

TIME-BASED MODELING OF HUMAN PERFORMANCE

Esa M. Rantanen and Brian R. Levinthal

Aviation Human Factors Division, University of Illinois at Urbana-Champaign
Savoy, Illinois

This paper presents a probabilistic approach to modeling human performance. Instead of focusing on mean performance, the effects of taskload on the distributions of performance variables are examined. From such data, probabilities of given levels of performance can be derived and methods of measurement that expand the analyses beyond those of the mean developed. Results from two experiments, one abstract, the other realistic, are presented in terms of timely performance on required tasks. As taskload increased, the participants were less likely to act on the experimental tasks at an earliest opportunity than under low taskload, resulting in increase of 'too late' errors. Measurement of taskload and performance in temporal terms also allowed for bracketing and making inferences about mental workload, which is not directly measurable.

INTRODUCTION

Human performance in controlling dynamic systems has a reasonably long history in psychological and engineering research, but time as a focal point in human performance studies has only recently gained in importance. From studies of operator performance in process control it is clear that operators do not merely react to information presented to them on panels and meters, but act on the basis of their understanding of the context of the information and the specific task they are performing (Hollnagel, 1993). The difficulties humans have in the control of dynamic situations and systems are well documented (Wagenaar & Sagaria, 1975; Wickens & Hollands, 2000) as are the often catastrophic consequences of these complications in many high-risk sectors (De Keyser, 1995). Examples of such occurrences are air traffic controllers who 'lose the picture' of traffic under their responsibility, pilots who 'fall behind' their aircraft, and control room operators who become overwhelmed by unanticipated events and are forced into a reactive mode of operation. Many operator errors can also be classified as temporal (Decortis et al., 1991).

Common to all examples mentioned above is time. More specifically, the ratio of time available to time required for performing some task, whether it is physical execution of some control action or diagnosis of plant state from available information or a combination of such activities, appears to determine the success of control of dynamic systems. Hollnagel's (1993) Contextual Control Model (COCOM) postulates a positive linear relationship between perceived time available and degree of control and specifies four control modes along this continuum: Strategic control (and high degree of control) is possible when time available is high; as time available decreases so does the operator's span of planning and he or she will be forced to adopt first a tactical control mode, then opportunistic control mode, and finally scrambled control mode. The first two control modes can also be described as proactive and the latter two as reactive. The performance implications are straightforward: success in the task is predicated by the degree of control, proactive behavior resulting in good and reactive behavior in poor performance.

The COCOM model and the inherent role of time in it suggest two important extensions to it. First, the demands of a multi-task system can be objectively measured as 'taskload,' and have a bearing on the ability of a controller to successfully

maintain a safe and efficient status of a dynamic system. Taskload is often equated to the ratio of time available and time required (Gawron, 2000; Stone et al., 1984; Gunning & Manning, 1980). Taskload has in turn been shown to be a workload driver (Raby & Wickens, 1994; Tulga & Sheridan, 1980; Moray et al., 1991). As taskload increases, the ability of an operator to successfully control a system generally decreases. The extent to which such a shift in performance differs from the expected is said to be a result of mental workload, a subjective interpretation of task demands that is related both to individual differences such as expertise and fatigue, and external factors such as time pressure and prior performance (Meshkati, 1988; Hart, 1986; Ogden et al., 1979). While mental workload cannot be measured directly, there often is a negative relationship between it and performance (Raby and Wickens, 1994; Hart, 1986; Sperandio, 1971, 1978). As the latter can be measured, inferences about the workload experienced by the operator may be made based on his or her performance. Hence, according to the above rundown, covert mental workload could be bracketed by measuring overt taskload and performance. Second, the notion of the four control modes implies some distinct, identifiable patterns of behavior or performance in each of them in terms of time. Errors and poor performance are often synonymous and many operator errors are classified as temporal (De Keyser, 1995; Decortis et al., 1991).

In this paper we pursue three distinct themes. First, manipulation of taskload was done in terms of the ratio of time required to time available and the actual taskload (as influenced by the participants actions in the task) as well as performance were also measured in terms of time. Specifically, we hypothesized that under high taskload the participants would shift from proactive to reactive mode, manifested in deterioration of planning and failures of anticipation and resulting in less-than-timely performance of the tasks. Second, we took a probabilistic approach to the measurement of human performance. Rather than mean performance we examined the effects of taskload on the distributions of selected performance variables. From such data, probabilities of given levels of performance can be derived. This approach can be used to develop methods of measurement that would expand the analyses beyond those of the mean (e.g., t-tests, ANOVA). The events important to human factors research typically involve the tails of the response distributions rather than the average responses (Wickens, 2001). The tails often represent the ex-

ceedances of tolerances or criteria (e.g., too fast, too slow, too early, too late, etc.) and are rightly classified as errors (Hollnagel, 1998). Because of the inherent difficulties in the study of typically very rare errors, instead of focusing solely on the means of responses, dependent variables in experiments can be derived from response distributions to capture also the ‘tail ends’ of human performance. Finally, we report two experiments: one was simple and involved an abstract task; the other extended the paradigm to a realistic air traffic control task. Both involved multiple simultaneous tasks between which participants had to time-share.

