NASA CONTRACTOR REPORT 177484

Human Operator Response to Error-Likely Situations in Complex Engineering Systems

Nancy M. Morris
William B. Rouse

(NASA-CR-177484) HUMAN OPERATOR RESPONSE TO ERROR-LIKELY SITUATIONS IN COMPLEX ENGINEERING SYSTEMS (Search Technology) 47 p CSCL 05I
N89-18008
Unclas
G3/53 0191748

CONTRACT NAS2- 12048
August 1988

NASA CONTRACTOR REPORT 177484

Human Operator Response to Error-Likely Situations in Complex Engineering Systems

Nancy M. Morris
William B. Rouse
Search Technology, Inc.

Prepared for
Ames Research Center
Under Contract NAS2-12048

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

SUMMARY

The research reported in this paper is the result of an effort directed at understanding the causes of human error in complex systems. First, a conceptual framework is provided, in which two broad categories of error are discussed: errors of action, or slips, and errors of intention, or mistakes. Conditions in which slips and mistakes might be expected to occur are identified, based on existing theories of human error. Regarding the role of workload, it is hypothesized that workload may act as a catalyst for error.

Two experiments are presented in which humans' responses to "error-likely" situations were examined. Subjects controlled PLANT under a variety of conditions and periodically provided subjective ratings of mental effort. A complex pattern of results was obtained, which was not consistent with predictions. Generally, the results of this research indicate that 1) humans respond to conditions in which errors might be expected by attempting to reduce the possibility of error, and 2) adaptation to conditions is a potent influence upon human behavior in discretionary situations. Subjects' explanations for changes in effort ratings are also explored.

INTRODUCTION

When an engineering system fails to perform its function successfully, there is often an investigation (formal or informal) to determine the reasons for the system failure. This is especially the case when the system failure leads to disastrous consequences. Frequently, the result of such investigations is the conclusion that the failure was at least partially the result of "human error." This conclusion was reached after the explosion at Flixborough, the crash of Eastern Air Lines flight 401, and the incidents at Three Mile Island and Chernobyl.

The potentially catastrophic consequences of human error have prompted researchers and practitioners to devote considerable effort to identifying ways to deal with the problem. One such approach which has gained some popularity is to consider the human as another component in the system. As with other system components, humans have "failure rates," which denote probabilities that errors will be made when interacting with other system components. The likelihood of system failure is determined by identifying the probability of failure of each system component (including the human), and combining the probabilities (Swain and Guttmann, 1980). Should the overall system failure rate be too great, a variety of steps may be taken to reduce the overall rate, such as by incorporating redundancy into the system.

A variety of criticisms have been leveled at this probabilistic approach to dealing with human error, particularly when the errors of interest involve "higher level" activities such as decision making and problem solving (e.g., Sheridan, 1980). For example, data on human error rates are frequently unavailable. One aspect of human error hampering the availability of error rate data is that it is often difficult to determine what constitutes an error. It is also extremely difficult to identify an appropriate denominator for estimating a "rate." Further, it has been argued that the way in which rules for combining probabilities/rates are typically used may be inappropriate, since human errors are often not independent. Finally, even though an error may be made, humans tend to be self-correcting; thus, although an error may occur, the consequences of the error may be averted.

In contrast to the above probabilistic approach, this report describes research which was conducted for the purpose of learning more about the causes of human error. It is hoped that greater understanding of the contributing factors will provide insights into ways to deal with the problem of human error. The conceptual framework upon which the work is based is presented first, followed by a discussion of the experimental approach and data obtained. Finally, the resulting insights into human behavior are discussed, and the implications for research and applications are noted.

CONCEPTUAL FRAMEWORK

An examination of available literature on human error reveals that most writers distinguish between at least two types of error. One common distinction is between errors of *action* (frequently discussed as "slips") and errors of *intention* ("mistakes"). A slip occurs when an intention is not executed as planned (e.g., an automobile driver turning on the windshield wipers instead of turning on the lights as intended). A mistake is the result of an inappropriate choice of intention (e.g., the hapless driver, stranded due to a lack of fuel, might seek to rectify the situation by replacing the battery).

Of the two types of error noted, more has been written concerning slips. The following discussion of slips and mistakes is based upon the work of Reason (1983; Reason and Mycielska, 1982), although Norman (1981) presented a discussion of slips that is conceptually similar. These authors have provided rather detailed conceptualizations to explain how slips and mistakes might arise, and the interested reader is referred to their work for further details. They are not discussed in depth here because the primary concern of the present research is with the implications of these models.

Characteristics of Slips

Relying primarily on information about errors recorded in diaries, Reason and Mycielska (1982, p. 21) discussed three general characteristics of slips:

1. Slips occur during the largely automatic execution of some well-established or routine sequence of actions....

2. Slips appear to be associated with distraction or preoccupation....

3. [Slips appear] to flourish in relatively familiar environments where there are few departures from the expected....

Perhaps the key contributor to the occurrence of slips is automaticity. Slips are the result of actions which are not consciously monitored. Thus, experts are not immune from slips; rather, in some cases, the expert may be *more* prone to such errors.

Of the many types of slip discussed, three were identified as being of interest to the present research. *Strong habit intrusions* appear to occur most frequently, and are the result of frequent behaviors replacing less frequent ones. Strong habit intrusions may occur when 1) there is a change in routine (e.g., failure to stop at the grocer's on the way home from work as intended); 2) one's routine has not changed, but other circumstances have (e.g., walking to where one's favorite chair once was, only to recall that the furniture has been rearranged); and 3) behavior is "captured" by features of the environment (e.g., putting on one's coat instead of retrieving the box stored on the shelf of the coat closet).

Unusual or ambiguous situations may lead to *misperceptions*. The accuracy of perceptions may be influenced by 1) frequency (unusual objects or events may be misperceived as more common ones); 2) incongruity (things which do not "belong" in a setting may not be accurately perceived); 3) context (the surrounding context may be used to clarify ambiguous details); and 4) need (e.g., the person who is hungry may perceive ambiguous stimuli as food).

Omissions or repetitions of steps in intended sequences of actions may be attributed to *place-losing errors*. This type of error may be described as "failure of the program counter." Factors which may influence the occurrence of place-losing errors include forgetting an action which has already been performed, anticipation of actions to come, and recollection of a previous unresolved intention.

Situations in Which Slips Are Likely

Based on this analysis, slips might be expected to occur in the following operational conditions:

1. Salient environmental cues are not relevant to the current intention.

2. Features of the environment have changed but the task to be accomplished has not (e.g., the arrangement of displays and controls has changed).

3. The intended routine has changed but environmental features have not (e.g., a slight change in a well-practiced procedure).

4. Environmental cues are unusual or ambiguous.

5. A long series of actions is required to accomplish a goal.

6. The time period between related actions is long and/or filled with other activity.

7. Procedures required to accomplish different goals are similar in places.

Characteristics of Mistakes

In discussing the origins of mistakes, Reason (1983) cites a number of biases in judgment and planning which have been well-documented in literature on decision making and problem solving. For example, decision makers appear to consider no more than two or three variables at a time. When retrieving information to be used in making a decision, some retrieval may be triggered by salient but irrelevant environmental cues. Attempts are then made to incorporate this irrelevant information into the decision.

Recall of information is also biased toward availability, and the influence of past successes may be inappropriately large. Missing pieces of information may be supplied by the decision maker, based on his/her "theories" of what that information should be; later, it may be impossible to distinguish such self-generated information from the real thing. Once a hypothesis has been selected, there is a tendency to seek confirming evidence and "explain away" counter evidence. Finally, humans have a tendency to be overconfident of the correctness of their state of knowledge.

In light of these and other biases, Reason has identified three general sources of mistakes. Some mistakes are due to *bounded rationality*, and are characterized by oversimplification of the problem. Thus, decisions made in situations in which the impact of several interacting variables should be considered may not be appropriate.

Some mistakes may be attributed to *imperfect rationality*. When searching for a solution, those approaches which have been successful in the past are most salient. The result is that decisions may be too bound to the past and too conservative.

Finally, some mistakes arise due to *reluctant rationality*. Sustained attention to weighing available evidence and considering alternatives is difficult to maintain for long periods of time and humans have a tendency to "jump to conclusions." Thus, decisions may be heavily influenced by salient environmental cues, particularly if they are familiar and indicate solutions which are well-tried. This process is efficient most of the time, but leads to mistakes when the salient environmental cues are not sufficient for an adequate decision.

Conditions in Which Mistakes Are Likely

If the preceding analysis is accepted, then the following conditions should lead to an increased likelihood of mistakes:

1. Making an appropriate decision requires the simultaneous consideration of more than two or three variables.

2. Salient environmental cues suggest a solution which is inappropriate.

3. An approach which is inappropriate for the current situation has been used successfully many times in similar situations.

4. Choice of a solution requires approaching the problem in a novel way.

The Role of Workload

In addition to the conditions listed above, another factor which may influence the occurrence of errors is imposed mental workload. Mental workload is not viewed as a cause of error, however, but rather as a *catalyst*. In other words, an increase in imposed mental workload may not lead to more errors unless other conditions are conducive to error. For example, incompatibilities in the design of displays and controls may not result in slips if the human operator is "careful" and monitors his/her actions closely. However, the requirement to perform more than one task may make it impossible to monitor one's actions as closely, and errors may result.

While the possible impacts of imposed mental workload can be anticipated fairly easily, it is somewhat more difficult to predict the effects of subjective mental workload on error. It is possible that low subjective load could be associated with more slips since the human may not feel a need to monitor lower-level actions. An increase in subjective mental workload might cause the human to "sit up and take notice," and monitor his/her actions more closely; on the other hand, high subjective load could be "distracting," and lead to even more slips. It is possible to make similar arguments for the effects of subjective load upon mistakes, dependent upon whether one views subjective load as motivational or stressful.

Of course, the occurrence of an error may affect subjective mental load. If the error goes unnoticed by the human, chances are that the need to deal with any consequences of the error as they become manifest may increase subjective load. It also seems that higher subjective mental load might be associated with knowing one has committed an error, particularly if error correction is required.

EXPERIMENT ONE

The research described in the following pages was designed to accomplish two goals. First, a greater understanding of the causes of human error was sought by creating circumstances in which errors might be expected to occur and evaluating human behavior in those conditions. Second, investigation of the relationships between error and workload was planned. Both imposed load and subjective load were of interest.

Subjects

The six subjects were recruited through a local area technical school and were paid for participation. An effort was made to obtain subjects with training and orientation similar to those of actual operators, although the pool of students in potentially relevant degree programs was small. Two of the subjects were sixth-quarter students in the school's electromechanical engineering program; two were first-quarter students in the electronics program. One subject was a sixth-quarter student in the data processing program, which placed heavy emphasis on troubleshooting of hardware and software. The sixth subject was not a student, but was employed full-time as the operator of a computer-based climate control system for a network of buildings.

Experimental Task Environment

Brief description of PLANT. The experimental task used was PLANT, a computer-driven simulation of a generic dynamic production process (Morris, Rouse, & Fath, 1985). During PLANT operation, both graphic and alphanumeric information is displayed to the PLANT operator via two color monitors. The graphic display for a sample PLANT problem is shown in Figure 1.

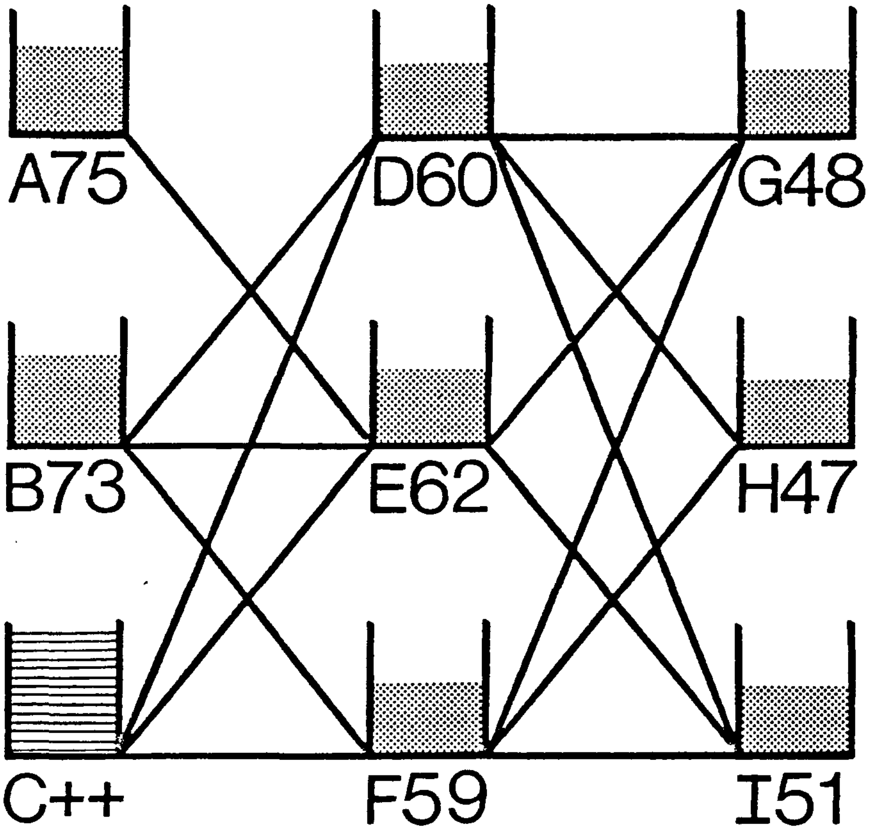

Figure 1. PLANT graphic display.

The system in Figure 1 contains nine tanks labeled A through I. Some of the tanks are currently connected by open valves, as indicated by lines between tanks. Numbers beneath tanks represent the current levels of fluid in them. (The "+ +" under tank C indicates a level that is too high to be displayed and, therefore, unsafe.) Fluid enters the PLANT system at the left and exits at the right as finished product.

In general, the PLANT operator's task is to supervise the flow of fluid through the series of tanks interconnected by pumps, valves, and pipes so as to produce an unspecified product. The operator may open and close valves, adjust system input and output, check flows between tanks, and order repairs of various PLANT components by typing commands at the keyboard. Maximizing production is the primary goal. However, as in real systems, the "physical" limitations of the system (such as tank capacity or reliability of system components) require that the PLANT operator be concerned with secondary goals as well. Among these secondary goals are stabilization of the system, and detection, diagnosis, and compensation for system failures. Instability is manifest by valve "trips" (automatic closing of valves by the PLANT safety system). The symptoms of failure are varied, as discussed below.

Modifications to PLANT. The current version of PLANT is implemented in Pascal on an IBM PC/AT and uses two Amdek 600 color monitors. Verbal ratings of mental effort were obtained via a Votan VPC 2000 speech recognition and playback device. A headset contained both the speaker and microphone. When ratings were requested, a horizontal ten-point scale was prominently displayed above the monitor containing the PLANT graphics display. "RATE YOUR MENTAL EFFORT" appeared above the scale; descriptors below the scale included "extremely low effort" below the number 1, "extremely high effort" below the number 10, and "moderate effort" centered below 5 and 6.

For purposes of this experiment, PLANT was modified in the following ways. First, the capability for forced-pacing was implemented so that PLANT could be updated automatically if desired. (In previous experiments, PLANT was self-paced only.) When forced-pacing was in effect, PLANT was automatically updated every 4 sec.

The way in which PLANT commands were entered was also changed. PLANT commands consist of two parts: a two-letter action (e.g., "ov" = open valve), plus a specified object ("ovad = open valve A-D) or quantity ("pi100" = pump in 100 units). Keys on the AT keyboard were redefined so that the two-letter commands could be entered with single keystrokes (but echoed on the alphanumeric screen as two letters). The number pad was used for specifying the remainder of the command. Digits retained their normal positions on the number pad; the nine tanks in PLANT (three rows by three columns) were also represented with the letters A through I on the keys corresponding to the digits one through nine.

Thus, issuing a PLANT command involved a minimum of three keystrokes: 1) action selection from the alphabetic keyboard, 2) selection of object or quantity from the number pad, and 3) a carriage return. Echoing a number or letter in response to pressing a key on the number pad was determined by the type of information required by the action selected. Necessary keys were clearly marked with adhesive labels, and all other keys were covered and disabled.

The most substantial change to PLANT involved increasing the number of failure types which could occur. Possible failures could be grouped into two broad classes. Valve and internal pump failures involved a stoppage of fluid flow through one or more valves, and were considered *simple failures*. Diagnosis of a valve or pump failure was based on a single unambiguous piece of information, a flow reading of 0.00 through one or more valves. The other types of failure that could occur were *complex failures*, and included failure of an input or output pump, tank rupture, display failure, and safety system failure. Accurate diagnosis of one of these failures was not possible with a single source of information, but rather required consulting two or three sources. There was, however, always sufficient evidence available to make an unambiguous diagnosis. Each piece of information used could be symptomatic of more than one failure, so that unambiguous diagnosis required integration of the available symptoms.

