Int. J. Man-Machine Studies (1993) **39**, 1015–1029

Monitoring behavior in manual and automated scheduling systems

YILI LIU

Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109-2117, USA

ROBERT FULD

ABB Combustion Engineering Nuclear Power, Windsor, CT 06095, USA

CHRISTOPHER D. WICKENS

Department of Psychology and Institute of Aviation, University of Illinois, Champaign, IL 68120, USA

(Received 15 January 1993 and accepted in revised form 27 July 1993)

Human monitoring behavior in manual and automated scheduling systems is examined through an experiment that required the subjects to perform scheduling and monitoring tasks. The task required the assignment of a series of incoming customers to the shortest of three parallel service lines. The subject was either in charge of the customer assignment (Manual Mode) or was monitoring an automated system performing the same task (Automatic Mode). In both cases, the subjects were required to detect the nonoptimal assignments that they or the computer had made. The results showed better error detection performance and lower subjective workload in the automatic mode. The subjects in the manual mode were both biased against declaring their own assignment errors and less sensitive to their misassignments. Results are compared with previous findings of monitoring behavior in manual control systems, and are discussed in terms of human decision making, reliability, workload and system design.

1. Introduction

Although the role of the human operator in modern systems is quickly changing from that of a manual controller to that of a monitor and supervisor, the data and knowledge base from which system design guidelines can be provided remains sparse. Our understanding of the effects of automation and the characteristics of human monitoring behavior remains limited at best; and as a result, the deployment of automation has often been haphazard and ill-conceived (Parsons, 1985; Price, 1985; Wiener, 1988). A clear understanding of the characteristics of human monitoring behavior and the nature of the human supervisory and monitoring role becomes of critical concern for the successful implementation and operation of many modern systems, including commercial and military airplanes (Chambers & Nagel, 1985; Pew, 1986), air traffic control (Parasuraman, 1987; Hopkin, 1992), advanced manufacturing systems (Sharit, 1985; Sanderson, 1989), and process control (Bainbridge, 1983; Sorkin & Woods, 1985).

A major decision regarding the operator's role in automated systems is one of

0020-7373/93/121015 + 15$08.00/0

human involvement and machine autonomy: should the operator serve as a passive monitor of failures and malfunctions with computers performing control functions more autonomously, or should the operator serve as an active participant in the control or decision loop with computers as intelligent assistants? Answers to these questions are not entirely clear. Both the concept of a passive monitor and that of an active participant have received support from arguments based on considerations of workload, safety, system familiarity, and trust.

One major argument that supports the role of the passive monitor is based on aviation accident analysis: since humans are the major cause of most accidents, the argument goes, air safety should be dramatically improved by removing humans from the control loop. A computer performing the same human function is usually more reliable and less susceptible to environmental and system factors (Chambers & Nagel, 1985; Nagel, 1988). Another argument that supports the concept of a passive monitor is based on the belief that automation will reduce the level of operator workload and improve operator performance on the remaining tasks, a belief that can be traced back to one of the earliest philosophies of automation, which stated that "man is best when doing least" (Birmingham & Taylor, 1954).

One major argument that supports the concept of an active participant is that no matter how much of a process is automated, should the process fail, the supervisor will be required to assume the original role as a controller: there is accumulating evidence and concern that removing the human from the control loop inhibits development and speeds the loss of control skill (Chambers & Nagel, 1985). This deficiency has been termed "out of the loop unfamiliarity", or OOTLU (Wickens, 1992). Given that control skills are acquired largely from task performance (Umbers, 1979) and that recall of knowledge from long-term memory requires frequent exercise to remain efficient, it is not surprising that anecdotal evidence for automation-induced OOTLU is accumulating (Wiener, 1988). Another argument that supports the concept of an active participant is that humans do not make terribly effective monitors of highly automated systems. Attentional factors such as vigilance decrements can limit human performance in complex monitoring tasks (Parasuraman, 1987). Furthermore, automation may simply shift the locus of workload by moving the operator to the higher level of a supervisory controller (Sheridan, 1987). There are also arguments that the operator's trust in automated devices, that is critical to the operation of autonomous systems, has yet to be fully established (Muir, 1988; Wiener, 1988; Moray & Lee, 1990).

Despite the increasing urgency of addressing these controversies associated with the changing nature of operator participation in modern systems, few experiments have systematically examined the effects of operators' mode of participation on their monitoring behavior in manual and automated systems. Of the limited number of reported studies, most have been in the realm of manual tracking and flight simulation. Young (1969) and Wickens and Kessel (1979, 1980) have demonstrated the superiority of man-in-the-loop monitoring performance in detecting dynamic system failures (where system failures were defined as sudden changes in controlled element dynamics). Using detection latency and accuracy as measures, both studies showed superior speed-accuracy characteristics for active controllers vs. passive monitors. Both studies attributed the better detection performance when humans are in the control loop to the added proprioceptive information available to the active controller. Similar results have also been reported recently by Bortolussi and Vidulich (1989).