EXPERIMENT 1

Method

Participants. Nine students from the University of Illinois at Urbana-Champaign, 6 female, 3 male, ages ranging from 22 to 30 years volunteered to take part in this experiment.

Apparatus and the experimental task. An experimental computer program presented the participants with a dynamic multi-task environment in which they were responsible for the monitoring, scheduling, and performance of four simultaneous tasks. The computer display was divided into four equal panes, of which only one could be viewed at a time by moving a cursor to it by a mouse. Each pane contained two elements, a progress bar and instructions (‘Enter Number to Reset’ followed by a random four-digit number; see Fig. 1).

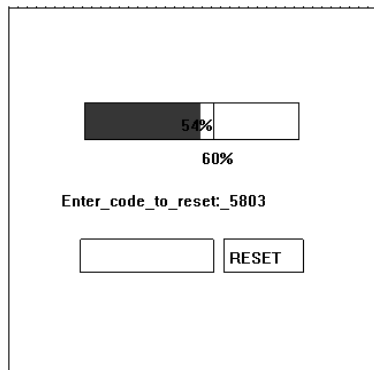


Figure 1. A sample pane from Exp. 1. The progress bar filled from left to right; participants completed the typing task after the bar passed the 60% marker but before it reached 100%.

Each task’s instructions were to be carried out within a specific window of opportunity (WO), which was represented visually as a portion of the progress bar. The bar filled from left to right at one of three varied speeds, and was marked at one of three locations to indicate the start of the tasks’ WO. After the participant entered the correct number and pressed ‘Enter’, a new task immediately replaced the completed task.

Independent variables. The combination of bar speed and percentage at which the WO opened resulted in the three durations of WO used in the experiment (the two extreme pairings of speed and window size were omitted). High and low taskload conditions were created by sampling tasks from two distinct distributions of WO. Both task distributions presented

participants with the same proportion of window durations, 3, 6, or 12 seconds. The taskload manipulation itself results from the fact that these two distributions used different bar speeds. The 1:2:4 ratio of bar speeds represented 60, 30, and 15-second bars, respectively. Hence, tasks appearing in the high taskload condition took half as long as those from the low taskload condition. A high-taskload epoch also had twice as many tasks as that of a low-taskload epoch. The nominal ratio of taskload in high vs. low taskload epochs was 2:1. Blocks of trials consisted of an epoch of high taskload and an epoch of low taskload. The order of presentation of these epochs (high to low or low to high) changed for each block.

Dependent variables. The main performance variable was time to first action (TFA), which was defined as the elapsed time from opening of WO to the time the task was performed, calculated by the first action on the task (i.e., first keystroke).

Results

Taskload. As the participants’ performance could impact the actual taskload experienced in the experiment, we had to quantify the taskload for each participant and for given epochs in the experimental blocks. This was done by the equation

$$TL_A = \frac{n \left(\sum_{i=1}^n TR_i \right)}{\sum_{j=1}^m WO_j} \tag{1}$$

where n is the number of tasks (timer resets) in an epoch, TR is the time required to perform the task, and WO is the duration of the window of opportunity for each task. The results show that the taskload manipulation was successful and the participants indeed experienced significantly different levels of taskload during the experiment. The planned shifts in taskload in the two experimental conditions were also clear (Fig. 2).

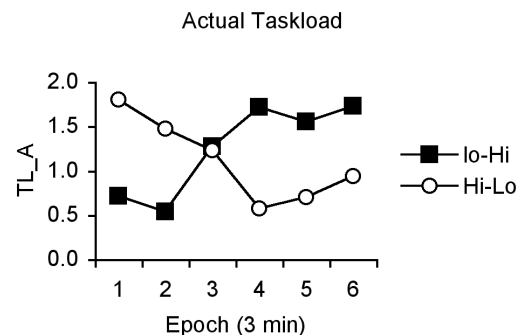


Figure 2. Mean actual taskload experienced by the participant by 3-minute epochs (by eq. 1); the shift from low to high and high to low taskload after about 9 minutes into the block is clearly visible.