Experimental Procedure

Subjects controlled PLANT for 26 periods (or "production runs") of approximately 20 min each. Each production run consisted of 350 system updates or iterations. Unless otherwise noted, production runs were self-paced. Production runs were grouped into 14 one-hour sessions. The first eight sessions were training sessions and consisted of production runs interspersed with instructions as described below. Experimental conditions were manipulated in the last six sessions. Subjects were allowed to complete as many sessions per visit as they wished.

Training sessions. At the beginning of session 1, subjects received written and verbal instruction in the basics of PLANT operation. The simple valve and pump failures were the only types of failure discussed at this time, and were the only types to occur in production runs until the later introduction of the complex failures. A formal test of the material was administered at the end of session 2 after the subjects had controlled PLANT for two production runs.

Procedures for PLANT operation were provided at the beginning of session 3, and were available in hard-copy form during all subsequent production runs. Subjects also began keeping logs in session 3 and continued for the remainder of the experiment. The experimenter explained that the logs would help her to understand "why they did what they did," and that they should make notes which might explain their actions while controlling PLANT. The subjects understood that these notes were to be made "at their convenience" and were not to interfere with PLANT control.

The effort-rating process was introduced in session 4. For each subject, the Votan speech recognition system was trained to recognize the spoken numbers "one" through "ten." Subjects were then asked to rate their mental effort (when prompted) on a scale of one to ten, with ten being the highest. All subjects indicated that they could comply with this request without difficulty, and there did not appear to be any confusion as to what was being asked of them.

Effort ratings were requested every 10 iterations in all subsequent production runs. Every tenth iteration, subjects were prompted through the Votan headset with the recorded word "effort." If a recognizable rating was not obtained within three iterations, the prompt "again" was heard. Subjects were informed that the second prompt was an indication that Votan had not understood them the first time, and that they should try to speak more clearly. If a recognizable rating was not obtained within eight iterations following the first prompt, the failure to obtain a rating was noted in the data file.

At the beginning of session 5, information about the nature and diagnosis of complex failures was provided. Each of the complex failures occurred at least once during the next seven production runs (runs 8-14), and the manner in which they should be diagnosed was reviewed frequently with subjects. Forced pacing was introduced in production runs 12-14.

Each production run contained five failures, with one occurring approximately every 75-80 iterations. Prior to production run 13, subjects had experienced no more than one complex failure per run. In run 13, there were three complex failures, and all subjects experienced at least two of them concurrently due to failure to repair one before another occurred. Immediately after run 13, subjects were given the opportunity to "recalibrate" their effort ratings. They were first asked to imagine and describe the easiest possible situation in PLANT (the most commonly noted was start-up at the beginning of a production run), and assign that situation a very low effort rating of perhaps one or two. Then, they described a very difficult situation (most referred to the production run just completed), and were asked to give that situation a very high effort rating close to ten. Before beginning run 14, it was suggested that they use the imagined situations as reference points when assigning effort ratings in the future. All subjects said they thought they could comply with this suggestion.

Experimental sessions. The experimental conditions were manipulated in sessions 9-14 (production runs 15-26). Three characteristics of PLANT were varied: 1) pacing, 2) display-control compatibility, and 3) the types of failure which could occur in a production run.

Imposed load was manipulated by making PLANT forced-paced (high load) or self-paced (low load). Variation of display-control compatibility was expected to alter the likelihood of slips occurring (i.e., errors in entering intended commands). In the "compatible" (low-slip) condition, the arrangement of tank labels on the number pad was isomorphic to the arrangement of tanks on the PLANT graphic display. In the "incompatible" (high-slip) condition, the tank labels on the number pad were inverted so that rows became columns and vice versa; digits, however, always occupied the same positions.

The expected likelihood of mistakes (in the form of incorrect diagnoses) was manipulated by controlling the types of failure which occurred during a production run. In the "simple failure" (low-mistake) condition, only simple failures occurred. Failures in the "complex failure" (high-mistake) condition included two complex failures and three simple failures.

Factorial combination of these attributes resulted in eight experimental conditions. Subjects controlled three production runs under each of the compatible, complex failure conditions (i.e., three each under forced- and self-pacing). So that the complex failure and incompatible situations would be more unusual and (hopefully) generate more errors, subjects controlled only one production run for each condition involving one of the high-error manipulations. The order of presentation for the twelve experimental production runs was determined pseudorandomly with the constraints that 1) compatible, simple failure runs were interspersed fairly evenly throughout the sequence, 2) no more than two runs of the same pacing occurred in sequence, and 3) high-error conditions occurred in self-paced runs before they occurred in forced-paced runs. The same order of presentation was used for all subjects.

Dependent Measures

Four classes of dependent measure were considered. These included 1) performance measures, 2) errors, 3) effort ratings, and 4) other behavioral measures.

Performance measures. The primary index of performance in controlling PLANT was the amount of production achieved. Based on previous experience with PLANT, a number of other indices of performance were used. Among these were 1) frequency of automatic valve closings on the part of the safety system ("trips"), 2) average number of open valves per iteration, 3) average levels of input and output specified, 4) average variance in fluid levels across the system, 5) average number of iterations to repair simple failures, and 6) average number of iterations to repair complex failures.

Errors. By examining the transaction files created during PLANT operation, it was possible to identify a number of actions which were erroneous. Only those actions which were unambiguously incorrect or inappropriate were designated as errors. All of the errors noted could be classified as such by considering the immediate PLANT context; there was no attempt to infer subjects' intentions from a sequence of actions and judge those intentions as appropriate or inappropriate. Thus, most of the errors discussed would be considered slips rather than mistakes.

Based on detailed analysis of transaction files from an earlier pilot study, five categories of error were distinguished. *Uncorrected syntax errors* ("UNCSYN") consisted of actions entered by the subject that were syntactically incorrect (e.g., an incomplete specification of a valve or a command to open a valve which did not exist). *Corrected syntax errors* ("CORSYN") were identical to uncorrected syntax errors, with the exception that the subject corrected them (via backspacing) before entering a carriage return.

A third class of error, *inappropriate actions* ("INAPP"), included actions which were syntactically correct, but to which PLANT could not respond. Examples of such errors were attempting to dispatch the repair crew while they were already occupied, opening a valve that was already open, or setting the input rate to its current value. Examination of the data revealed that the primary source of INAPP errors was opening valves which were already open. This appeared to be due to the subjects' open-loop control of PLANT (i.e., entering a sequence of commands without waiting for the system to update between commands).

The fourth class of actions considered in the error analysis consisted of commands which were syntactically correct but were changed by the subject to other actions. It was impossible to determine whether these *"okay" changes* ("OKCHNG") were corrections of typographical errors (which happened to be syntactically acceptable), or were the results of changes in subjects' intentions (which might not have been errors at all). Thus, such actions were noted as interesting behavior and retained as a separate category when analyzing errors.

The final category of error included only errors of diagnosis: failing to repair a fault or repairing a PLANT component that had not failed. Careful examination of the data in this experiment revealed that inappropriate repair of a component could be clearly construed as a slip in only one instance: an experienced subject observed a flow reading of zero (an unambiguous indication of a valve failure) and then repaired the wrong valve. In all other cases, inappropriate repair appeared to be the result of misdiagnosis. This slip was included in the UNCSYN category (as suggested by the subject) and the others were considered mistakes ("MISTAKES").

Effort ratings. Two characteristics of effort ratings were noted: 1) the value of each rating, and 2) the number of ten-iteration intervals in which ratings were not recognized by Votan. Each production run consisted of 350 iterations, so a maximum of 35 effort ratings per run was possible. Occasionally no rating was received within a ten-iteration interval. Examination of data files revealed that Votan was quite sensitive to extraneous low-level speech (e.g., subjects talking to themselves), and that multiple ratings were frequently obtained (due to inappropriate classification of extraneous speech as digits by Votan). In light of these multiple ratings and the fact that subjects were prompted twice if necessary, a missing rating was interpreted as absence of a response from the subject rather than a failure of Votan to understand a verbal rating.

Multiple ratings were resolved by blind experimenter judgment, with no knowledge of the experimental conditions associated with any given series of ratings. In the rare cases in which it was not possible to be confident of the true rating, the rating was recorded as missing.

Other behavioral measures. Elapsed real time between actions was also noted, as was the frequency with which each type of PLANT command was issued successfully (i.e., excluding errors).

RESULTS

The following discussion of statistical analyses presents a complex set of results. The fact that the results require a complicated description is not too surprising, in light of the opportunities for complexity afforded by the number of dependent measures examined. Following the detailed presentation of significant effects, a more coherent and succinct interpretation of the results is provided.

Based on examinations of data files and interviews with subjects, it was determined that the control behavior of two subjects was erratic and departed radically from prescribed approaches to controlling PLANT. These subjects (coincidentally, the two first-quarter electronics students) achieved much less production, kept fewer valves open, and made many more errors than did the other subjects. Conversations with these subjects revealed that departures from prescribed PLANT control were not the result of differing strategies, but rather reflected a lack of understanding of PLANT. (For example, they could not answer questions about PLANT drawn from the written instructions they had received at the beginning of the experiment.) Thus, the decision was made to exclude their data from further analysis, and the results reported here are based on data from four subjects.

The analyses performed fell into two categories. First, the effects of experimental conditions on the dependent measures were investigated. Three-way analysis of variance with repeated measures was the primary statistical tool used for this purpose. In discussing the results of these analyses, the factors are designated as follows: Compatibility (compatible vs. incompatible keyboard arrangement), Failure Complexity (simple failures only vs. both simple and complex failures), and Pacing (self- or forced-pacing). In addition to examining the effects of experimental conditions, the relationships between dependent measures were also explored using time series analysis.

Effects of Experimental Conditions

Performance and other behavioral measures. Of the many performance measures recorded, only two were significantly affected by experimental conditions. There was a significant interaction effect of Compatibility and Pacing on production ($F(1,3) = 9.96, p = .05$). Production was lower in the compatible, forced-paced condition than in the other three combinations of Compatibility and Pacing (94,171 vs. a mean of 109,240).

There was a three-way interaction effect ($F(1,3) = 36.64, p = .01$) of Failure Complexity, Compatibility, and Pacing on the average level of output specified per output command. In the simple failure conditions, the effects of Compatibility and Pacing were similar to those described for production: lower levels of output were specified in the compatible, forced-paced condition than in the other three combinations of Compatibility and Pacing (99.52 vs. a mean of 109.57). However, the effects were different in the complex failure conditions: average output was approximately equal in the two forced-paced conditions (a mean of 89.75), and lower than in the compatible, self-paced (101.78) and incompatible, self-paced (110.90) conditions. The latter two were also different from each other.

Command usage across experimental conditions is presented graphically in Figures 2 and 3. Four categories of command are represented in these figures: 1) repair of PLANT components ("reprs"); 2) flow readings ("flows"); 3) adjustments to input or output ("pi/po," short for "pump in/pump out"); and 4) opening or closing valves ("op/cl"). The three-letter labels on the abscissa represent the eight experimental conditions. These letters refer to the Failure Complexity (Simple vs. Complex), Compatibility (Compatible vs. Incompatible), and Pacing (Self vs. Forced) manipulations, respectively.

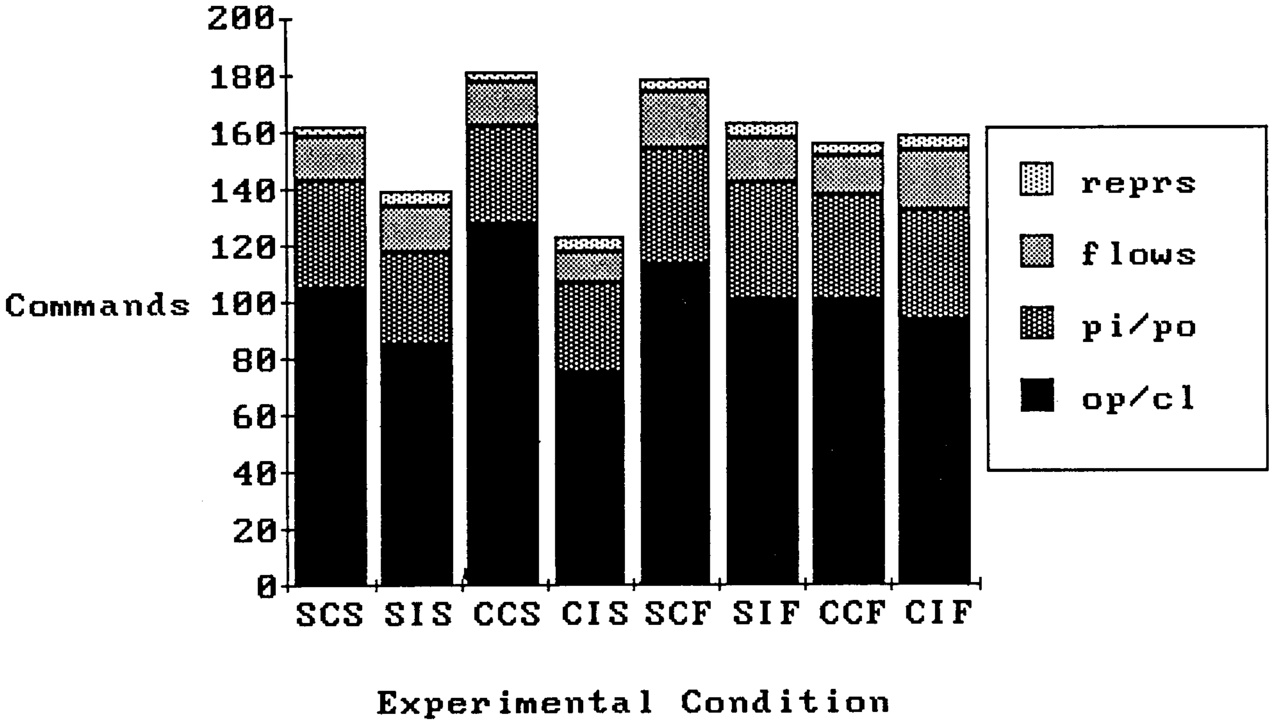

Figure 2. Command usage across experimental conditions--Experiment One.

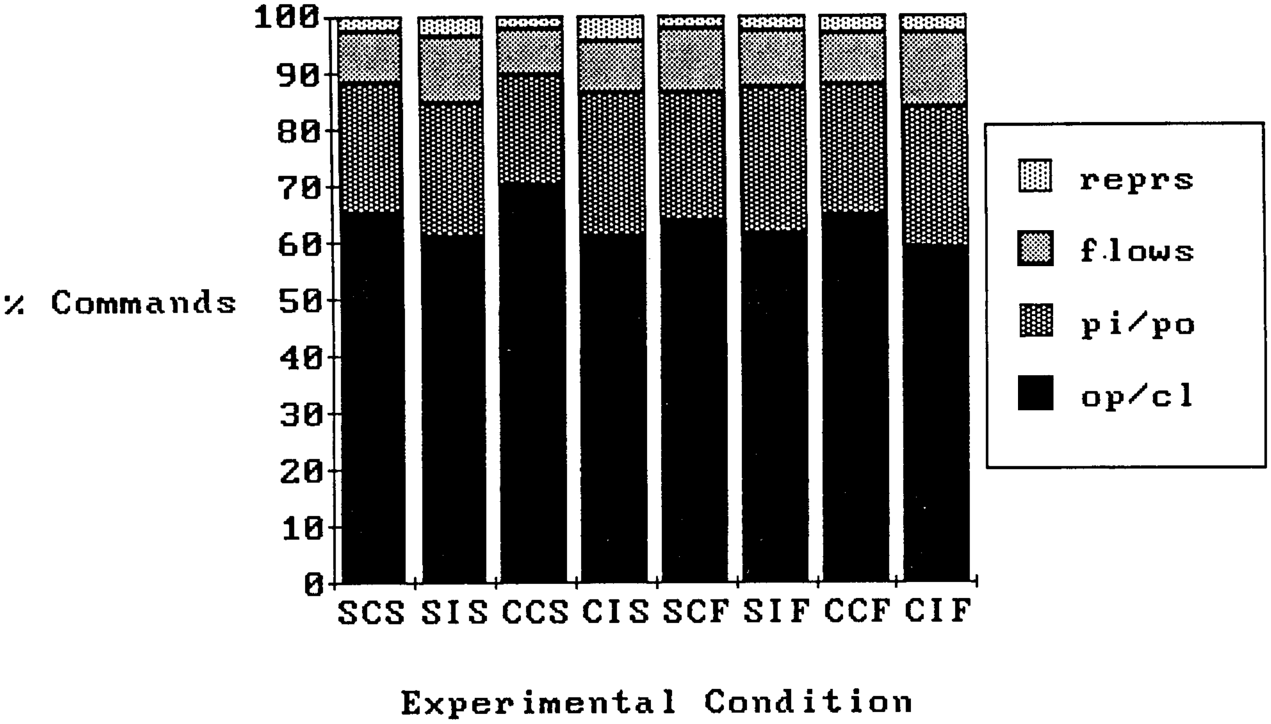

Figure 3. Percent of each type of command used across experimental conditions--Experiment One.