Evidence that demonstrated superior failure detection performance in the automatic mode was provided by Ephrath and Curry (1977), who used professional airline pilots to perform a complex, multi-loop landing approach simulation. In this study, failures consisted of slow, steadily diverging errors in either the yaw or pitch axis. Since failures in this study were not defined by a change in control laws, no adaptation in control behavior was required by the occurrence of a failure. The proprioceptive channel thus did not provide unique cues for the manual mode pilots, and their detection performance declined with the higher workload of concurrent manual control.

While these studies identified proprioceptive information and workload as two critical factors that determine an operator's monitoring performance, the results were derived from manual tracking and flight deck environments that have a significant manual control component. There is a significant lack of understanding of the critical factors that influence monitoring behavior in automated systems that mainly perform decision functions rather than motor control. One important example is advanced scheduling systems, in which schedulers make decisions according to some scheduling rules about the timing and sequencing of performing operations on various tasks to reach certain criteria. In advanced scheduling systems, much scheduling is done by computers equipped with sophisticated scheduling algorithms, and the human's responsibility is to monitor the operation of the computer scheduler. Although we have begun to accumulate knowledge about human scheduling abilities (Sanderson, 1989; Moray, Dessouky, Kijowski & Adapathya, 1991), no study has been reported which examined the characteristics of human monitoring behavior in scheduling systems. Our knowledge accumulated in the manual tracking studies is not readily generalizable to the scheduling environments and other cognitive systems, because of the great differences in the nature of the tasks.

The objective of this study was to examine the characteristics of human monitoring behavior in manual and automated scheduling systems through an experiment that required the subjects to perform scheduling and monitoring tasks. The task required the assignment of a series of incoming customers to the shortest of three parallel service lines. The subject was either in charge of the customer assignment (Manual Mode) or was monitoring an automated system performing the same task (Automatic Mode). In both cases, the subjects were required to detect the nonoptimal assignments that they or the computer had made. Performance indices of the signal detection paradigm were used to compare the error detection sensitivity and decision criterion when the human functions as a active scheduler or a passive monitor. NASA subjective workload ratings (Hart & Staveland, 1988) and time estimation performance as a secondary task were also collected to analyse the workload demands and their effects on monitoring performance.

One important factor that has been identified as having significant influence on human performance in automated systems is the operators' level of self-confidence in their own judgments and their trust in automated devices (Moray & Lee, 1990). A general conclusion from previous studies is that people tend to be overconfident in their judgments of their own abilities in performing various tasks such as diagnosis, forecasting, and memory retrieval. This bias has been observed in several forms and in several contexts (Fischhoff & MacGregor, 1982; Pitz & Sachs, 1984). For example, in the context of automotive troubleshooting, Mehle (1982) observed that subjects are unjustly confident that they have entertained all possible diagnosis hypotheses.

In the context of decision aids, there is also evidence that decision-makers' level of self-confidence is often overestimated, and that this overconfidence in their own judgments leads them to mistrust the outcomes of various automated decision aids designed to supplement their judgments (Kleinmuntz, 1985, 1990). This overconfidence is often a manifestation of a well documented decision bias called confirmation bias: decision makers tend to seek (and therefore find) information that confirms their chosen hypothesis (Einhorn & Hogarth, 1978; Kahneman, Slovic & Tversky, 1982). This bias produces a sort of "cognitive conceit" (Edwards, 1968) or "cognitive tunnel vision" (Sheridan, 1981), in which decision makers fail to encode or process information that is contradictory to or inconsistent with their initial hypothesis.

These results have significant implications for the analysis of automated cognitive systems, and also provided us the basis for a prediction about human performance in scheduling systems. We predicted that, besides workload and information requirements, a third dimension—that of possible biases in decision making—is an important factor that determines human monitoring performance. The effects of the three factors on human monitoring behavior in manual and automated scheduling systems were investigated in this study.

2. Method

2.1. TASKS

A dynamic scheduling and monitoring task was developed to simulate a supermarket "customer" assignment situation. The task required the assignment of a series of incoming "customers" to the shortest of three parallel service lines displayed on a computer display (see Figure 1). The display area of each "customer" was directly

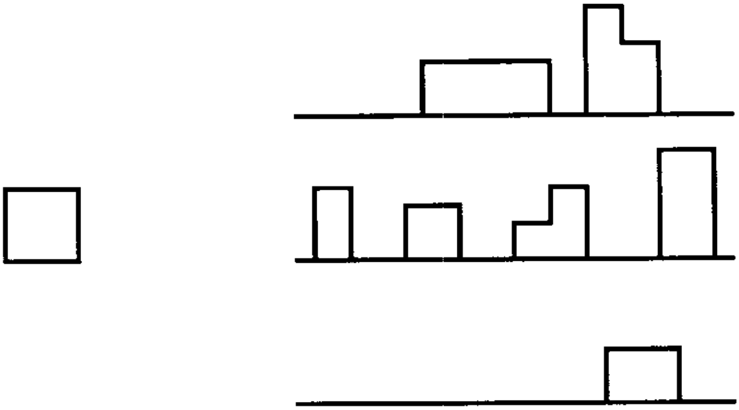

FIGURE 1. A pictorial representation of the task display. The "customer" arriving at the left is waiting to be assigned to one of the three lines at the right. In this example the optimal choice with the minimal waiting time is clearly line 3, which has the least sum of areas.

proportional to the service time required by the given "customer". Thus the shortest line refers to the line with the least sum of areas. The shape of a customer did not provide information for customer assignment. One customer arrived at a time, and the interval between the arrival of a new customer and the assignment of the previous customer was randomly distributed between 3 and 4 seconds.