Performance. The main performance variable was the timeliness in resetting the timers in the experimental task. According to our hypothesis, participants who were able to main-

tain good temporal awareness should have been able to reliably attend the next open WO, manifested in short elapsed time from opening of the WO to the first key press to reset the timer and small variance in these times. These aspects of timing performance were examined by fitting a known distribution to the data and comparing the parameters of the distributions between the experimental conditions. The Weibull distribution was chosen due to its versatility and overall good fits to the data.

As taskload increased the participants became increasingly late in their resetting task and also exhibited substantial variability in their performance (Fig. 3). The results showed a significant positive relationship between taskload and TFA. Consequently, success rate, measured by the number of timers reset within the WO, was significantly different between the two taskload conditions, decreasing from 90% to 80% as taskload increased. These phenomena may attest to a loss of temporal awareness under high taskload and a lapse into a reactive mode of operation.

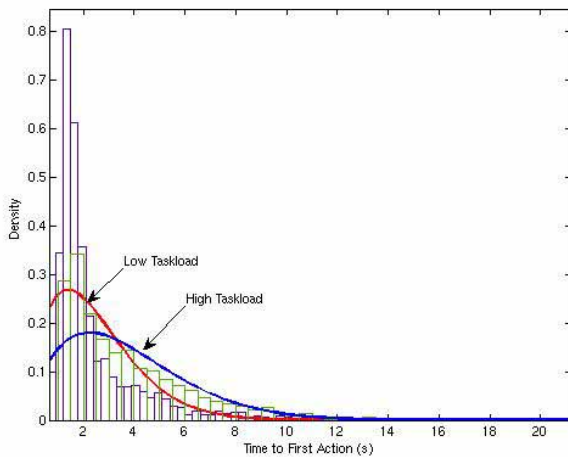


Figure 3. Shift in the time elapsed from opening of WO to first action in a timer resetting task in Experiment 1 as taskload increased. This shift from low to high taskload is evident in the means (from 2.51 to 3.79), variances (from 2.83 to 5.97) and the distributions' scale and location parameters (from 2.79 to 4.22 and 1.52 to 1.58, respectively).

EXPERIMENT 2

Method

Participants. A total of 11 retired FAA controllers and supervisors assigned to the FAA Technical in Atlantic City, NJ, volunteered to participate in the experiment. All participants were male, with a mean age of 55.64 years (range 38 to 66 years) and with a mean experience as a controller of 23.45 years (range 11 to 33 years).

Apparatus. The experimental apparatus was a custom-built ATC simulator, allowing for accurate and reliable recording and time stamping of all specified events concurrently as the program ran, and flexibility in the "scripting" and timing of the scenarios in each pane by the experimenters. The simulation program was ran on two laptop computers, and it mimicked the display system replacement (DSR), including

data link (DL) capability, allowing for more accurate timing of participant interactions with the DL interface than would have been possible via voice communications.

Experimental task. The experimental task mimicked the job of air traffic controllers. The participants viewed air traffic scenarios on four separate quadrants, or panes, on a single computer display. The scenarios could be viewed only one at a time by moving a cursor to the desired pane (Fig. 4). This task balanced the requirements of realism and experimental control, and it allowed for accurate measurement of times of the different events unfolding in the experimental scenarios as well as timing of the participants' actions in response to them.

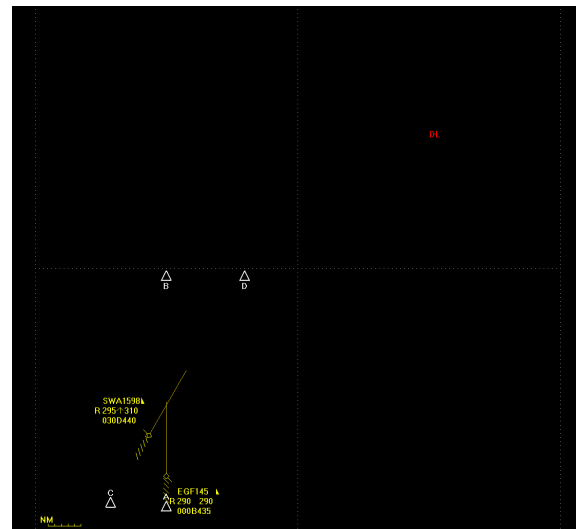


Figure 4. Screenshot from Experiment 2. The cursor is in the lower left-hand pane and the two aircraft in that pane visible; note a DL request alert in the upper right-hand pane

Design. The primary independent variable was taskload, which was manipulated by both the time available and the time required to perform the tasks, which in turn were manipulated through several other variables over which the experimenter had complete control. Note, however, that control over these variables was constrained by the participants' actions after the onset of the experiment, that is, the eventual sequence and timing of the tasks depended on individual participants' different time management skills and strategies as well as other individual performance differences. One of the explicit objectives of this experiment was to develop measures that can be used to quantify taskload as a predictor variable for analyses of performance data.