Analysis of command usage revealed the following significant effects. More commands were issued in compatible runs than in incompatible runs (169 vs. 146, F(1,3) = 20.61, *p* = .02). Closer examination indicated that this difference was attributable primarily to a difference in the number of "open" and "close" commands issued in those runs (109 vs. 88, F(1,3) = 20.67, *p* = .02). The number of input and output commands issued was significantly affected by Pacing (34 for self-paced runs vs. 39 for forced-paced runs, F(1,3) = 12.13, *p* = .04).

Frequency of flow readings was affected by the interaction of Failure Complexity, Compatibility, and Pacing (F(1,3) = 55.56, *p* = .005). Fewer flow commands were issued in the complex failure, incompatible, self-paced condition than in any others (10.75). The greatest number of flow commands were issued in the complex failure, incompatible, forced-paced condition and in the simple failure, compatible, forced-paced condition (which were approximately equal, with a mean of 20.58). There were no differences in flow readings among the other experimental conditions (a mean of 15.06).

As a final observation relative to the behavioral measures, subjects completed self-paced runs in 21% less time than forced-paced runs (18.5 vs. 23.3 minutes, F(1,3) = 10.14, *p* = .05).

Error measures. Errors observed in each experimental condition are presented graphically in Figures 4 and 5. Significant differences may be summarized as follows. Forced-pacing resulted in more CORSYN errors (4.73 vs. 2.59 with self-pacing, F(1,3) = 34.31, *p* = .01) and more total syntax errors (i.e., UNCSYN + CORSYN) (11.27 vs. 5.51 with self-pacing, F(1,3) = 12.14, *p* = .04). Also in the expected direction were the marginal effects of Pacing on UNCSYN errors (7.11 with forced-pacing vs. 2.86 with self-pacing, F(1,3) = 8.26, *p* = .06) and total changes (i.e., CORSYN + OKCHNG) (9.76 with forced-pacing vs. 5.72 with self-pacing, F(1,3) = 9.21, *p* = .06).

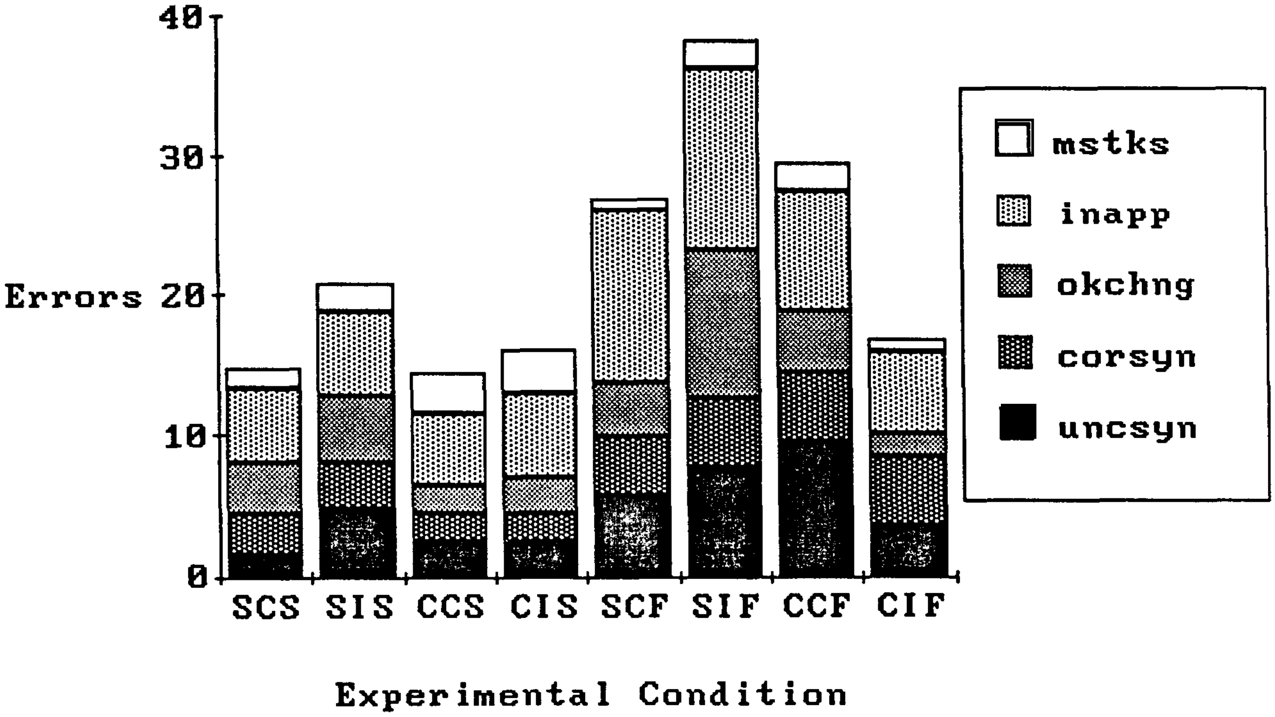

Figure 4. Errors across experimental conditions--Experiment One.

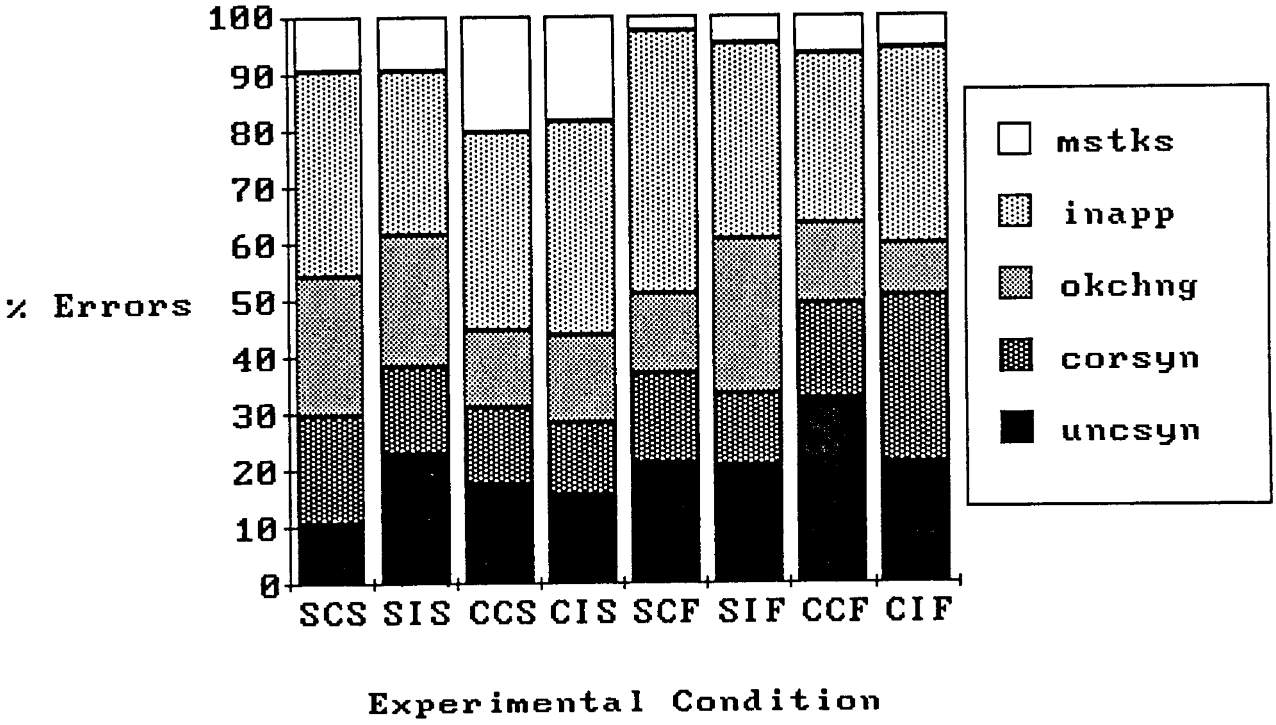

Figure 5. Percent of each type of error occurring across experimental conditions--Experiment One.

Two effects of Failure Complexity were noted. First, the frequency of OKCHNG was greater in the simple failure conditions than in complex failure conditions (5.62 vs. 2.56, F(1,3) = 15.19, *p* = .03). Second, there were marginally more MISTAKES associated with complex failure conditions (2.37 vs. 1.69 with simple failure conditions, F(1,3) = 8.89, *p* = .06).

The remaining effects involved the interaction of Failure Complexity with either Compatibility or Pacing. The Failure Complexity x Compatibility interaction had a significant effect on UNCSYN errors (F(1,3) = 20.67, *p* = .02). Fewer UNCSYN errors occurred in the complex failure, incompatible condition (3.0) than in the other three combinations of Failure Complexity and Compatibility, and fewer occurred in the complex failure, compatible condition (4.7) than in either the simple failure, incompatible (6.25) or complex failure, compatible (6.0) conditions.

Both the Failure Complexity x Compatibility and Failure Complexity x Pacing interactions had significant effects upon INAPP errors. Regarding the Failure Complexity x Compatibility interaction (F(1,3) = 35.17, *p* = .01), more INAPP errors occurred in the simple failure, incompatible conditions than in the other three combinations of Failure Complexity and Compatibility (12.0 vs. a mean of 7.1). Considering the Failure Complexity x Pacing interaction (F(1,3) = 21.03, *p* = .02), more INAPP errors occurred in the simple failure, forced-paced conditions than in the other combinations of Failure Complexity and Pacing (15.14 vs. a mean of 6.0).

Surprisingly, there was no significant main effect of Compatibility on errors. However, examination of errors occurring during the first ten iterations of production runs revealed that a total of 12 errors occurred in the first ten iterations of incompatible runs, whereas only 6.3 errors occurred in the first ten iterations of compatible runs. (The number cited for compatible runs is not an integer because it has been adjusted for multiple simple failure, compatible runs.)

Effort ratings. Significant differences due to Failure Complexity and Pacing in the number of effort ratings received were observed. Fewer ratings were obtained in the complex failure conditions than in the simple failure conditions (29.5 vs. 31.1, F(1,3) = 11.04, p = .05), and fewer ratings were received in the forced-paced conditions than in the self-paced conditions 29.5 vs. 31.1, F(1,3) = 21.17, p = .02). When missing values were treated as missing data (i.e., the analysis included compensation for unequal cell sizes), no differences in the values of ratings were found. In light of the pattern of differences in number of ratings obtained, missing ratings were assigned a value of 10 and the analysis was repeated; however, no differences in magnitude of ratings were observed.

Relationships Between Measures

As noted, relationships between dependent measures were investigated using time series analysis. Four variables were compared: production, number of open valves, total errors, and effort ratings. Recall that effort ratings were requested every ten iterations. Values of the other three variables were computed for each ten-iteration interval during the production run; comparisons with effort ratings involved values for the ten iterations preceding the rating.

Full analyses were performed for each individual production run, resulting in 48 analyses (4 subjects x 12 sessions). Autocorrelations for each variable and cross-correlations for each pair of variables were obtained. Although several significant relationships were noted, it was apparent in light of the large number of results and plots that some aggregation was needed to facilitate interpretation.

"Averaged" correlation functions were created by collapsing across subjects (i.e., computing mean values for each point in the function), and the results were plotted. For some production runs the averaged functions exhibited clear relationships between variables, whereas for others the relationships were not strong. However, these differences in relationships between variables did not appear to be related to any of the experimental treatments.

Averaged functions were then created for each subject by collapsing across conditions. Observation of the plots of the resulting functions revealed strong similarities across subjects. Thus, as a final step, a single plot was created for each variable and pair of variables by collapsing across both subjects and conditions. Those plots which exhibited clear relationships (i.e., autocorrelations or cross-correlations exceeding ±0.3) are shown in Figures 6 through 10.

It is easier to obtain an overall perspective of the results obtained by considering the summary plots rather than the original results of the time series analyses. However, it is important to note the extent to which the summary is representative of individual analyses, as discussed here.

Autocorrelations. Referring to Figure 6, the relationship of production to itself was strong and positive, and observed to a greater or lesser degree in all 48 analyses. The strongest relationship was found with a lag of one interval (i.e., one ten-iteration period of approximately 31-40 sec, dependent on pacing condition). As may be ascertained from Figure 6, the strength of the relationship was substantially diminished with a lag of two.

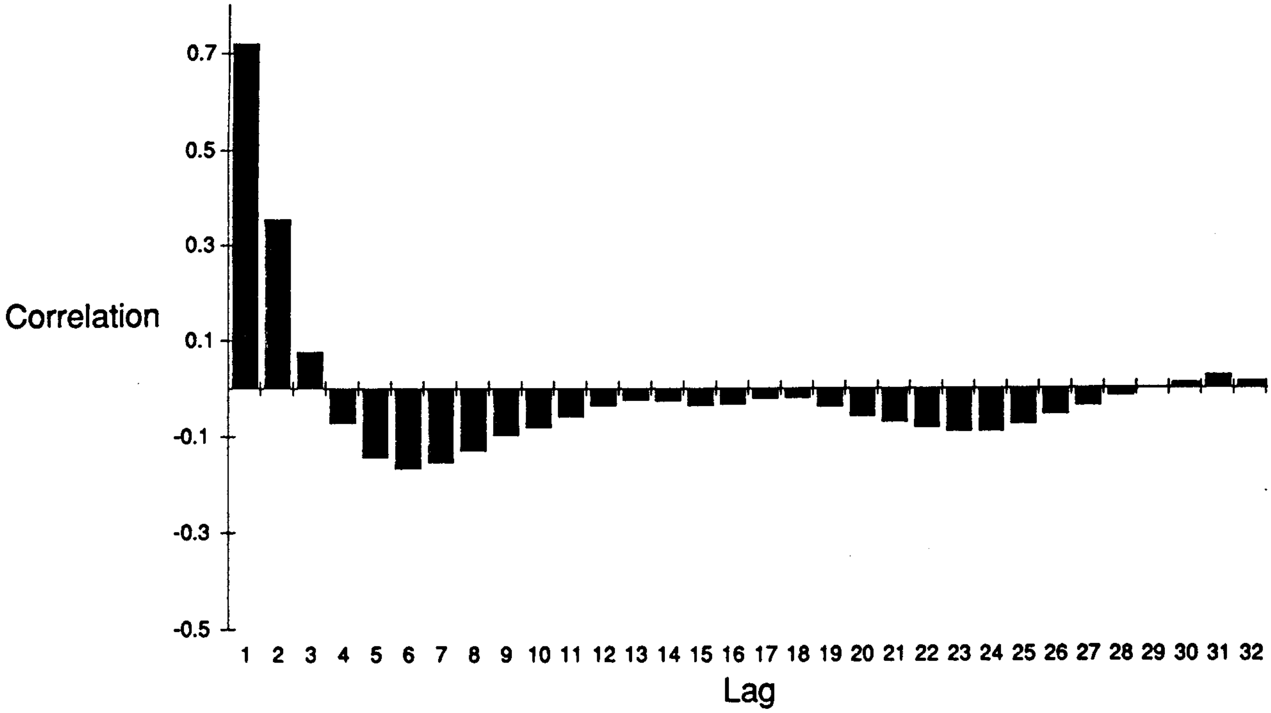

Figure 6. Autocorrelation function for PLANT production--Experiment One.

A smaller but equally consistent positive autocorrelation was observed for number of open valves. (See Figure 7.) The strongest relationship was also noted with a lag of one, and the reduction of the relationship with a lag of two was even greater than in the case of production.