During different experimental sessions, the new "customers" were assigned either by the subjects (Manual Mode), or by the computer (Automatic Mode). In both cases the subjects were required to monitor whether the assignments that they or the computer had made were optimal (i.e. whether the "customer" was assigned to the line with the shortest expected wait), and indicate nonoptimal assignment by pressing a left-hand key (an "oops" response). In the manual mode, assignments were made by pressing one of three left-hand keys corresponding to the three service lines. The subjects were instructed that speed and accuracy were equally important for making customer assignment. However, they were required to make customer assignment immediately if they heard a high-pitched tone, which was used as a "time-out" signal to create a time pressure of customer assignment by subjects. The time interval between "customer" arrival and tone presentation was varied to create fast and slow conditions. The high-pitched tone was presented to a subject only if he/she failed to make an assignment within the time interval. In the automatic mode, the subjects were actually monitoring the "shuffled" playback of a computer recording of their own manual sessions. Thus, the frequency as well as the specific stimulus appearance of the nonoptimal assignments in the automatic mode was yoked to the frequency and appearance with which they occurred in the manual mode. Yet the sequence of automatic trial segments never replicated the sequence of the manual segments. By doing so, the two modes can be compared at the same difficulty level of monitoring task. The subjects were never informed of this identity before the end of the experiment, nor did they report that they ever realized it when they were informed of this identity after completing the experiment.

In addition to the scheduling and monitoring task, the subjects were required to perform a time estimation secondary task, which was employed as an objective measure of task workload. For the time estimation task, the subjects were asked to estimate elapsed time by the time production method (Hart, 1975): They were instructed to estimate 10 s intervals and press a right-hand time estimation key whenever they felt that 10 s had elapsed since the previous keypress. The subjects were also instructed not to use counting, tapping or any sort of direct timing procedures.

2.2. EXPERIMENTAL DESIGN AND PROCEDURE

In order to provide adequate training to the subjects, each subject was given 10 "customer assignment" training sessions of 1 h each. These 10 training sessions were held on 10 separate days. Nine right-handed subjects participated in the experiment and were paid four dollars per hour as remuneration. In each training session, visual and auditory feedback was presented on a trial-by-trial basis to develop proficiency and consistency at lane size discrimination. The subjects were instructed to reach an assignment accuracy of at least 70%, which was over twice what would be expected by chance (33%). The mean assignment time of the last five training sessions of each subject (RT) was used as the criterion for calculating the time pressure of fast and slow conditions of the experimental sessions for that subject. The time pressure was twice the mean assignment time (2·00 * RT) for the slow condition and 0·75 times the mean assignment time (0·75 * RT) for the fast condition.

At the end of the "customer assignment" training sessions, subjects performed one 12·5 min time estimation block with no concurrent activities, followed by three 12·5 min time estimation blocks with only manual "customer assignment". Then, subjects performed another 12·5 min time estimation block with no concurrent activities. The average time estimation performance of the two blocks in which subjects performed only time estimation was used as the baseline for time estimation performance. Subsequently, the subjects were given a 1-h training session with error detection and time estimation requirements imposed onto the routine customer assignment task. This session was designed to familiarize subjects with all tasks in the experimental situation.

The last 4 days were the experimental sessions (two manual and two automatic mode sessions). Because of the "yoked" nature of the experiment and thus the recording requirement for the automatic condition, the first and the third days for each subject were "manual mode" sessions, and the second and fourth days were "automatic mode" sessions. The trial segment orders were "shuffled" between modes, and the fast and slow trials were counterbalanced between sessions. Each session consisted of four blocks (two fast and two slow) of 12·5 min each. The subjects were instructed to give top priority to customer assignment and error detection, and give second priority to time estimation.

After each block, the NASA–TLX workload scale was used to collect the subject's subjective workload experience in that block. There are two main components to the procedure: the rating scales themselves and the assignment of importance weights to the different attributes. This procedure resulted in a weighted workload rating and six subscale ratings for the subject in each experimental condition. The six subscales include three task-related factors (physical requirement, mental requirement, time requirement), and three subject-related factors (amount of effort, frustration, success and failure). For a detailed description of the rationale and implementation of the NASA workload scale, see Hart & Staveland (1988).

