammatic displays can be proposed, based on results from the above experiments.

Topological structure
Show the topological and causal structure of the system, such as the pathways between components, controls and indicators using conventions that are visually clear. Structural relationships involved in understanding system states must appear on the diagram.

Control and indicator states
Echo the topological effects of external controls, and show indicator states at the corresponding topological points on the system diagram.

Internal states
If information on the states of internal components is reliable and available, show the states that are significant to the user, so that there are no hidden states and no

inferences are required to deduce significant component states. Provide the state information at the corresponding topological point in the display.

Causal relationships
Show the pathway of causality through the topological structure, such as the color-coding of energized connections. Distinguish component states from other state information that may be on the displays (for example, by using different color–codes).

Malfunctions
Show failures of causal flow, such as malfunctions, in a perceptually salient way (for example, bright yellow for a component that fails to produce output when it should). Exactly what display properties are most suitable requires more investigation.

This research was supported by NASA under Grant No. NAG–9–139 to the author.

References

BISHOP, Y. M. M., FIENBERG, S. E. & HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT.

GOODSTEIN, L. P. (1981). Discriminative display support for process operators. In J. RASMUSSEN & W. B. ROUSE, Eds. *Human Detection and Diagnosis of System Failures.* New York: Plenum Press.

GOODSTEIN, L. P. (1982). An integrated display set for process operators. In G. JOHANNSEN & J. E. RIJNSDORP, Eds. *Analysis, Design and Evaluation of Man–Machine Systems. Proceedings of the IFAC/IFIP/IFORS/IEA Conference,* Baden-Baden, Germany.

GOVINDARAJ, T. (1988). Intelligent computer aids for fault diagnosis training of expert operators of large dynamic systems. In J. PSOTKA, L. D., MASSEY, S. A. & MUTTER, Eds. *Intelligent Tutoring Systems: Lessons Learned.* pp. 303–321. Hillsdale, NJ: Erlbaum.

HEPPNER, D. B., DAHLHAUSEN, M. J. & SCHUBERT, F. H. (1983). *Electrochemical carbon dioxide concentrator subsystem development: final report.* Technical Report No. 482–4. Cleveland: Life Systems, Inc.

HOLLAN, J. D., HUTCHINS, E. L. & WEITZMAN, L. (1984). STEAMER: an interactive inspectable simulation-based training system. *The AI Magazine,* **Summer,** 15–27.

HOLLAN, J. D., HUTCHINS, E. L., MCCANDLESS, T. P., ROSENSTEIN, M., & WEITZMAN, L. (1987). Graphic interfaces for simulation. *Advances in Man–Machine Systems Research,* **3,** 129–163.

KIERAS, D. E. (1988*a*). What mental model should be taught: choosing instructional content for complex engineered systems. In J. PSOTKA, L. D. MASSEY & S. A. MUTTER, Eds. *Intelligent Tutoring Systems: Lessons Learned.* pp. 85–111. Hillsdale, NJ: Erlbaum.

KIERAS, D. E. (1988*b*). *Diagrammatic displays for engineered systems: effects on human performance in interacting with malfunctioning systems* Technical Report No. 29. Ann Arbor: Univeristy of Michigan, Technical Communication Program.

KIERAS, D. E. (1990). The role of cognitive simulation models in the development of advanced training and testing systems. In N. FREDERIKSEN, R. GLASER, A. LESGOLD & M. SHAFTO, Eds. *Diagnostic Monitoring of Skills and Knowledge Acquisition.* pp. 51–73, Hillsdale, NJ: Lawrence Erlbaum Associates.

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

LARKIN, J. H. & SIMON, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science,* **11,** 65–99.

MALIN, J. T. & LANCE, N. (1987). Processes in construction of failure management expert systems from device design information. *IEEE Transactions on Systems, Man, and Cybernetics,* SMC-**17,** 956–967.

PEDHAZUR, E. J. (1982). *Multiple Regression in Behavioral Research.* 2nd edn. New York: Holt, Rinehart & Winston.

RASMUSSEN, J. & GOODSTEIN, L. P. (1988). Information technology and work. In M. HELANDER, Ed. *Handbook of Human–Computer Interaction.* pp. 175–202. Amsterdam: Elsevier.

REYNOLDS, H. T. (1977). *The Analysis of Cross-Classifications.* New York: The Free Press.

SHERIDAN, T. B. (1987). Supervisory control. In G. SALVENDY, Ed. *Handbook of Human Factors.* pp. 1243–1268. New York: Wiley.

WISE, J. A. (1986). Display systems for electrical system control centers. *Proceedings of the Human Factors Society-30th Annual Meeting* pp. 1264–1268.

WOODS, D. D. (1984). Visual momentum: a concept to improve the cognitive coupling of person and computer. *International Journal of Man–Machine Studies,* **21,** 229–244.

WOODS, D. D. & ROTH, E. M. (1988). Cognitive systems engineering. In M. HELANDER, Ed. *Handbook of Human–Computer Interaction.* pp. 3–43. Amsterdam: Elsevier.

WOODS, D. D., O'BRIEN, J. F. & HANES, L. F. (1987). Human factors challenges in process control: the case of nuclear power plants. In G. SALVENDY, Ed. *Handbook of Human Factors.* pp. 1724–1770, New York: Wiley.

WOOLF, B., BLEGEN, D., HANSEN, J. H. & VERLOOP, A. (1986). Teaching a complex industrial process. *Proceedings of the Fifth National Conference on Artificial Intelligence.* pp. 722–728. Philadelphia, PA: Morgan Kaufmann.