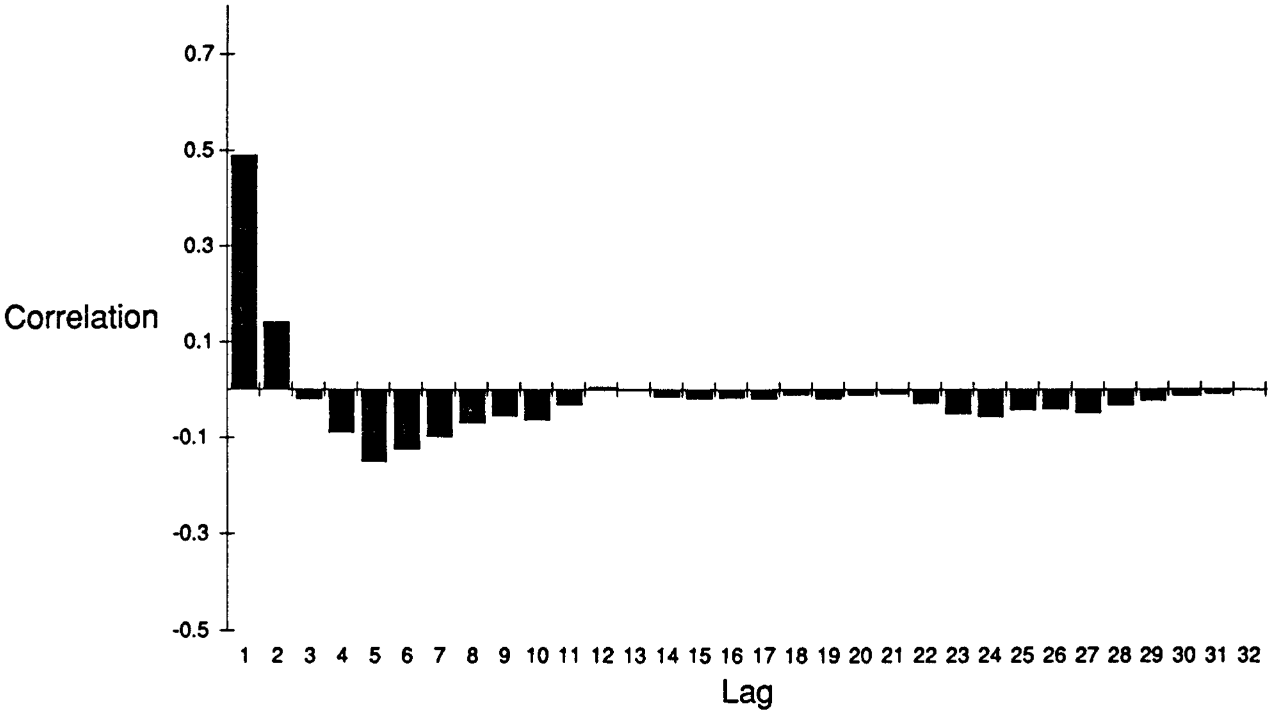

Figure 7. Autocorrelation function for number of open valves--Experiment One.

Significant positive autocorrelations of effort ratings were found in 75% of the analyses. As with production and number of open valves, the strongest relationship was noted with a lag of one. (This is indicated in Figure 8.) However, the largest autocorrelation was not as great as those observed for production and number of open valves, and the time span for the relationship was longer.

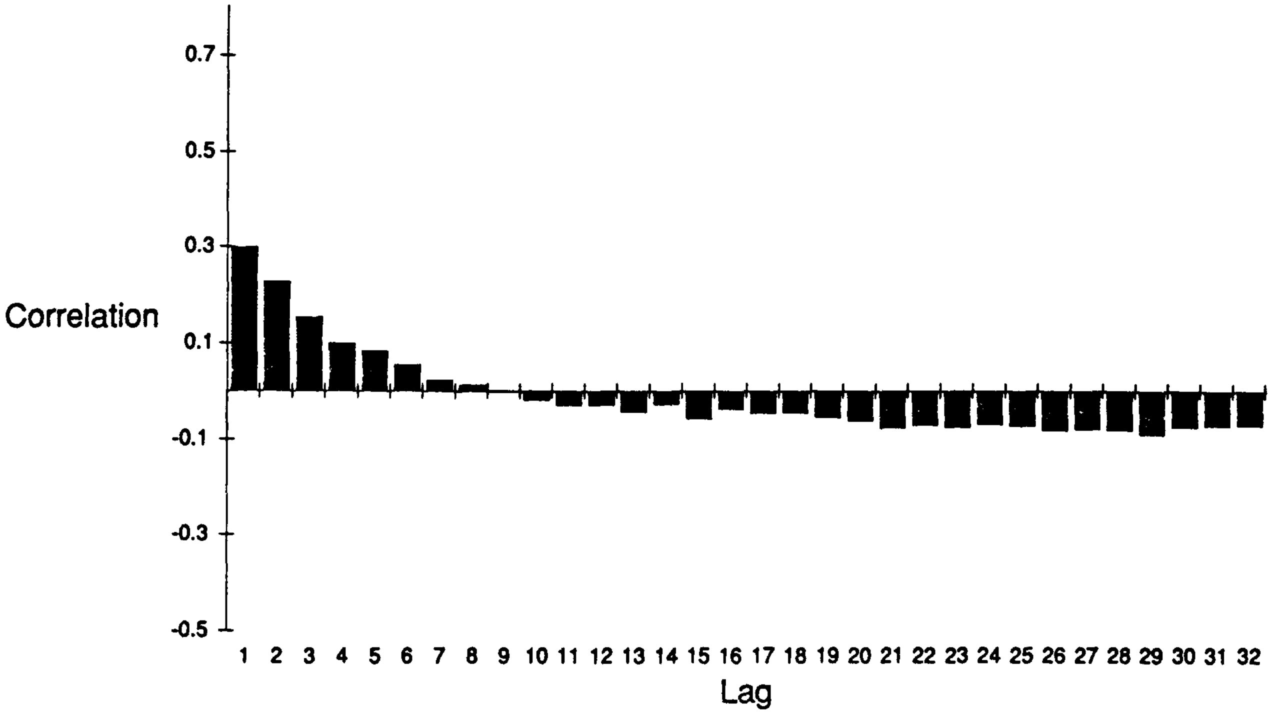

Figure 8. Autocorrelation function for effort ratings--Experiment One.

Autocorrelations of errors were consistently low and non-significant.

Cross-correlations. Consistent relationships were observed with only two of the possible pairs of variables. First, as shown in Figure 9, the number of open valves was positively related to production. As might be expected, this was true only for positive lags (i.e., increased number of open valves preceded increased production). The strongest relationship was noted with a lag of one, but there were also noteworthy relationships with lags of two and three.

Referring to Figure 10, negative cross-correlations were observed between effort ratings and production in 44 of the original analyses. In all of these cases, increased effort ratings preceded decreased production. The strongest relationship was usually observed with lags of one or two, and this relationship often persisted through lags of three or four.

As a note, no consistent relationships between effort ratings and errors were noted. Significant correlations were observed in only 18 of the 48 analyses. Of these, however, 17 of the correlations were positive, with a mean of .348. Eight of the significant correlations involved changes in effort preceding changes in errors, with a mean lag of 1.6. In ten of the cases, changes in error preceded changes in effort, with a mean lag of 0.8.

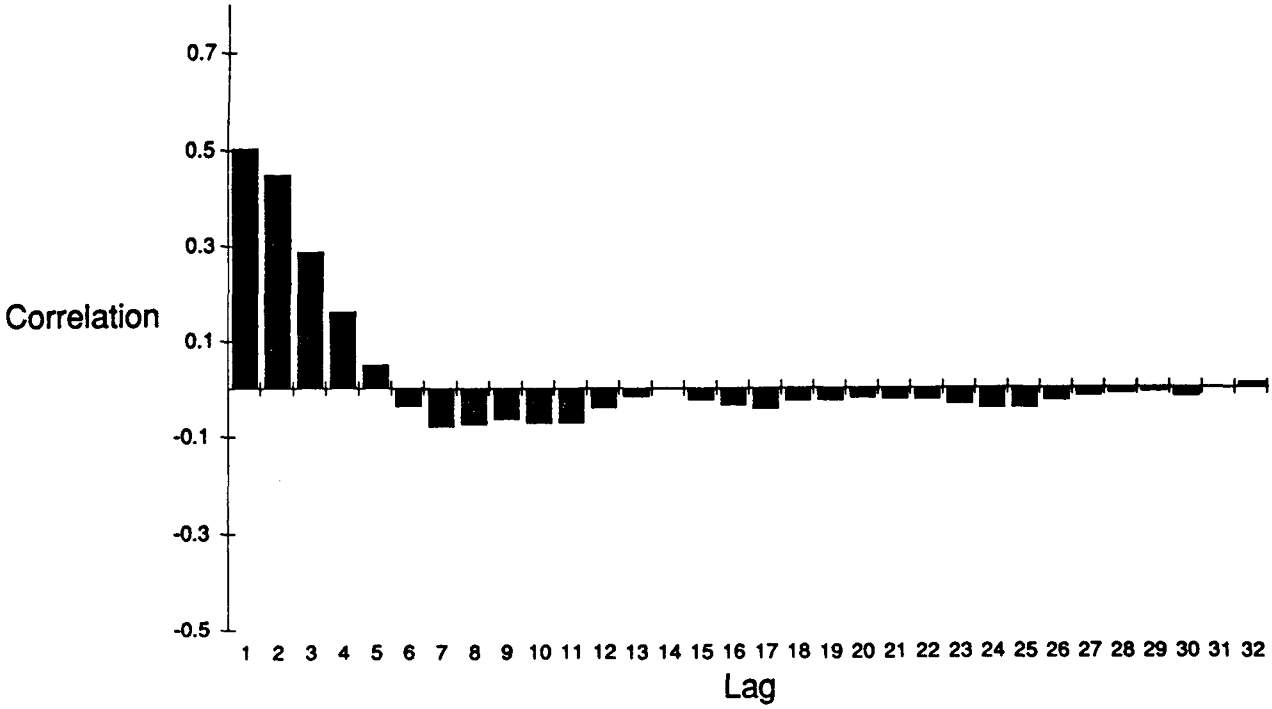

Figure 9. Cross-correlation function for number of open valves x production--Experiment One.

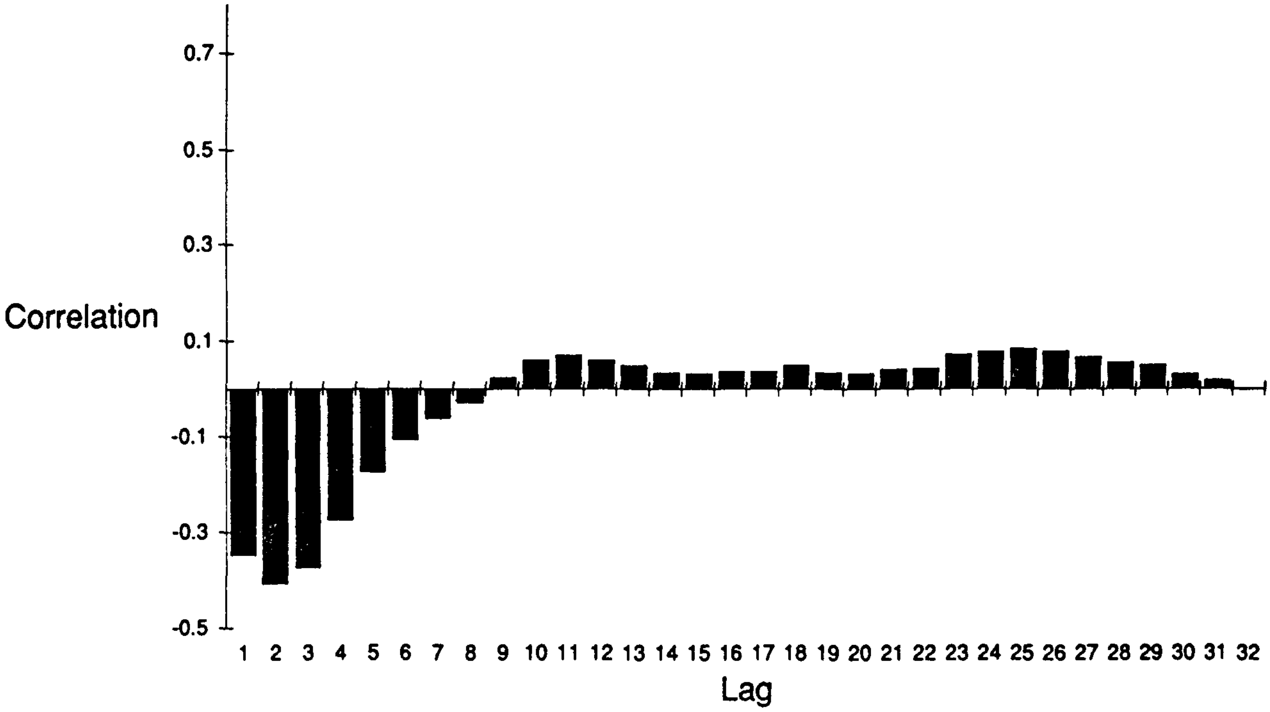

Figure 10. Cross-correlation function for effort ratings x production--Experiment One.

DISCUSSION

The following discussion seeks to provide an interpretation of the results that is consistent with trends observed in the data. As noted with the presentation of statistical analyses, a rather complicated pattern of results was obtained. In order to develop a coherent explanation of the findings, some elaboration of the results was required. Thus, although this discussion is firmly based on the data obtained, not all statements refer to statistically significant relationships. Further, the picture painted here does not take into account every significant effect noted and is, therefore, not complete.

Effects of Experimental Conditions

First, consider the effects the experimental manipulations had upon subjects' behavior. Regarding the effects of the incompatible keyboard layout, the absence of a significant main effect of Compatibility indicates that this manipulation did not pose a great problem for the subjects. Apparently, subjects had a few problems at the beginning of incompatible production runs (as suggested by the higher incidence of errors during the first ten iterations), but they adapted to the incompatible keyboard fairly quickly. Judging from the reduction in commands issued, one way in which they adapted was to avoid using the keyboard. The incompatible condition did have some detrimental effects, as indicated by the interactions with other experimental factors and the increase in errors over compatible conditions, but these effects were overshadowed by the impacts of the other manipulations.

In contrast, the problems created by Pacing manipulations were more evident, as illustrated by higher frequencies of syntax errors. Thus, the hypothesis that increased imposed load would be associated with more errors was supported. In light of the finding that self-paced runs were completed in less time than forced-paced runs, the effects of Pacing were due to the forced-paced nature of the task rather than to a reduction in time allowed. The apparent response of subjects to forced pacing was to become more conservative. In the absence of other manipulations, forced pacing led subjects to scale back on output and hence achieve less production.

The manipulations of Compatibility and Pacing interacted in their effects on subjects' behavior. For example, in the incompatible conditions there was no reduction in output associated with forced pacing. Additionally, forced pacing seemed to enhance the detrimental effects of the incompatible keyboard.

The effects of the Failure Complexity manipulation were more complicated and difficult to interpret, largely due to the interactions of this factor with the other experimental variables. If we consider the effects of the complex failure manipulation in isolation, the apparent impact was to cause subjects to monitor the system more closely, resulting in more closed-loop control. (Recall the lower frequency of INAPP errors, which were largely associated with open-loop control, in the complex failure conditions.) The higher incidence of UNCSYN errors associated with the complex failure, compatible conditions suggests that, while subjects focused more on the state of the system, they paid less attention to the details of entering their actions at the keyboard. As with forced pacing, the complex failure manipulation had a conservative influence upon subjects' specification of output.

When combined with other experimental variables, the complex failure manipulation served to reduce the effects of those variables upon observed performance. Consider the Failure Complexity x Compatibility interaction. There was an apparent increase in INAPP errors associated with incompatible conditions. However, this increase did not occur when the incompatible keyboard was combined with complex failures, presumably because subjects were monitoring the system more closely. Interestingly, the compatibility manipulation seemed to have a reciprocal effect upon errors associated with the complex failure conditions. Apparently the incompatible keyboard caused subjects to focus on the details of command entry enough to avoid the increase in UNCSYN errors found in complex failure, compatible conditions.

The complex failure condition also attenuated the effects of forced pacing. Again, this effect was most evident in the frequency of INAPP errors. Of the two variables, however, Pacing was more potent in its effects upon behavior. The effects of forced pacing were reduced with complex failures, but not eliminated.

Consider now the complex failure, incompatible, forced-paced condition. Prior to collecting data, it was hypothesized that this would be the most difficult condition for subjects. However, the low incidence of errors in this condition (as illustrated in Figure 4) might lead one to believe that this was not the case. It seems unreasonable to conclude that this condition was easier than, for example, either of the simple failure, incompatible conditions, so an alternative interpretation is offered. It is hypothesized that subjects were more careful in this condition, and, thus, avoided making errors. This "greater care" hypothesis is further supported by the high percentage of errors which were self-corrected (CORSYN) in this condition. Interestingly, the reduction in system output which usually accompanied complex failure or forced-paced conditions did not occur when combined with the incompatible keyboard. One possible conjecture is that subjects focused more on details of command entry and events occurring within the system (i.e., failures), and were less aware of the more global system output.