3. Results

3.1. SCHEDULING PERFORMANCE

The "customer" assignment performance of the subjects in the manual mode was measured as customer assignment error frequencies, which are presented in the first and the fourth rows of the data in Table 1. An error in customer assignment occurred when a customer was assigned to a line that did not have the least sum of areas. Error frequency was calculated as the ratio of the number of assignment errors to the total number of assignments. There are two major results that have implications for analysing the results of monitoring performance, which is the focus of the current study: first, the time pressure manipulation did not produce

TABLE 1

Performance of each subject on the scheduling and the monitoring tasks recorded as "customer" assignment error frequencies and error detection frequencies

Subject No.		1	2	3	4	5	6	7	8	9	Mean
Slow Speed	Assignment error	0·344	0·250	0·280	0·329	0·258	0·291	0·230	0·393	0·328	0·300
	Manual detection	0·207	0·104	0·093	0·202	0·080	0·058	0·050	0·130	0·157	0·120
	Automatic detection	0·227	0·237	0·165	0·361	0·239	0·202	0·161	0·172	0·363	0·236
Fast Speed	Assignment error	0·356	0·302	0·254	0·246	0·279	0·298	0·180	0·376	0·296	0·287
	Manual detection	0·312	0·131	0·076	0·246	0·119	0·057	0·054	0·149	0·189	0·148
	Automatic detection	0·260	0·285	0·147	0·412	0·295	0·172	0·158	0·151	0·387	0·252

a significant effect on assignment accuracy ($t(8) = 1{\cdot}017, p > 0{\cdot}10$). Second, two subjects (S1 and S8) were not only the least accurate of the group in terms of the assignment task, but were also the only subjects to fall consistently below the accuracy criterion set by the experimenter.

3.2. MONITORING PERFORMANCE

Monitoring performance was analysed with the performance indices of the signal detection paradigm: error detection sensitivity and decision criterion.

3.2.1. Error detection sensitivity

The hit and false alarm rates at each level of mode and speed were first calculated for each subject. A hit was an event in which a subject made a detection response when an error actually occurred. A false alarm was defined as a detection response when no error occurred. A′ scores, a related measure of sensitivity (Pollack & Norman, 1964), were then computed and subjected to statistical analysis. This method does not make any assumptions about the form of the underlying signal and noise distributions. It requires only that the receiver operating characteristic (ROC), a Cartesian plot of hit probability p(H) against false alarm probability p(FA), be monotonically increasing. This means that for a receiver of constant sensitivity, no increase in hit rate will be accompanied by a decrease in false alarm rate, and vice-versa. For a detailed description of the calculation and rationale of A′ scores, see Pollack and Norman (1964), or Wickens (1992).

While the current experimental paradigm depended on subjects assigning customers to the best of their ability, nothing precluded them from producing intentional and therefore more easily detected errors in the manual mode. To the

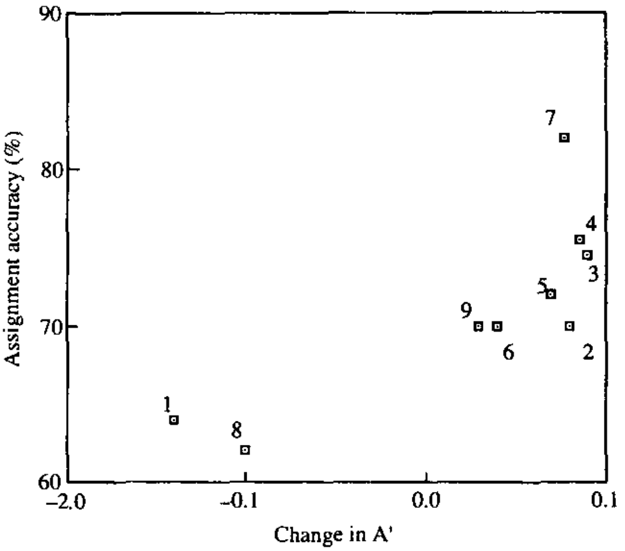

FIGURE 2. Scatterplot of the nine subjects' assignment accuracy vs. mean A′ increment from manual-fast to automatic-fast conditions. The numbers in the scatterplot correspond to the subject numbers.

extent that intentional errors in assignment provided information to the subject about error likelihood, it would create a bias that favored manual mode detection.

To investigate this possibility, a scatterplot showing assignment accuracy against A′ difference is shown in Figure 2. It is evident that two subjects (S1 and S8) are more different from the group than the rest of the group is from itself. They were the same two subjects whose "customer assignment" accuracy fell consistently below the accuracy criterion set by the experimenter. For these reasons, it was felt that these subjects may have adopted a strategy of achieving higher detection accuracy in the manual mode by intentionally sacrificing their assignment performance. Thus, the data of these two subjects are excluded from further analysis.

The A′ scores were presented in Figure 3, and examined in a two-way repeated measures analysis of variance. Detection sensitivity was found to be significantly higher in automatic than in manual conditions ($F(1, 6) = 48{\cdot}2, p < 0{\cdot}001$). No significant main effect for task pace was found ($p > 0{\cdot}10$), nor was there any interaction between the two factors of mode and pace ($p > 0{\cdot}10$).