It is curious that no effects of experimental conditions were observed relative to the magnitude of effort ratings. If subjects used greater care (i.e., exercised greater effort) in some conditions than in others, then it might be expected that differences in effort ratings would be found. Reliable differences in the frequency of missing ratings suggest that there were differences in subjective effort, but these were not reflected in rating magnitude. Further discussion of effort ratings is deferred to a later section.

Relationships Between Measures

Now consider the observed relationships among the dependent measures. Autocorrelations and cross-correlations involving production and number of open valves require little interpretation. They simply reflect the orderly dynamics of PLANT, and do not offer insights into human behavior. Therefore, the focus here will be on those relationships involving errors and/or effort ratings.

First, recall that no consistent autocorrelations of error were found. Thus, there is no evidence to support the intuitively appealing notion that errors lead to more errors. It is feasible that the typically low frequencies of error could have made it difficult to find statistically reliable relationships; although nearly 40 errors were noted in the worst condition (simple failure, incompatible, forced-paced), there were many intervals in which no errors occurred. It is also possible that relationships could have been obscured if the ten-iteration sampling interval was too large. However, use of smaller intervals was avoided in light of the low frequencies of error. As a result, little can be said about any time-varying characteristics of error which might exist.

No reliable relationships were found between errors and indices of PLANT performance (production and number of open valves). Most likely, this was because

PLANT is very forgiving. In fact, most of the categories of error considered in this experiment are "trapped" by the interface to PLANT and can have no impact upon the system. Additionally, the measures of production and number of open valves are rather global measures. Thus, the failure to find any reliable relationships between errors and these measures is not difficult to accept.

There were also no consistent relationships between errors and effort ratings. All significant correlations were positive, but significant relationships were not noted reliably across conditions or subjects. This does not appear to be a failure of several marginal relationships to achieve the traditionally accepted significance level of .05; in at least half of the cases, no relationship at all was evident. Among the cases in which significant correlations were noted, increases in effort ratings preceded increased errors in approximately half of them and followed error increases in the other half. Thus, relative to the questions of whether perceived effort serves as a catalyst for error or error contributes to perceived effort, the present results suggest that either relationship is possible.

Recall the other relationships involving effort ratings. Two are of interest here: 1) consistent autocorrelations, and 2) prediction of changes in production. Regarding the autocorrelations, when one considers the factors which might contribute to such a result, a number of interesting possibilities emerge.

For example, subjects' perceptions of effort could be closely related to the current PLANT state, and changes in ratings merely reflect the cyclical nature of PLANT operation associated with less-than-optimal control strategies (i.e., a cyclical pattern of losing and regaining system stability). Alternatively, subjects could have recalled their last rating when deciding what the current rating should be and used the previous rating as an anchor for the current one. Yet another possibility, which is purely conjecture at this point, is that changes in perceived effort do not occur abruptly (at least in benign situations such as those encountered in PLANT), but rather evolve over time. Upon encountering a problem, subjective effort may increase as the problem persists; similarly, subjective effort could dissipate gradually once the problem is resolved.

These possibilities are not mutually exclusive; a given rating could be influenced by any or all of them. Understanding the reasons for the autoregressive nature of the effort ratings noted here will require a greater understanding of the factors on which the effort ratings were based. Unfortunately, the data obtained in this experiment do little to enhance such an understanding.

In light of the fairly consistent negative relationship between effort ratings and production, it may be stated with some confidence that the effort ratings did reflect something relevant to PLANT operation. The discovery that increases in effort ratings foreshadowed decreased production during the next 20-40 iterations was quite interesting. It was also puzzling, however, because there were no other relationships observed which could account for increases in effort ratings. Ratings were not related to experimental condition, frequency of error, or PLANT state (i.e., number of open valves). Solving the puzzle is difficult without extrapolation. A hypothesis is offered, but first it is necessary to step back and consider the overall pattern of results obtained.

A Broader Perspective

As noted near the beginning of this report, this research was conducted with two goals in mind: 1) investigating the causes of human error, and 2) investigating the relationships between error and mental workload. As often happens in research, several outcomes of this experiment were not as expected. Manipulating imposed load via pacing resulted in more errors, as predicted; however, the other experimental manipulations failed to produce the anticipated effects, and the resulting data is difficult to interpret. Nevertheless, if we examine the manner in which subjects responded in this experiment, the results are informative.

It was hypothesized that humans would be more likely to make errors under certain conditions. As an evaluation of this idea, we created those conditions, placed people in them, and waited for the human subjects to provide a demonstration of our hypothesis. However, what we received was not verification of our predictions, but rather a demonstration of human adaptability. If the interpretation of results offered thus far is accepted, then the subjects realized they were in situations in which errors were likely, and took steps to compensate for those situations. In incompatible conditions, they reduced their use of the incompatible keyboard; in complex failure and forced-paced conditions, they scaled back on system output (which would have the effect of making PLANT more stable). In short, the subjects' response to the error-likely conditions was to try to render them less error-likely.

The possibility that subjects compensated for troublesome conditions by scaling back also sheds new light on the relationship between effort ratings and PLANT production and the puzzling absence of effects due to experimental conditions. The negative correlation between effort ratings and PLANT production indicates that subjects responded to increased subjective effort by reducing system output. As noted, this would result in a more stable PLANT, and if subjective effort was affected by PLANT state, a lower level of perceived effort.

Hence, to a certain extent subjects could regulate their level of subjective effort by altering their control strategies. Note that apparently compensatory control strategies occurred in the complex failure and forced-pacing conditions--those conditions in which increased subjective effort was expected. If subjects were able to reduce subjective effort in those conditions, this could explain the failure to note differences in effort ratings associated with experimental manipulations.

Similar arguments could be use to explain the absence of relationships between effort ratings and other measures. For example, suppose that subjective effort and error are in fact positively related, and that increases in effort precede increased error. By adopting a compensatory strategy, a subject could avoid committing errors by responding to increased perceived effort as a signal to scale back before the errors occurred. In the resulting data, no relationship between effort and errors would be apparent.

EXPERIMENT TWO

In light of the results of the first experiment, the second experiment was conducted with two goals in mind. First, an effort was to be made to curtail subjects' adaptation to experimental conditions by changing PLANT control strategies (primarily, scaling back on production). It was hoped that a production bonus would help achieve this objective. The second goal was to gain greater understanding of the factors leading to increases in perceived effort and the effects of these perceptions on subsequent behavior. The approach used was in-depth questioning of subjects as they observed "instant replays" of their behavior in controlling PLANT. As manipulation of display-keyboard compatibility appeared to have little impact upon behavior in the first experiment, this manipulation was omitted.

Subjects

Subjects in this experiment were the four persons whose data were analyzed in the previous experiment. During the one-year interval since Experiment One was conducted, the three students had received their degrees from the technical school, and all were employed in jobs that were at least moderately related to the training they had received. One worked in a data processing department for a small company, one maintained personal computers, and one maintained copy machines. The fourth subject was still employed as an operator of a climate-control system.

Experimental Procedure

Subjects controlled PLANT for a total of 13 production runs. These production runs were a subset of those encountered by subjects during Experiment One. Although all of the problems seen had been presented earlier, subjects gave no indication that the production runs were familiar to them. Since the production runs differed only in the location and timing of generic failures (e.g., failure of pump A at time 73 vs. failure of pump C at time 84), it is unlikely that the specifics of a given run could have been recalled from the year before.

The first seven sessions were training, in which relevant features of PLANT (i.e., the basics of PLANT operation, operational procedures, and simple and complex PLANT failures) were successively reintroduced. The effort-rating procedure was also reintroduced during training. Subjects were told at the beginning of the experiment that they would be asked to explain their actions in detail at a later time.

At the beginning of the seventh production run (the last training session), subjects were informed that a production bonus would be awarded for each of the last six production runs (8-13). It was explained that they were being told this in advance so that, if they wanted to change strategies as a result of this information, they could try out the new strategy before it counted. The precise calculation of the bonus was not discussed; however, each subject was told that he/she would definitely get some money for each run, and that the amount would depend on how much he/she produced and how much the other subjects produced. It was further suggested that the total amount of bonus money received by each subject could be "anywhere from $5 to $50." (In actuality, each subject received approximately $30.)

At the end of the seventh production run, the technique for eliciting subjects' explanations of their behavior was introduced. First, the experimenter generated a hard-copy plot of the effort ratings obtained over the course of the production run, and showed the plot to the subject. The subject was told that the experimenter was interested in "why they did what they did," and "what they were thinking and feeling" as they controlled PLANT. It was then explained that the production run would be replayed as the subject watched, and the subject was to "talk through" what was happening. Explanations were to include, but not be limited to, the reasons for effort ratings being what they were. The production run was replayed in self-paced mode as the subject watched, and the ensuing discussion between the subject and experimenter was taped.

Experimental manipulations were presented in the last six production runs. The two experimental factors were Pacing (self- vs. forced-paced) and Failure Complexity (simple failures only vs. both simple and complex failures). Subjects controlled PLANT twice under each of the simple failure conditions (simple failure, self-paced and simple failure, forced-paced), and once under each of the complex failure conditions. Experimental production runs were presented pseudorandomly, as in Experiment One.

Dependent measures were the same as for the first experiment: performance measures, errors, effort ratings, and other behavioral measures.

RESULTS

Statistical techniques used to analyze the data in Experiment Two were similar to those used in Experiment One. The effects of experimental conditions were evaluated with two-way analysis of variance with repeated measures. Factors in these analyses were Failure Complexity and Pacing. Relationships between measures were explored using time series analysis, in the same manner as described for Experiment One. In the following presentation, effects noted in Experiment Two are reported first, followed by a comparison of the two experiments.

Effects of Experimental Conditions

Performance and other behavioral measures. Two performance measures were significantly affected by Pacing. First, the average level of input specified was higher in the self-paced condition than in the forced-paced condition (117.17 vs. 104.32, F(1,3) = 12.63, p = .038). Second, subjects kept more valves open in the self-paced condition than with forced-pacing (16.57 vs. 15.99, F(1,3) = 79.07, p = .003).

There was a significant main effect of Failure Complexity on trips and on the average variance of tank levels. More trips occurred in the simple failure conditions than the complex failure conditions (238.62 vs. 151.50, F(1,3) = 26.81, p = .014).

The Failure Complexity manipulation had significant effects upon command usage as well. (Command usage across experimental conditions is summarized in Figures 11 and 12.) More "open" and "close" commands were issued in simple failure conditions (150.38

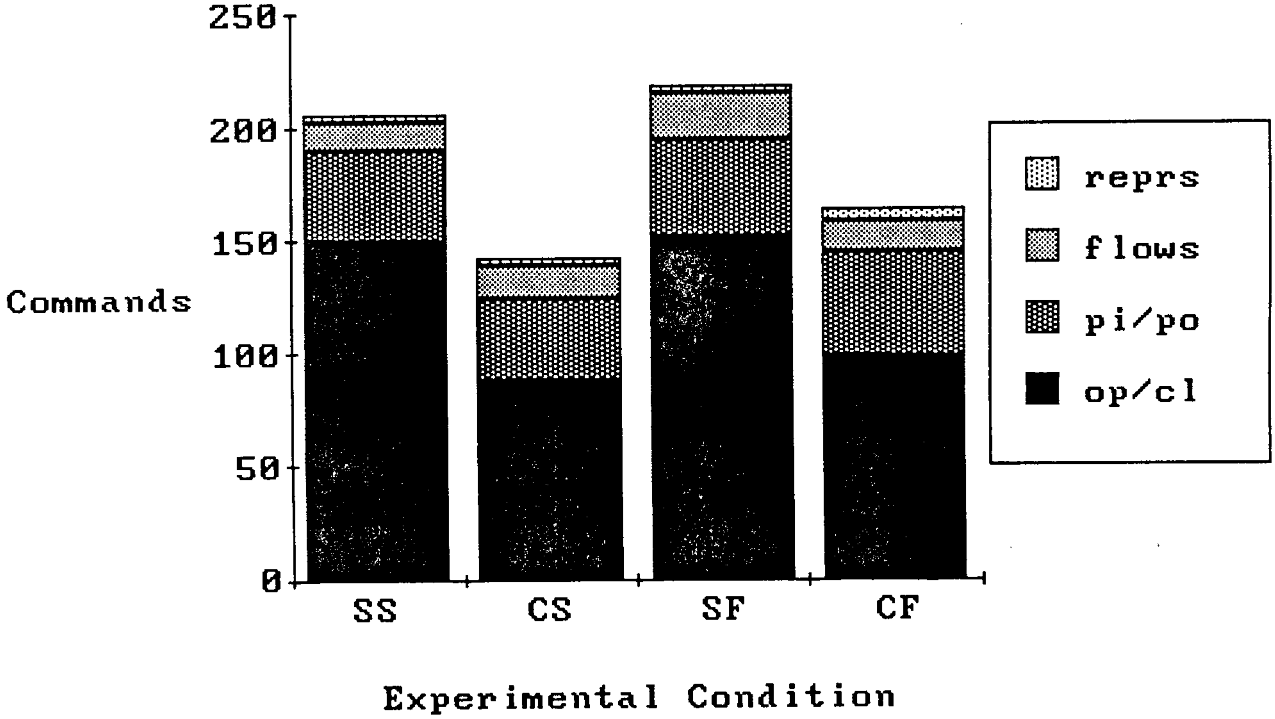

Figure 11. Command usage across experimental conditions--Experiment Two.

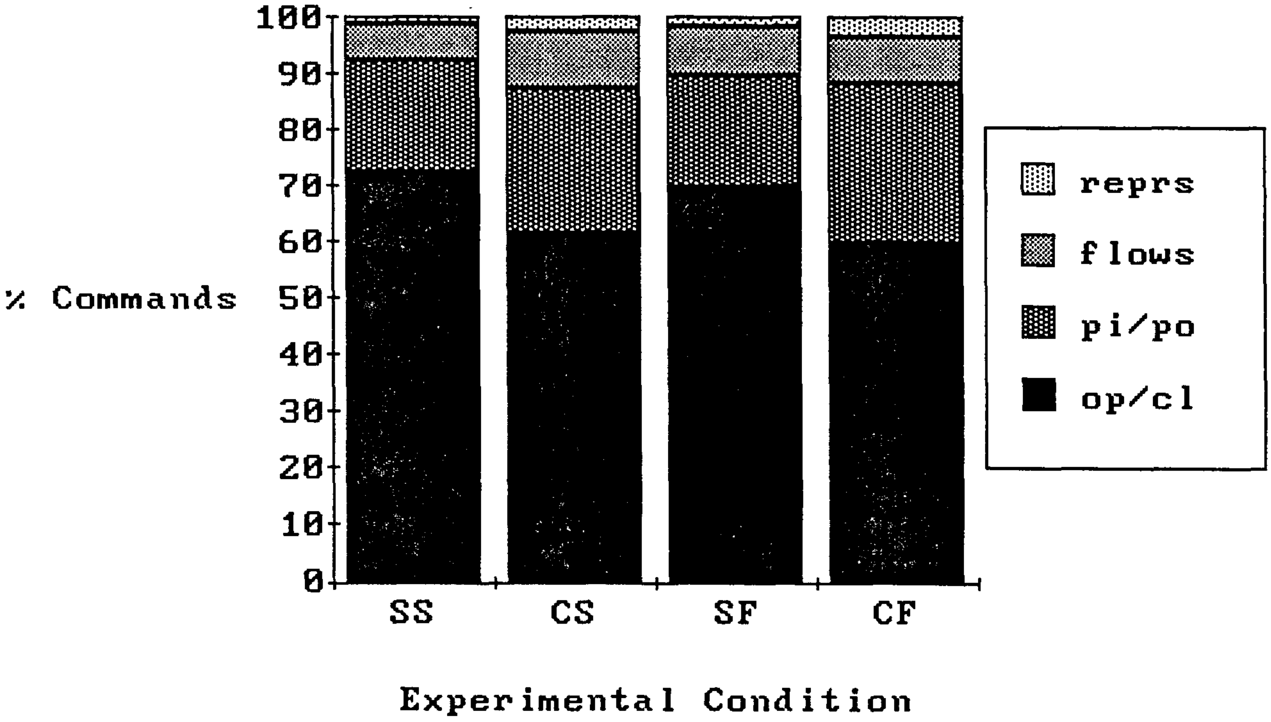

Figure 12. Percent of each type of command used across experimental conditions--Experiment Two.

vs. 92.75 for complex failures, F(1,3) = 22.31, p = .018). There were also more total commands issued in simple failure runs (211.88 vs. 153.38, F(1,3) = 64.88, p = .004). This difference in total commands can be largely attributed to the difference in open and close commands.

As with the first experiment, subjects completed self-paced runs in less time than forced-paced runs (16.9 vs. 23.3 minutes, F(1,3) = 15.19, p = .03).

Error measures. Of the two experimental factors, only Failure Complexity had an effect upon errors. (See Figures 13 and 14 for summaries of errors across experimental conditions.) In the simple failure conditions, subjects made more CORSYN errors (4.25 vs. 1.75 for complex failure conditions, F(1,3) = 10.47, p = .048), more INAPP errors (7.50 vs. 4.88, F(1,3) = 16.46, p = .027), and more total errors (21.12 vs. 13.13, F(1,3) = 14.12, p = .033).

Effort ratings. In contrast to Experiment One, there were no differences in the number of ratings obtained across experimental conditions. However, there was a small but significant effect of Pacing upon the magnitude of ratings. Effort ratings were lower with self-pacing than with forced-pacing (4.015 vs. 4.560, F(1,3) = 10.51, p = .048).

Relationships Between Measures

Investigation of relationships between measures was conducted with time series analysis in the same manner as described for Experiment One. Full analyses were conducted for each of the 24 experimental production runs (4 subjects x 6 sessions), and then averaged correlation functions were created to facilitate interpretation. Those plots

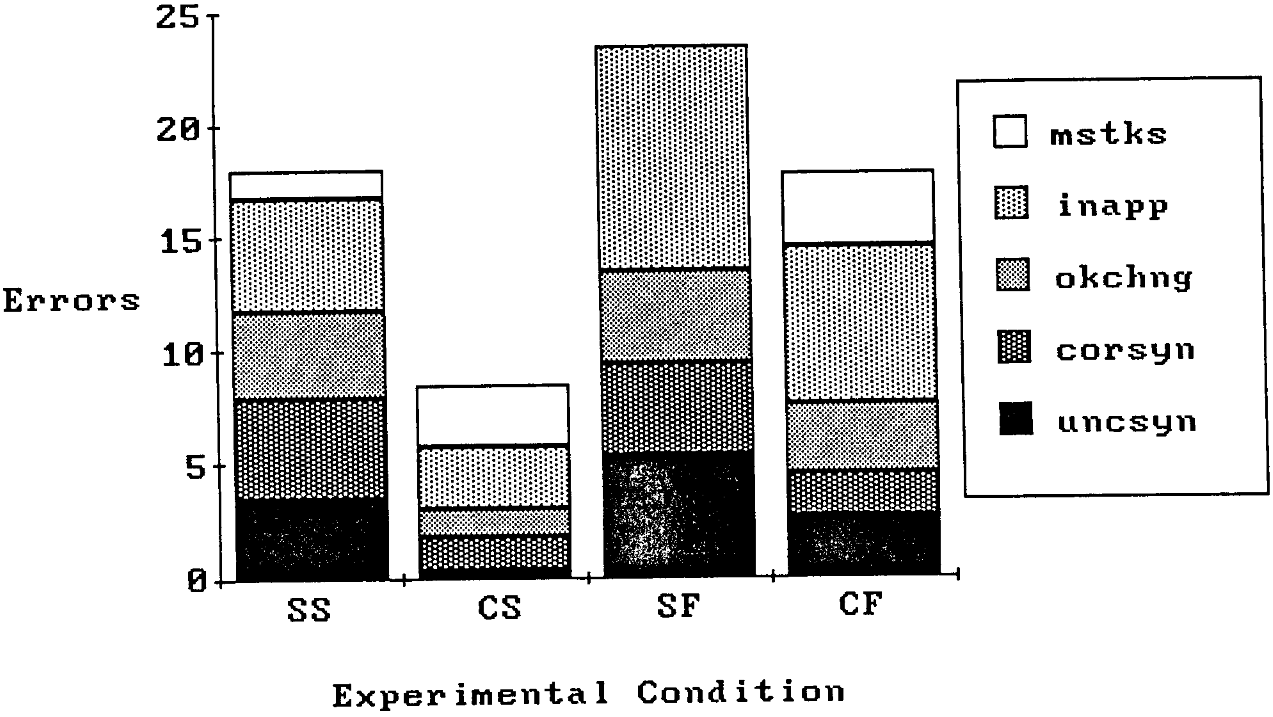

Figure 13. Errors across experimental conditions--Experiment Two.

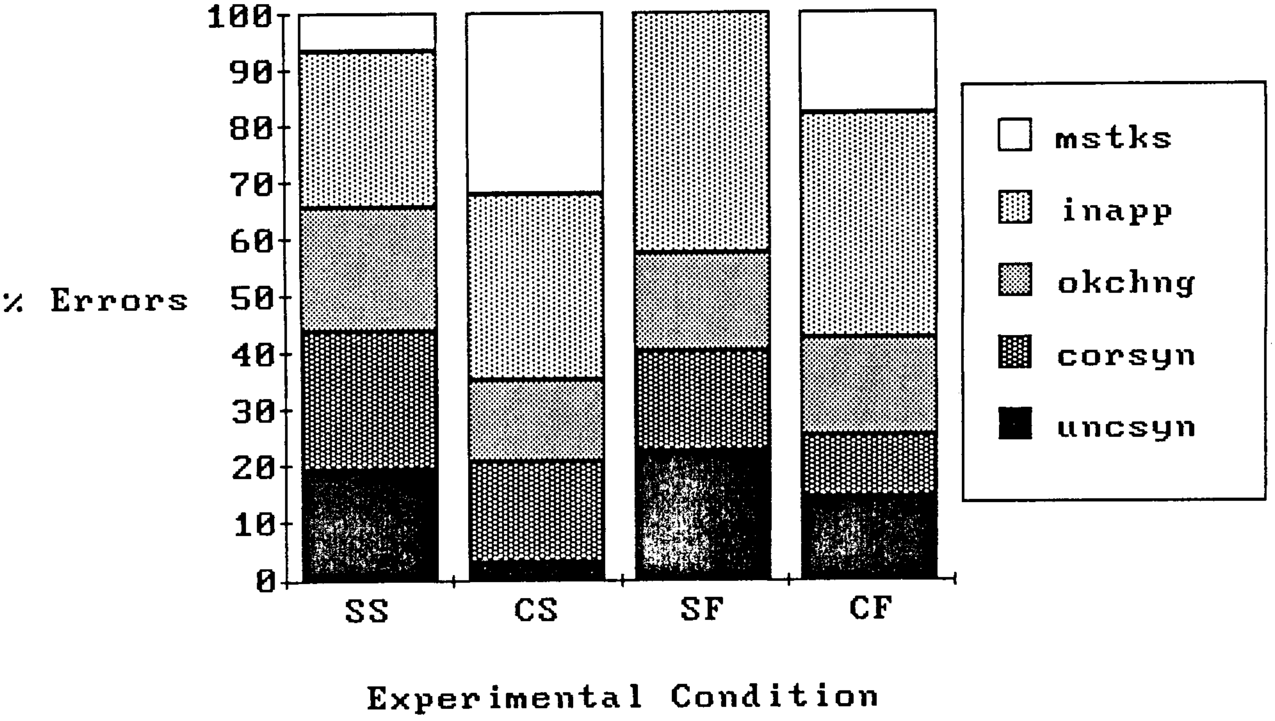

Figure 14. Percent of each type of error occurring across experimental conditions--Experiment Two.

of functions containing correlations greater than ±0.3 are shown in Figures 15 through 18.

Autocorrelations. As illustrated in Figures 15 and 16, strong positive autocorrelations of production and number of open valves were found. As in Experiment One, these relationships were observed consistently in all 24 analyses.

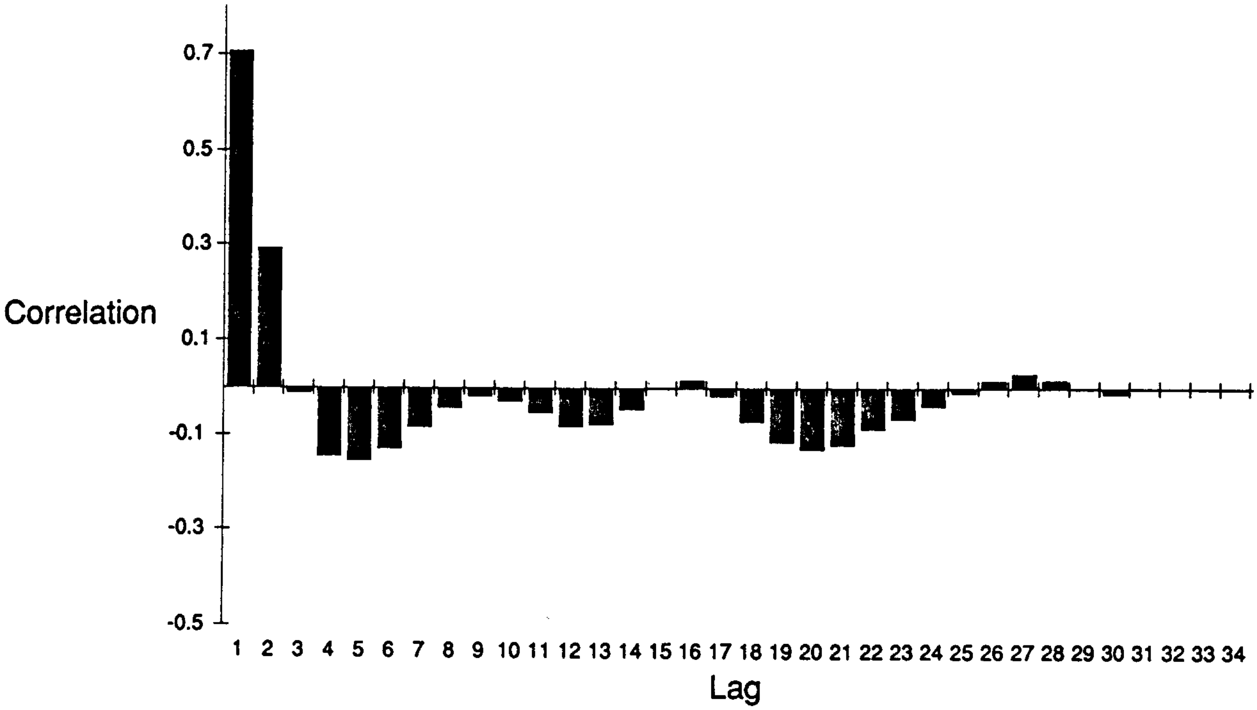

Figure 15. Autocorrelation function for PLANT production--Experiment Two.

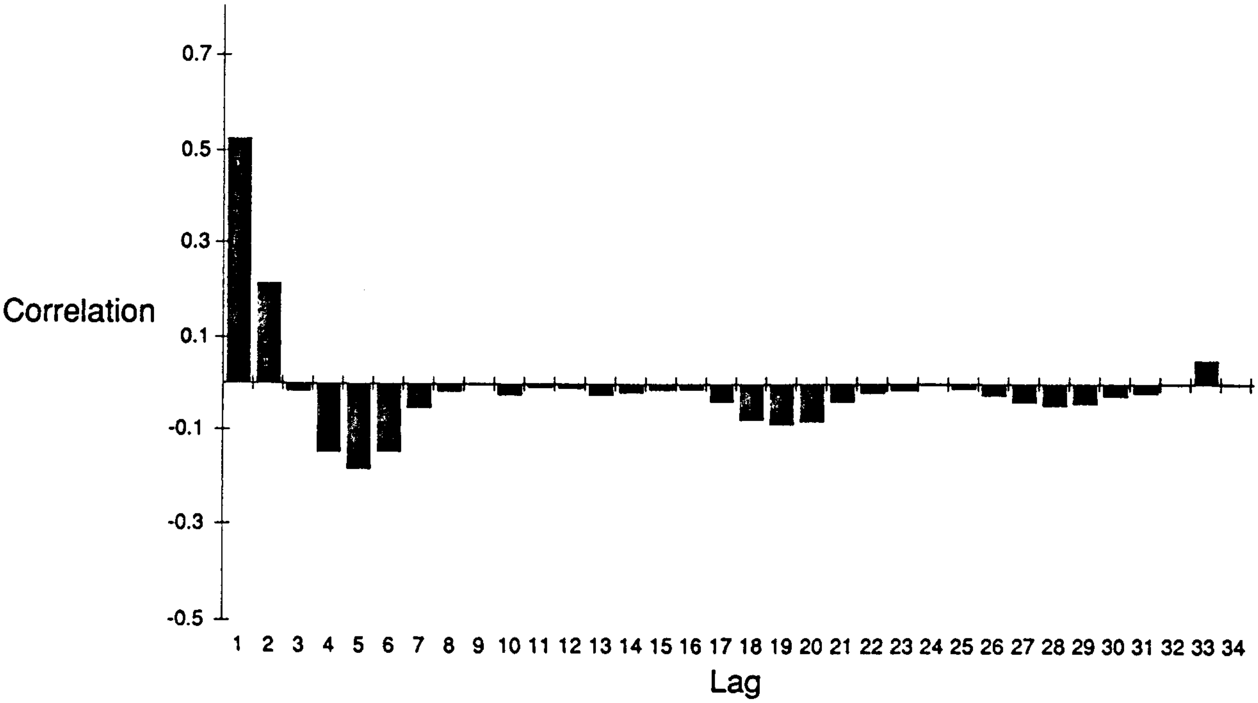

Figure 16. Autocorrelation function for number of open valves--Experiment Two.

Positive autocorrelations of effort ratings were also noted. (See Figure 17.) Also consistent with the results of the first experiment, these relationships were found in 75% of the analyses.

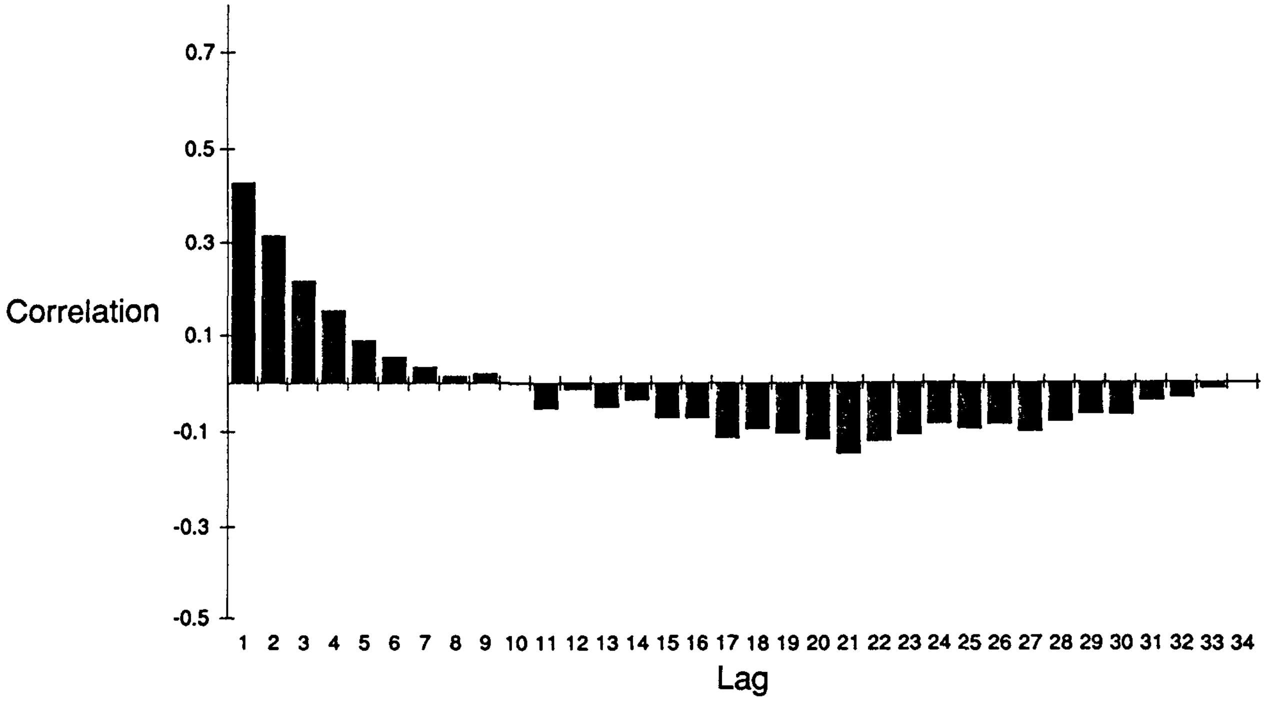

Figure 17. Autocorrelation function for effort ratings--Experiment Two.