3.2.2. Response Bias

One of the reasons for the popularity of the signal detection theory is the proposition that the receiver's sensitivity and response criteria are orthogonal dimensions that can be examined separately. According to the signal detection theory (Green & Swets, 1966), when there are no differences in payoffs between the different events, subjects optimally should match their response frequency to the frequency with which signals occur. Signals are, in this case, assignment errors. Any mismatch between the two rates reflects a response bias.

To examine the effects of response bias, the system's signal rates were compared with the subject's response ("oops") rate in each mode (Table 1). Since the

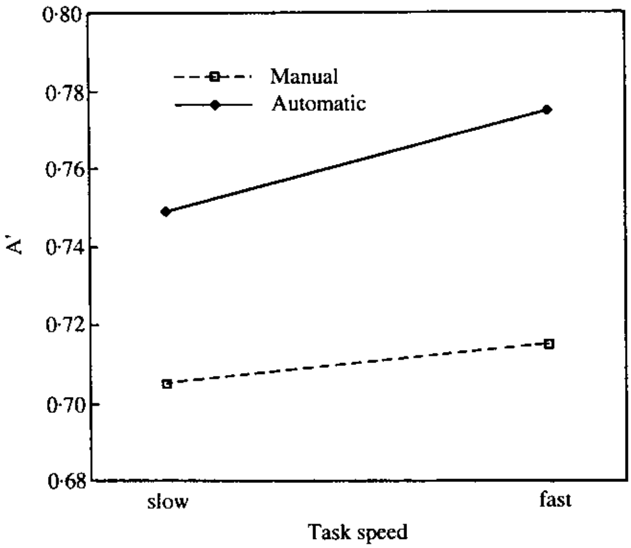

FIGURE 3. Mean A′ data as a function of participatory mode and task speed.

differences of interest were within and not between the two paces, two separate one-way repeated measures ANOVAs with three levels (error frequency, manual error detection frequency, and automatic error detection frequency) were performed, corresponding to the two levels of task pace. The result of each analysis showed significant differences ($p < 0{\cdot}01$ for all); thus paired comparisons of means were performed within each of the two ANOVAs.

Since multiple *post hoc* comparisons on a single set of data produce a cumulative "family-wise" error, these single-*df* comparisons were evaluated against Tukey's *F* as a correction (Keppel, 1982). The results of the paired comparisons, as presented in Table 2, showed that rate of responses in the automatic mode did not differ significantly from the signal rate for either subset of the sample at either task pace. This indicates that subjects' error detection response rates tended to be optimal in the automatic mode. However, in the manual mode, subjects were significantly less frequent in responding relative to both the optimal prescription of signal rate and to the automatic mode ($p < 0{\cdot}01$ for all, see Table 2), indicating a conservative response bias.

TABLE 2
Paired comparison F values for signal vs. response frequency

Pace	Signal vs. Manual	Signal vs. Automatic	Manual vs. Automatic
Slow	182·74*	2·33	78·46*
Fast	24·46*	0·00	110·07*

$^*p < 0{\cdot}01$

3.3. MONITORING WORKLOAD

The means of the NASA-TLX workload ratings under the four experimental conditions are shown in Figure 4. A two-way repeated measures analysis of variance demonstrated that workload in the manual mode was significantly higher than in the automatic mode ($F(1, 6) = 7{\cdot}97, p < 0{\cdot}05$), and that there was a significant interaction between the factors of mode and speed ($F(1, 6) = 20{\cdot}10, p < 0{\cdot}01$), but no significant speed effect ($F(1, 6) = 3{\cdot}45, p > 0{\cdot}10$). A Fisher test further indicated that the subjective workload rating was significantly higher in the manual-fast condition than in any of the other three conditions, while no significant differences existed among these three means.

The time estimation data are summarized in Table 3. Earlier research has found that the median of the length of the 10 s intervals was a more representative measure of central tendency than the mean, and the average absolute deviation of scores from the median was a more representative measure of dispersion than the standard deviation from the mean (Hart, 1975). Thus, the median and the average

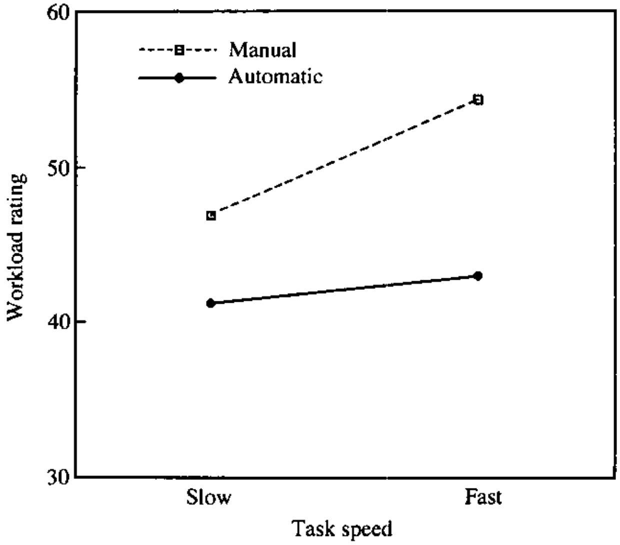

FIGURE 4. Mean workload ratings as a function of participatory mode and task speed.

TABLE 3
Time estimation performance (s)

	Manual fast	Manual slow	Automatic fast	Automatic slow	Assignment only	Time estimation only
Median	15·56	16·01	16·58	17·22	12·55	13·03
Absolute deviation	3·86	4·09	4·38	4·41	2·22	2·40

absolute deviation were used in the current analysis. Both the length and variability of the 10 s intervals were approximately the same for the single task time estimation sessions and the sessions with only the customer assignment task ($t = 0{\cdot}38$ for the median t-test, $t = 0{\cdot}49$ for the absolute deviation). However, a repeated measures one-way ANOVA revealed that the added requirement to monitor and detect assignment errors increased the length and variability of time estimation ($F(5, 35) = 6{\cdot}46$ for the median length, $p < 0{\cdot}01$; $F(5, 35) = 11{\cdot}20$ for the average deviation, $p < 0{\cdot}01$). Fisher tests revealed that both measures of time estimation performance under the four experimental conditions had significantly larger values than under the baseline condition, yet the four experimental conditions did not show significant differences among themselves.

4. Discussion

The objective of the current study was to examine the characteristics of human monitoring behavior in manual and automated scheduling systems. Indices of the signal detection paradigm—namely, detection sensitivity and response criterion—were used to analyse subjects' monitoring performance. Several interesting results regarding subjects' monitoring performance were observed in this study: subjects showed greater sensitivity to assignment errors in the automatic than in the manual condition; and they adopted conservative, nonoptimal response criteria in the manual mode of the error detection task. Each of these issues will be discussed in turn.

4.1. ERROR DETECTION SENSITIVITY

As discussed in the introduction, previous studies on operator performance in manual and automated tracking and control systems have identified the availability of task information and the level of workload as two critical factors that influence monitoring performance. Manual performance of a task will aid the operator in detecting system failures to the extent that it provides the operator with information about the system, but will degrade detection performance insofar as it increases the operator's workload (Wickens & Kessel, 1979; Ephrath & Young, 1981; Rouse, 1981). To use this heuristic as a basis for predicting whether a manual or automated detection environment will prove more favorable thus requires differentiating the effects of information and workload on task performance.

To disambiguate these effects, the present study employed objectively similar stimuli between modes to partially control for differences in detection task information available to the subject. In the automatic mode, the subjects actually monitored the yoked playback of their own manual sessions, although they were never informed of the identity, nor did they ever realize it. It is very difficult to see how the present automatic environment could offer unique information about error probability (i.e. information not similarly available in the manual mode) to the operators. In contrast, in the manual mode, the task of assigning the customers may have allowed some subjects to gain some knowledge and even control of how error probabilities were changing from moment to moment (and in fact, two subjects very likely capitalized on this control). It therefore seems safe to say that to whatever extent an information inequality existed, it could not have favored the automatic mode. We thus have a point of departure for arguing that reduced workload,.and not an information advantage, accounted for the greater detection sensitivity observed in the automatic mode. The subjective workload measures indeed provided indication that higher levels of workload were perceived in the manual mode.

This finding of a sensitivity decrement produced by higher workload is in contrast to that of Wickens and Kessel (1979). In their study, the added workload of controlling a primary tracking task did not hinder the detection of failures in that task. To explain the different results of the two experiments, we need to examine the various subtasks' demands for attentional resources. In Wickens and Kessel's study, the resources required by tracking concerned mainly the response mechanisms, whereas the resources involved in their failure detection task were primarily related to perceptual and decision-making mechanisms. Their failure detection task required a comparison of visual feedback with expectancies of the qualitative way in which the tracking cursor would respond to a given control input. Since the response processes and perceptual/decision-making processes draw from different processing resources that are not mutually available (Wickens, 1992), it is not surprising that little significant interferences were observed between tracking and detection.

However, in the present study, the "customer" assignment and the detection tasks are both perceptual/decision-making tasks, and thus draw from the same perceptual/cognitive processing resources. This competition for common processing resources between the assignment task and the monitoring task resulted in the interference between the assignment and the error detection tasks and thus the lower detection sensitivity.

4.2. RESPONSE BIAS IN MONITORING

As a dimension that is independent of the sensitivity of the observer, response bias allows a comparison between the frequency of to-be-detected signals and the frequency of the monitor's response. This distinction allows description of a monitor as risky (declares an error too frequently), optimal (declares errors with the same probability as that of an erroneous customer assignment), or conservative (declares an error too infrequently).

In the automatic mode, subjects produced response rates that were not significantly different from the optimal frequency of response. However, in the manual mode, subjects were significantly less frequent in responding relative to the optimal prescription of signal rate, indicating a conservative response bias. This conservatism, or "cognitive conceit" (Edwards, 1968; Fischoff, 1977), has been demonstrated in studies of human decision making. The current study, adopting a unique approach of employing "yoked" manual and automatic conditions, provides new evidence of this conservatism in a monitoring environment.