Autocorrelations of errors were consistently low and non-significant.

Cross-correlations. The strong positive relationship between number of open valves and subsequent production that was noted in the first experiment was replicated in Experiment Two. This relationship is illustrated in Figure 18.

Negative cross-correlations between effort ratings and production were also found in 79% of the analyses. However, these relationships differed from those noted in Experiment One, in that changes in effort ratings preceded production changes in only 6 of the cases; in 13 analyses, changes in production preceded changes in effort ratings.

Negative cross-correlations were also noted in 79% of the analyses between effort ratings and number of open valves. As with the effort-production relationships, changes in one measure did not consistently precede changes in the other. Changes in effort preceded changes in production in 8 of the 19 cases in which the relationship was observed.

Significant cross-correlations between effort ratings and errors were noted in only 6 of the 24 analyses.

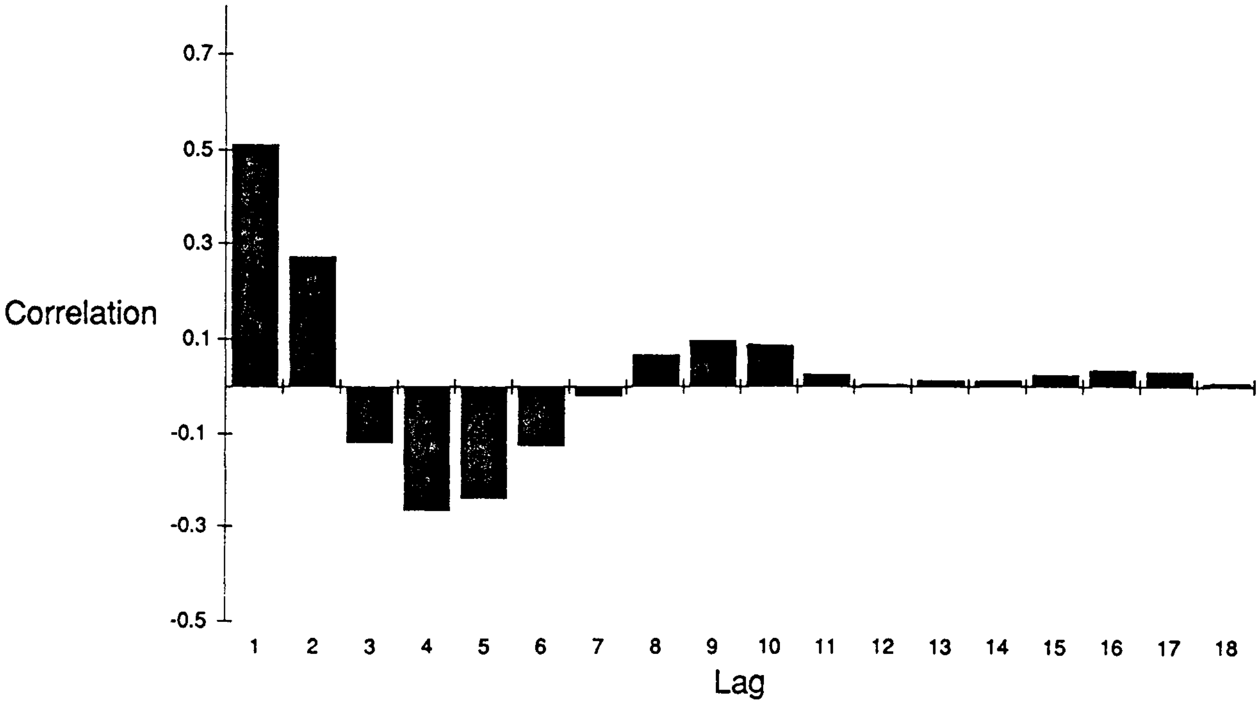

Figure 18. Cross-correlation for number of open valves x production--Experiment Two.

Differences Between Experiments

In order to make direct comparisons between the two experiments, all dependent measures were evaluated with analyses of variance in which Experiment was included as a factor. More specifically, three-way analyses with repeated measures were performed using the factors of Experiment, Failure Complexity, and Pacing. Data from the incompatible conditions in the first experiment were excluded from this analysis. The Experiment factor was involved in the following significant effects.

Performance and other behavioral measures. There was a significant interaction of Experiment and Failure Complexity in their effects on production (F(1,3) = 16.01, *p* = .028). In Experiment One, subjects produced significantly less in the complex failure condition (95,218 units) than in the simple failure condition (115,075 units). Production in both Failure Complexity conditions in Experiment Two was equal to the simple failure condition in Experiment One (117,192 units).

The interaction of Experiment and Failure Complexity was also significant for valve trips (F(1,3) = 12.63, *p* = .038). There was no significant difference in trips associated with Failure Complexity in Experiment One (146.13 vs. 124.38 for simple and complex failure conditions, respectively). However, there were significantly more trips in the simple failure condition (238.38) than in the complex failure condition (151.50) in Experiment Two.

Regarding the number of open valves, the interaction of Experiment and Failure Complexity was once again significant. Subjects kept fewer valves open in the simple failure condition in Experiment Two than in the other three combinations of Failure Complexity and Experiment (15.95 vs. a mean of 16.36, F(1,3) = 11.68, *p* = .042).

Subjects' choices of input level were also affected by the interaction of Experiment and Failure Complexity (F(1,3) = 15.58, *p* = .029). In Experiment One, input was higher in the simple failure condition than in the complex failure condition (105.29 vs. 94.53). Average input in the two Failure Complexity conditions in Experiment Two did not differ from the simple failure condition in Experiment One (108.10 and 113.38 for the simple and complex failure conditions, respectively).[1]

Average output was affected in the same manner as input. Output was significantly higher in the simple failure condition (105.49) than in the complex failure condition (95.86) in Experiment One, but no differences were observed in Experiment Two (107.22 vs. 119.00, F(1,3) = 12.89, *p* = .037)

There was a significant interaction effect of Experiment and Failure Complexity upon the total number of commands issued (F(1,3) = 14.12, *p* = .033). In Experiment One, there was no difference in total commands associated with Failure Complexity (170.04 vs. 150.75 for simple and complex failure conditions, respectively). In Experiment Two, however, more commands were issued in the simple failure condition (211.88) than in the complex failure condition (153.38). As noted in the presentation of results from Experiment Two, this difference appeared to be influenced largely by the number of "open valve" commands.

Error measures. No significant effects of Experiment upon the specific error measures (i.e., UNCSYN, CORSYN, INAPP, OKCHNG, and MISTAKES) were observed. However, there was a significant interaction effect of Experiment and Failure Complexity upon total errors (F(1,3) = 34.27, *p* = .010). Subjects made fewer errors in the complex failure condition in Experiment Two (13.13) than with the other three combinations of Experiment and Failure Complexity (21.44, 22.25, and 21.12).

Effort ratings. There was a significant main effect of Experiment upon effort ratings. Effort ratings obtained in Experiment One were higher than those obtained in Experiment Two (5.36 vs. 4.29, F(1,3) = 55.57, *p* = .005).

Analysis of Post-Session Interviews

The recorded discussions with subjects were treated as data and analyzed as follows. First, plots of subjects' effort ratings over time were examined to identify points at which ratings had changed (i.e., gone up or down one at least one point). Transcripts of the discussions were then reviewed to ascertain the reasons subjects had given for changing their ratings. Occasionally, there were multiple reasons for a particular change. In such a case, all reasons were noted. In a few cases (approximately 2%), subjects explained why their ratings had *not* been affected by situations that normally would have caused a change in rating. Reasons that a rating did not go down were recorded in the "reasons for effort ratings going up" category, and reasons that a rating stayed lower than normal for a given situation were classified as "reasons for effort ratings going down."

----

[1] It may be recalled that the measures of production, trips, and input were not found to be significantly affected by Failure Complexity in the original analyses of data from Experiment One. A reexamination of that data indicates that there were differences for each of these measures consistent with the results reported here, but the differences failed to reach significance in Experiment One due to inter- and intra-subject variability.

A total of 354 reasons for changes in effort ratings were offered by subjects, which were classified into 15 categories by experimenter judgment. In an effort to avoid altering subjects' meaning and forcing responses into categories that were inappropriate, most of the classification process was accomplished using the verbatim statements from subjects. Category labels were chosen as the last step. The resulting frequencies in each category are presented in Table 1. A brief explanation of each category listed in Table 1 is provided in Table 2.

Table 1

Reasons for Changes in Effort Ratings

REASONS FOR EFFORT RATINGS GOING DOWN	
Acceptable situation	
problem over	61
no problem	2
strange but manageable	6
total acceptable situation	69
Acceptable plan for recovery	35
Acceptable explanation	13
Miscellaneous	3
TOTAL REASONS FOR RATINGS GOING DOWN	120
REASONS FOR EFFORT RATINGS GOING UP	
Execution difficulties	
errors	16
unacceptable system state	62
total execution difficulties	78
Inadequate explanation	75
Operating on edge	29
No acceptable plan for recovery	
goal conflict	18
no plan	7
total no acceptable plan	25
Psychological inertia	9
Surprise	8
Miscellaneous	11
TOTAL REASONS FOR RATINGS GOING UP	234

Table 2

Elaboration of Reasons for changes in Effort Ratings

CATEGORY	ELABORATION
Acceptable situation	The current situation is acceptable for one of the following reasons.
problem over	There was a problem, but it is no longer present.
no problem	The earlier perception of a problem was inaccurate.
strange but manageable	Although an acceptable explanation for current PLANT behavior is not available, the situation is controllable.
Acceptable plan for recovery	Although the current situation is not acceptable, the course of action to be taken to rectify it is known.
Acceptable explanation	Events which were puzzling at first are now understandable.
Execution difficulties	There are problems in controlling PLANT for one of the following reasons.
errors	One or more earlier actions were inappropriate.
unacceptable system state	Control of PLANT is difficult due to the current unstable PLANT state (as manifest by frequent valve trips).
Inadequate explanation	Some current events in PLANT are not understood.
Operating on edge	Although the current situation is acceptable, the situation would deteriorate rapidly if something went wrong.
No acceptable plan for recovery	The current situation is unacceptable and there is no plan for recovery, due to one of the following reasons.
goal conflict	More than one action is appropriate in the current situation, but only one action may be taken at a time.
no plan	It is not clear what actions would be appropriate in the current situation.
Psychological inertia	The current rating is high because the subjective impact of an earlier bad experience has not worn off.
Surprise	Development of the current unacceptable situation was not noticed.
Miscellaneous	The reason does not fit into one of the above categories (e.g., "I don't know.").

DISCUSSION

As with the discussion of the results of Experiment One, an attempt will be made here to summarize the results of Experiment Two in a coherent manner. Commonalties and differences in the two experiments will also be considered. Rather than discuss each statistically significant effect obtained, the approach will be to synthesize an overall perspective of how subjects responded.

Effects of Experimental Conditions

The effects of the experimental factors in Experiment Two may be summarized as follows. Regarding the effects of Pacing, forced-pacing resulted in lower input, fewer valves open, and higher effort ratings than did self-pacing. The Failure Complexity manipulation led to differences in trips and variance in tank levels. However, recall that performance was worse (i.e., there were more trips and larger deviations in tank levels) with the *simple failures* condition. In contrast to Experiment One, there were no significant interaction effects of the two factors.

The reasons that the variables mentioned were affected by one experimental factor and not the other are not clear. Therefore, no attempt will be made here to distinguish the effects of Pacing and Failure Complexity. It is instructive, however, to look at the overall pattern of effects obtained.

Consider the six performance measures affected by the experimental manipulations in the two experiments: 1) production, 2) average output, 3) average input, 4) trips, 5) number of open valves, and 6) variance in tank levels. Only two of these were significantly affected by the manipulations in Experiment One: production and average output. In contrast, production and average output were not significantly affected in Experiment Two, but all of the others were.

This pattern of results suggests the following interpretation. In both experiments, subjects were told that the primary goal of PLANT operation was to produce. Recall that the importance of this goal was emphasized with a production bonus in Experiment Two. It appears that the promise of a bonus led to a shift in focus across the two experiments. In Experiment One, subjects concentrated on maintaining a stable PLANT and responding to problems within the system. One way that they maintained stability was to scale back on output when it appeared that there might be problems. In Experiment Two, however, they "protected" production and did not use the strategy of reducing output to avoid problems. The significant effects of the experimental manipulations upon the other performance measures reveal that, although production was maintained, stability was sacrificed.

At first glance, it would appear that subjects made a tradeoff between production and stability, opting for stability in Experiment One and production in Experiment Two. It is important to note, however, that the dynamics of PLANT do not require such a tradeoff at the production levels achieved by subjects. The differences observed reflected the subjects' choices of strategy rather than PLANT dynamics.

This assertion is supported with a closer examination of the results. Subjects did not scale back on production in the complex failure condition in Experiment Two as they had in Experiment One; however, stability measures in the complex failure condition were not affected. Rather, system instability was greater in the *simple failure* condition in Experiment Two, with no corresponding increase in production.

An alternative interpretation is required. It is hypothesized that subjects interpreted the added emphasis on production as "permission" to allow more instability than they had in the past. This could account for the sizable increase in trips in the simple failure condition in Experiment Two. Consistent with the earlier interpretation of Experiment One, however, they appeared to be more "careful" in problematic situations (i.e., the complex failure condition). Thus, they maintained system stability and committed fewer errors in the complex failure condition.

Another effect of experimental conditions which must be addressed here is the discovery that effort ratings were lower in Experiment Two than in Experiment One. One interpretation is that the subjects were more experienced in Experiment Two and generally less affected by the problems of PLANT control. Further, the activities associated with participating in an experiment were no longer new to them, and they were, perhaps, more confident and less anxious to please the experimenter.

Although this interpretation is certainly plausible and descriptive of subjects' apparent attitudes toward the experimental situation, the comments of subjects during the course of the experiment suggest that another less obvious "force" was operating as well. Consider again the interpretation that there was a shift in subjects' focus from stability to production across the two experiments, and the further possibility that subjects viewed system stability as less important in Experiment Two. Such a change in focus and attitude may have allowed the subjects to simplify the problem of PLANT control.

If stability is an important goal independent of production, then valve trips or deviations from the "desired" tank level (which occur frequently in PLANT) are noteworthy events to be avoided for the sake of stability. If, on the other hand, the only important goal is production, instability is a problem only when it interferes with production. It is suggested that subjects adopted the latter position and experienced less effort as a result. This point is best illustrated by the comments of one subject as she observed a replay of one of her sessions during a post-session interview. "Just look at that system! It was in pretty bad shape then, but I didn't care. Look how much I was producing!"

The explanations offered thus far appear plausible when the effects are considered in isolation. They are still acceptable when these effects are considered as a group, but none of the explanations is sufficient to account for all of the results. Specifically, recall the following effects:

1. More production was achieved in the complex failure condition in Experiment Two than in Experiment One.

2. Fewer errors were made in the complex failure condition in Experiment Two than in Experiment One.

3. Effort ratings were lower in Experiment Two than in Experiment One, and there was no apparent increase associated with the complex failure condition.

These results present a complex picture that is not readily understood. Providing an adequate explanation of what happened requires a great deal of extrapolation. The following discourse is offered as a hypothesis.

As discussed earlier, subjects in Experiment Two appeared to focus more on production and less on trips. The emphasis on production resulted in more production, and the lowered concern for trips led to more trips, fewer open valves, and (less directly) lower effort ratings. In the complex failure condition, however, subjects were more careful. The presence of a problem that was difficult to diagnose, and, perhaps, the fact that they had allowed so many trips in the simple failure condition, may have prompted subjects to "straighten up." This resulted in fewer errors in the complex failure condition.

The most puzzling aspect of these results is the lack of a relationship between errors and effort. If subjects were more careful in the complex failure condition, then why weren't effort ratings higher in that condition? Recall that no relationship between these two measures was found in Experiment One either.

An answer may be found by examining the reasons subjects gave for changes in effort ratings. As will be discussed later, execution difficulties were associated with only a third of the reasons offered by subjects; yet, most of the errors noted were errors of execution. If execution difficulties did not contribute to most of the increases in subjective effort, then it is not surprising that the two measures were not correlated.

Relationships Between Measures

Most of the relationships observed in Experiment Two were consistent with those in Experiment One. Autocorrelations and cross-correlations of production and number of open valves continued to illustrate the orderly dynamics of PLANT. Once again, no relationships involving errors were found.

Three relationships require further discussion. First, as in Experiment One, significant autocorrelations of effort ratings were noted. It is curious that the proportion of production runs in which this relationship was observed was 75% in each experiment. No explanation for this is offered; the coincidence is highlighted as a point of comparison for future research.