One theoretical explanation of this phenomenon is the expectancy theory, which was originally proposed by Baker (1961) to explain the vigilance decrement phenomenon. This theory considers decrements in detection performance during a watch to be due to a conservative adjustment of the response criterion in response to the observed infrequency of events. Subjective expectancy, then, is reduced every time a signal is missed by the observer. This leads to a sequential increase in conservatism until, in concert with other factors in the detection environment, a new level of equilibrium for the criterion is achieved. Thus a lower initial expectancy may lead to a spiraling increase in missed signals and response bias. This phenomenon is also labeled as the "vicious circle" hypothesis (Broadbent, 1971).

While the current experiment was not a vigilance paradigm (i.e. trial sessions were not lengthy nor signals infrequent), this theory does shed some light on the interpretation of the current results. It illustrates that the cognitive conceit of an overconfident monitor may plant the seed of conservatism, and this conservatism can be further self-reinforced in the process of monitoring. In addition to the observation that subjects could not detect many of their own nonoptimal assignments, another important observation is that they did not realize the existence of this limitation in their ability to detect their own errors. Subjects expressed great surprise when they were informed, at the end of the experiment, that the computer assignment errors in the automatic mode were actually the subject's own errors recorded in the manual mode and that they detected more of these errors in the automatic mode than in the manual mode. These observations have important implications for flexible automation systems. Operators in these systems should be made aware of the existence and the effects of their overconfidence bias or cognitive conceit in their training procedures and when they use automated decision aids. Knowledge of performance results should be added to the system if possible.

The current results also have important implications for certain human reliability issues. Adams (1982) has suggested that one attribute of human operators is that they are often self-correcting with respect to their own errors. However, the present results suggest that where the system does not provide performance feedback, some kinds of errors are less, rather than more, likely to be detected by an operator that is "in the loop". With the terminology of error analysis provided by Norman (1983), the present study provided evidence that some *mistakes* (errors of perception of situation or of the choice of intentions) are more, rather than less, obscure to skilled, confident operators, although skilled operators are shown to be good at detecting their own *slips* (errors of executing intentions) (Rabbitt, 1978; Woods, 1984; Reason, 1990). Future research should identify adequate feedback mechanisms and training procedures to help operators realize and overcome these problems in monitoring tasks.

In summary, the present study indicates that, besides workload and information requirements, a third dimension—that of possible biases in decision making—is an important consideration in the analysis of human monitoring behavior in automated cognitive systems and in the allocation of monitoring functions to human operators. While previous studies have demonstrated the superiority of man-in-the-loop monitoring performance in motor control and tracking systems because of the added proprioceptive information, the results of the current study seem to support the concept of man-out-of-loop monitoring in automated cognitive systems that perform decision functions rather than motor control. Future research should further examine these important issues of operators' role in modern systems, and identify adequate feedback mechanisms, training procedures and support tools to improve operators' monitoring performance.

This research was supported by Grant #NASA NAG-2-308 from the NASA Ames Research Center. Sandra Hart was the technical monitor. We would like to acknowledge the support of Gunnar P. Seaburg in programming the task. We are also indebted to comments made by three anonymous referees.

References

ADAMS, J. A. (1982). Issues in human reliability. *Human Factors,* **24,** 1–10.

BAINBRIDGE, L. (1983). Ironies of automation. *Automatica,* **19,** 775–779.

BAKER, C. H. (1961). Maintaining the level of vigilance by means of knowledge of results about a secondary vigilance task. *Ergonomics,* **4,** 311–316.

BIRMINGHAM, H. P. & TAYLOR, F. V. (1954). *A human engineering approach to the design of man-operated continuous control systems* (Report No. 433). Washington, DC: US Naval Research Laboratory.

BORTOLUSSI, M. R. & VIDULICH, M. A. (1989). The benefits and costs of automation in advanced helicopters: an empirical study. *Proceedings of the 5th International Symposium on Aviation Psychology,* pp. 594–599. Columbus: Ohio State University, Department of Aviation.

BROADBENT, D. E. (1971). *Decision and stress.* New York: Academic Press.

CHAMBERS, A. B. & NAGEL, D. C. (1985). Pilots of the future: human or computer? *Communications of the ACM,* **28**(11), 1187–1199.

EDWARDS, W. (1968). Conservatism in human information processing. In B. KLEINMUNTZ, Ed. *Formal representation of human judgment,* pp. 17–52. New York: Wiley.

EINHORN, H. J. & HOGARTH, R. M. (1978). Confidence in judgment: persistence of the illusion of validity. *Psychological Review,* **85,** 395–416.

EPHRATH, A. R. & CURRY, R. E. (1977). Detection by pilots of system failures during instrument landings. *IEEE Transactions on Systems, Man, and Cybernetics,* **SMC-7,** 841–852.

EPHRATH, A. R. & YOUNG, L. R. (1981). Monitoring vs. man-in-the-loop detection of aircraft control failures. In J. RASMUSSEN & W. B. ROUSE, Eds. *Human detection and diagnosis of system failures,* pp. 143–154. New York: Plenum Press.

FISCHOFF, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance,* **3,** 349–358.

FISCHOFF, B. & MACGREGOR, D. (1982). Subjective confidence in forecasting. *Journal of Forecasting,* **1,** 155–172.

GREEN, D. M. & SWETS, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

HART, S. G. (1975). Time estimation as a secondary task to measure workload. *Proceedings of the 11th Annual Conference on Manual Control,* pp. 64–77. Pasadena, CA: JPL Publications.

HART, S. G. & STAVELAND, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In P. A. HANCOCK & N. MESHKATI, Eds. *Human mental workload,* pp. 139–183. Amsterdam: North-Holland.

HOPKIN, V. D. (1992). Human factors issues in air traffic control. *Human Factors Society Bulletin,* **35,** 1–4.

KAHNEMAN, D., SLOVIC, P. & TVERSKY, A., Eds. (1982). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

KEPPEL, G. (1982). *Design and analysis.* Englewood Cliffs, NJ: Prentice Hall.

KLEINMUNTZ, B. (1985). Cognitive heuristics and feedback in a dynamic decision environment. *Management Science,* **31**(6), 680–702.

KLEINMUNTZ, B. (1990). Why we still use our heads instead of formulas: toward an integrative approach. *Psychological Bulletin,* **107**(3), 296–310.

MEHLE, T. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychologica,* **52,** 87–116.

MORAY, N., DESSOUKY, M. I., KIJOWSKI, B. A. & ADAPATHYA, R. (1991). Strategic behavior, workload, and performance in task scheduling. *Human Factors,* **33,** 607–629.

MORAY, N. & LEE, J. (1990). *Trust and allocation of function in the control of automatic systems* (Technical Report EPRL-90-05). Urbana, University of Illinois, Engineering Psychology Research Laboratory.

MUIR, B. M. (1988). Trust between humans and machines, and the design of decision aids. In E. HOLLNAGEL, G. MANCINI & D. D. WOODS, Eds. *Cognitive engineering in complex dynamic worlds*, pp. 71–83. London: Academic Press.

NAGEL, D. C. (1988). Human error in aviation operations. In E. L. WIENER & D. C. NAGEL, Eds. *Human factors in aviation*, pp. 263–303. New York: Academic Press.

NORMAN, D A. (1983). Design rules based on analyses of human error. *Communications of the ACM*, **26**, 254–258.

PARASURAMAN, R. (1987). Human–computer monitoring. *Human Factors*, **29**, 695–706.

PARSONS, H. M. (1985). Automation and the individual: comprehensive and comparative views. *Human Factors*, **27**, 99–112.

PEW, R. W. (1986). Human performance issues in the design of future air force systems. *Aviation, Space, and Environmental Medicine*, **10**, 78–82.

PITZ, G. F. & SACHS, N. J. (1984). Judgment and decision: theory and application. *Annual Review of Psychology*, **35**, 139–163.

POLLACK, I. & NORMAN, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, **1**, 125–126.

PRICE, H. E. (1985). The allocation of function in systems. *Human Factors*, **27**, 33–45.

RABBITT, P. M. A. (1978). Detection of errors by skilled typists. *Ergonomics*, **21**, 945–958.

REASON, J. (1990). *Human error*. New York: Cambridge University Press.

ROUSE, W. B. (1981). Human computer interaction in the control of dynamic systems. *Computing Surveys*, **13**, 71–99.

SANDERSON, P. (1989). The human planning and scheduling role in advanced manufacturing systems: an emerging human factors domain. *Human Factors*, **31**, 635–666.

SHARIT, J. (1985). Supervisory control of a flexible manufacturing system. *Human Factors*, **27**, 47–60.

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

SHERIDAN, T. B. (1981). Understanding human error and aiding human diagnostic behavior in nuclear power plants. In J. RASMUSSEN & W. B. ROUSE, Eds. *Human detection and diagnosis of system failures*, pp. 19–35. New York: Plenum Press.

SORKIN, R. D. & WOODS, D. D. (1985). Systems with human monitors: a signal detection analysis. *Human–Computer Interaction*, **1**, 49–75.

UMBERS, I. G. (1979). Models of the process operator. *International Journal of Man–Machine Studies*, **11**, 263–284.

WICKENS, C. D. (1992). *Engineering psychology and human performance*. New York: Harper Collins.

WICKENS, C. D. & KESSEL, C. (1979). The effects of participatory mode and task workload on the detection of dynamic system failures. *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-9**, 24–34.

WICKENS, C. D. & KESSEL, C. (1980). Processing resource demands of failure detection in dynamic systems. *Journal of Experimental Psychology, Human Perception and Performance*, **6**, 564–577.

WIENER, E. L. (1988). Cockpit automation. In E. L. WIENER & D. C. NAGEL, Eds. *Human factors in aviation*, pp. 433–461. San Diego: Academic Press.

WOODS, D. D. (1984). Some results on operator performance in emergency events. *Institute of Chemical Engineers Symposium Series*, **90**, 21–31.

YOUNG, L. R. (1969). On adaptive manual control. *IEEE Transactions on Man–Machine Systems*, **MMS-10**, 292–331.

Paper accepted for publication by Associate Editor Professor R. C. Williges.