The second noteworthy relationship is the autocorrelation of effort ratings and production. Recall that changes in effort ratings preceded changes in production in Experiment One, and that this relationship was interpreted as subjects' scaling back on production in response to increased mental effort. The relationship observed in Experiment Two was less consistent, with changes in effort *following* changes in production in 56% of the analyses, and preceding production changes in only 26% of the cases. This change in the effort-production relationship is further evidence of subjects' change in focus in response to the production bonus. When PLANT stability was more important, increased mental effort led to reduced production in the hopes of maintaining control. However, with production as the primary goal, changes in production led to changes in mental effort.

The third relationship to be discussed is the autocorrelation of effort and number of open valves. In Experiment One, no clear relationship between these measures was observed. In Experiment Two, however, changes in the number of open valves preceded changes in effort in 48% of the analyses, and followed effort changes 35% of the time. No strong interpretation of this relationship is offered. It is merely cited as further suggesting that changing the emphasis on production led to the emergence of relationships involving secondary measures.

Post-Session Interviews

Looking at the subjects' explanations for changes in effort ratings summarized in Table 1, a number of observations may be made. The most noticeable characteristic of the data is the fact that there were more reasons for increases in effort ratings than there were for decreases, both in frequency and variety. At least two factors seemed to contribute to the difference in frequency. First, large increases in perceived effort were usually not abrupt; rather, rating magnitude typically increased gradually over time until reaching a peak. On the other hand, recovery (i.e., decrease in perceived effort) seemed to happen more quickly. Second, subjects usually began the session with an extremely low level of perceived effort, which was not reported again after problems had occurred.

Consider now the reasons given for decreases in effort ratings. Not surprisingly, by far the most common explanation offered was that an unacceptable situation had improved. The incidence of the other two explanations (having an acceptable explanation for the situation or an acceptable plan for recovery) is worth noting, however. According to a prescriptive view of problem solving in a variety of control situations, problem solution involves three stages: situation assessment, planning and commitment, and execution and monitoring. The third stage is completed when the problem is solved. Judging from Table 1, perceived effort often decreases upon completion of the third stage (i.e., the situation is again acceptable). These results indicate that a partial solution (i.e., completion of one or two of the stages) can sometimes result in a reduction of perceived effort as well.

The largest frequency of reasons given for increased effort ratings is in the category of "execution difficulties." Due to the varied nature of responses in this category, the high frequency may be misleading. Two distinct types of response were classified as execution difficulty. The first, "errors," could be described as "chagrin" at having committed an error. Thus, in a sense, these responses could be considered a form of "psychological inertia" (which will be discussed in a moment).

The second type of execution difficulty, "unacceptable system state," occurred as a result of problems in controlling PLANT, usually due to valve trips. In the PLANT environment, valve trips can be problematic for two reasons: 1) they are annoying because the valves must be reopened, and 2) they may signify problems. The difference in interpretation is important, because the first is commonly associated with emergence from a problem state (i.e., keep opening valves until stability is regained), whereas the second signals the onset of a problem. It is not possible to distinguish these alternative interpretations of valve trips from subjects' responses.

The single most frequent reason subjects gave for increased effort ratings was "inadequate explanation." Higher ratings were frequently associated with confusion as to why events in PLANT were happening. The third category in the list of reasons for rating increases, "operating on edge," might have been cited less often if there had been no production bonus. Subjects occasionally mentioned in Experiment Two that they felt they were pushing PLANT harder to achieve higher production.

"No acceptable plan for recovery" was offered as an explanation for approximately 10% of the increases in effort ratings. The low frequency in this category relative to difficulties in explanation or execution probably reflects PLANT characteristics rather than inherent differences in the three stages of problem solving. In the PLANT environment, plans for recovery from most unacceptable operational situations are provided in the form of procedures. Thus, much of the subject's task involves identifying which procedure applies (i.e., assessing the situation) and executing the appropriate procedure, with little need for generating plans.

"Psychological inertia," for lack of a better term, refers to carryover effects from previous rating periods. In other words, subjects reported that some ratings were high because they had not "settled down" from earlier experiences. Explanations in this category described two types of inertia: 1) persistence of perceived effort (continued high ratings after a situation had improved), and 2) renewal of perceived effort (because the current situation reminded them of an earlier one).

Occasionally, subjects reported that subjective effort increased as a result of "surprise." Subjects' ratings suggest that the effects of surprise, though infrequent, can be quite potent. The most marked increases in ratings (e.g., jumping from a "2" to a "10") were associated with surprise, and at least two reports of psychological inertia referred to a surprise event.

A great deal of interpretation is not required to synthesize the results presented in Table 1 into a transition diagram such as the one in Figure 19. The predominant force affecting increases in effort ratings in the PLANT environment is the transition from an acceptable to an unacceptable situation. This is illustrated by the large arrow in the center of the figure. The other arrows provide more explicit representations of the reasons offered by subjects. Solid arrows indicate reasons for effort ratings going up, and broken arrows represent reasons for effort ratings going down.

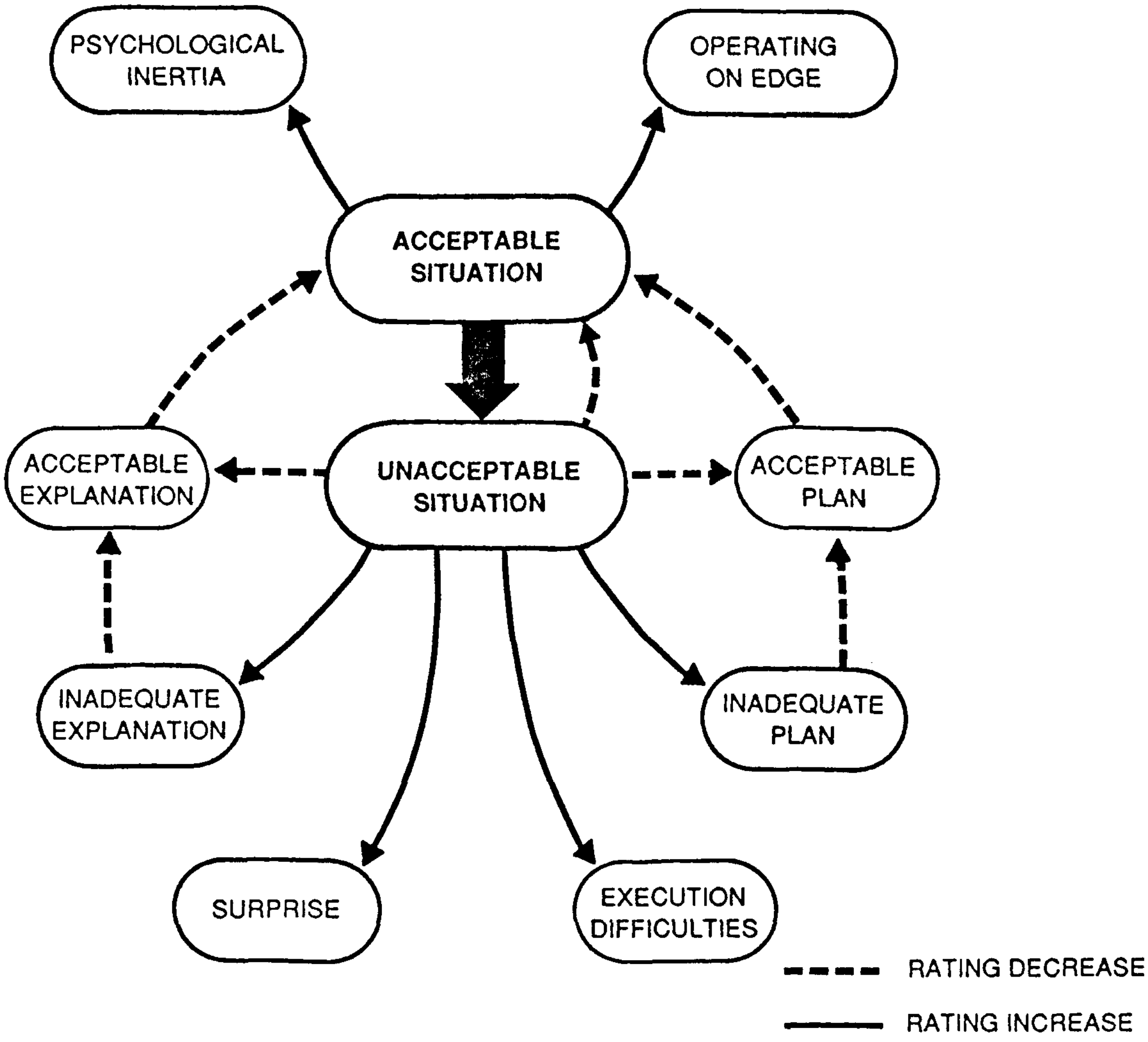

Figure 19. Factors influencing changes in effort ratings, as inferred from subjects' explanations.

CONCLUSIONS

The results of the two experiments reported here suggest a few general comments. First, the comprehensive approach to data analysis made it possible to identify a number of interesting patterns in the data which would have gone unnoticed if analysis had been confined to performance measures. In particular, the use of time series analysis to examine time-varying characteristics of subjective effort and performance is a promising approach. Since it is reasonable to expect that cause-and-effect relationships involving effort may not be contemporaneous, the capability to incorporate time lags into the analysis is appealing.

The technique employed for eliciting subjects' thoughts over the course of the session was also very useful, and appears to have been a good compromise between on-line generation of verbal protocols and post-event interviews. Ratings of effort were requested on line to obtain as accurate a representation as possible. In-depth exploration of the events that contributed to the effort ratings was possible in the subsequent interviews without interfering with subjects' control behavior. Replaying the session for the subjects allowed them to reconstruct the production run (apparently with little difficulty), and the ratings jogged their memories and helped them to recall the timing of events.

Consider now the failure to find a relationship between subjective effort and error. The PLANT environment allows substantial operator discretion and is largely forgiving of slips. Tolerance of slips arises from the "slip-tolerant" nature of the interface and the possibility of correcting many erroneous actions before the consequences of such actions become severe. The results of this research suggest that perceived effort is not reliably related to slips in such environments. This is worth noting because operator discretion and self-correction is possible in many of the operational environments in existence.

It is possible, however, that perceived effort in these environments could be related to mistakes. If we consider the reasons subjects gave for increases in effort ratings, many of them could be associated with some cognitive activity (such as figuring out what was happening or trying to develop a plan). Mistakes occurred if the products of such cognitive activity were inappropriate intentions. In fact, in other work (van Eekhout & Rouse, 1981; Johnson & Rouse, 1982) the mere fact that a subject did not understand what was happening was considered to be an error.

Information about the occurrence of mistakes could not be extracted easily from the data obtained in this research, but it is not difficult to imagine an association between mistakes and effort. Two alternatives are feasible given the data available. First, effort may increase in anticipation of a situation in which a mistake is likely. Second, effort may increase as a result of having to deal with the consequences of the mistake. It is not possible to determine which occurred most often here, but the alternatives could be evaluated with the time-series methodology if appropriate data were available.

Considering the overall results of these two experiments, they provide a demonstration of the adaptability of the human operator. Adaptation to three different factors is tenable in light of these results: 1) likelihood of error, 2) perceived effort, and 3) reward contingencies. This might be viewed as a failure of the experimental paradigm to control for unwanted subject variation, thereby preventing the identification of the "true" relationships between the experimental factors.

It is argued, however, that subjects' behavior in these experiments did reflect the way the operator of an actual system might behave. In many situations, adaptation on the part of the operator is possible; in fact, the human element persists in some systems precisely because of its adaptive capabilities. As seen in the results reported here, it may be impossible to understand operator behavior in less constrained situations if the possibility of adaptive behavior is not taken into consideration.

The observation that humans adapt to reward contingencies is not particularly novel, and does not merit further discussion here. It is worthwhile, however, to consider the other forms of adaptation noted. Two general "principles" are offered.

First, as suggested by the results of Experiment One and, to a lesser extent, Experiment Two, people encountering a situation in which errors appear likely respond to that situation by trying to reduce the likelihood of error. They do this by either controlling the situation or controlling themselves (e.g., they are more careful or they change strategies). Errors result if people are placed in situations to which they cannot adapt for some reason. One possibility, which underlies much of the analysis in the conceptual framework, is that people do not perceive the need to adapt until errors have already occurred. In the present research, this does not seem to have been the case.

Second, when experiencing an increase in perceived effort above some "acceptable" threshold, people attempt to reduce the level of perceived effort. The options of controlling the situation or controlling themselves once again apply. This interpretation is consistent with common reports of human "economizing" in decision making.

The implications of these results are interesting. For example, the problems associated with attempting to identify human error rates, as discussed in the beginning of this paper, are underscored. In light of humans' adaptive tendencies, the concept of human error rate seems rather ephemeral. Thus, rather than questioning the likelihood of error in a statistical sense, a more important issue to be addressed is the identification of factors which limit humans' ability to adapt a situation to themselves or vice versa.

Also at issue is the generalizability of results obtained in constrained situations to situations in which more discretion is possible. If research is to provide understanding of human behavior in less constrained environments, then discretion must be possible in the experimental paradigms used. Placing constraints on the environment for the sake of experimental tractability may be necessary, but it must be done with great care. Human adaptation is the norm rather than the exception. Since it appears that adaptation is likely, effort should be devoted to identifying the precipitating conditions and ways in which a human operator is likely to adapt.

REFERENCES

Johnson, W.B., & Rouse, W.B. (1982). Analysis and classification of human errors in troubleshooting live aircraft power plants. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-12,* 389-393.

Norman, D. A. (1981). Categorization of action slips. *Psychological Review, 88,* 1-15.

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

Reason, J. (1983, September). On the nature of mistakes. In N. Moray and J. W. Senders (Eds.), *Preprints of the NATO Conference on Human Error,* Bellagio, Italy.

Reason, J., & Mycielska, K. (1982). *Absent minded? The psychology of mental lapses and everyday errors.* Englewood Cliffs, NJ: Prentice-Hall.

Rouse, W. B., & Morris, N. M. (1985). Conceptual design of a human error tolerant interface for complex engineering systems. *Automatica, 23,* 231-235.

Sheridan, T. B. (1980, February). Human error in nuclear power plants. *Technology Review,* 22-23.

Swain, A. D., & Guttmann, H. E. (1980, August). *Handbook of human reliability analysis with emphasis on nuclear power plant applications* (Report No. NUREG/CR-1278). Washington, DC: Nuclear Regulatory Commission.

van Eekhout, J.M., & Rouse, W.B. (1981). Human errors in detection, diagnosis, and compensation for failures in the engine control room of a supertanker. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-11,* 813-816.

1. Report No. NASA CR-177484	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle HUMAN OPERATOR RESPONSE TO ERROR-LIKELY SITUATIONS IN COMPLEX ENGINEERING SYSTEMS	5. Report Date August 1988 6. Performing Organization Code
7. Author(s) Nancy M. Morris and William B. Rouse	8. Performing Organization Report No.
9. Performing Organization Name and Address Search Technology, Inc. 5550A Peachtree Parkway, #500 Norcross, GA 30092	10. Work Unit No. T3227 11. Contract or Grant No. NAS2-12048
12. Sponsoring Agency Name and Address National Aeronautics and Space Administration Washington, D.C. 20546	13. Type of Report and Period Covered Contractor Report 14. Sponsoring Agency Code

15. Supplementary Notes

Point of Contact: Technical Monitor, Sandra G. Hart, MS: 239-3
Ames Research Center, Moffett Field, CA 94035

16. Abstract

The research reported in this paper is the result of an effort directed at understanding the causes of human error in complex systems. First, a conceptual framework is provided, in which two broad categories of error are discussed: errors of action, or slips, and errors of intention, or mistakes. Conditions in which slips and mistakes might be expected to occur are identified, based on existing theories of human error. Regarding the role of workload, it is hypothesized that workload may act as a catalyst for error.

Two experiments are presented in which humans' responses to "error-likely" situations were examined. Subjects controlled PLANT under a variety of conditions and periodically provided subjective ratings of mental effort. A complex pattern of results was obtained, which was not consistent with predictions. Generally, the results of this research indicate that 1) humans respond to conditions in which errors might be expected by attempting to reduce the possibility of error, and 2) adaptation to conditions is a potent influence upon human behavior in discretionary situations. Subjects' explanations for changes in effort ratings are also explored.

17. Key Words (Suggested by Author(s)) Mental Workload Human Error	18. Distribution Statement Unclassified - Unlimited STAR Category 53

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 41	22. Price* A03

*For sale by the National Technical Information Service, Springfield, Virginia 22161