Time required was manipulated primarily by differential difficulty of conflict situations, based on findings of Rantanen and Nunes (in press) and the number of tasks to be performed during the experimental scenarios. The actual time required to execute the control actions necessary to resolve conflicts, respond to pilot requests, and accept and initiate handoffs, were determined by a pilot study. Time available (TA) consisted of the individual windows of opportunity (WO) for each task encountered per trial. The ratio of time required and time available was the basis of the definition and computation of nominal taskload, calculated at the outset of the scenarios. A

total of 31 different scenario files were created and nominal taskload was calculated for each scenario by summing the times required for each action required and dividing this by the sum of times available (i.e., WO) for each task. These individual nominal taskload scores for each scenario allowed for ordering of the scenarios in the experimental blocks to manipulate taskload within the block as desired (i.e., low-to-high, and high-to-low).

Design. The basic design of the experiment was a 3 (taskload, Low, Transition, High) x 2 (order Lo-Tr-Hi, Hi-Tr-Lo) x 2 (replicates) factorial design. In the analyses only low and high taskload conditions were considered, the transitional scenarios split between the two conditions. Four scenario files, one file per quadrant (pane) on the display, started the experimental blocks. At the end of each scenario, a new scenario files filled the pane. An experimental block was comprised of 4 (panes) x 5 scenarios, which followed each other in a seamless sequence.

Dependent variables. The main performance variable was time to first action (TFA), which was defined as the elapsed time from opening of WO to the time the task was performed, calculated by the first action on the task (e.g., mouse click on a flyout menu).

Results

Taskload. We wanted to determine the actual taskload as influenced by the participants’ control actions and strategies, as we anticipated the actual taskload to be different from the nominal one determined from the outset of the experiment. An index of taskload (TL_A) was provided by the equation

$$TL_A = \frac{n(TR_{avg})}{TE} \tag{2}$$

where n is the total number of tasks present in an epoch and TR_{avg} the average time required to perform these tasks. The TE is the duration of the epoch, in this case 300 s (5 min). It is acknowledged that many tasks had zero time required to perform them, for example, acceptance and initiation of handoffs and transfer of communication only required a single mouse-click. Furthermore, it is clear that physically performing the task, by keyboard entries or clicking through menus with a mouse, only constitutes a small fraction of the total time required to perform the task, that is, the overt actions do not reveal planning and decision-making processes, which almost certainly require most of the controller’s time. Nevertheless, multiplication of the time required by the number of tasks compensate to some degree the very short (i.e., 0) performance times in an epoch, and indeed this index showed clear differences between the different taskload conditions (Fig. 5). The differences between taskload were significant (two-sample t-test, $p < .05$) for all but the transition epoch.

Performance. The results were analyzed for all 6 different ATC tasks performed by the participants; however, results for only conflict resolution are reported here. Analysis of TFAs for conflict resolution task showed very clear patterns according to our hypothesis: As taskload increased, the participants were less likely to act on conflict resolution at an earliest op-

portunity than under low taskload conditions (Fig. 6). On average, in low taskload conditions the participants first acted on an impending conflict 59 s after the WO opened, but waited nearly twice as long under high taskload.

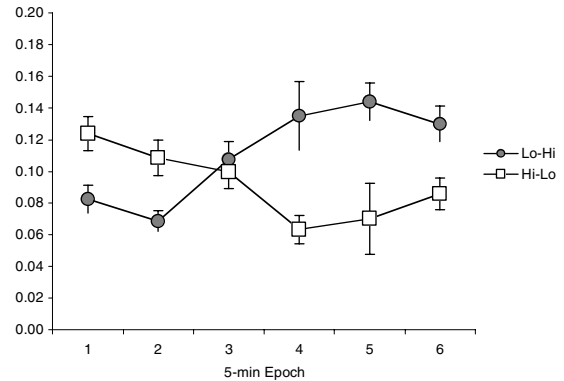


Figure 5. Mean actual taskload index values (by eq. 2) by taskload condition and 5-minute epochs; the switch between taskload conditions in the middle of the block is apparent.

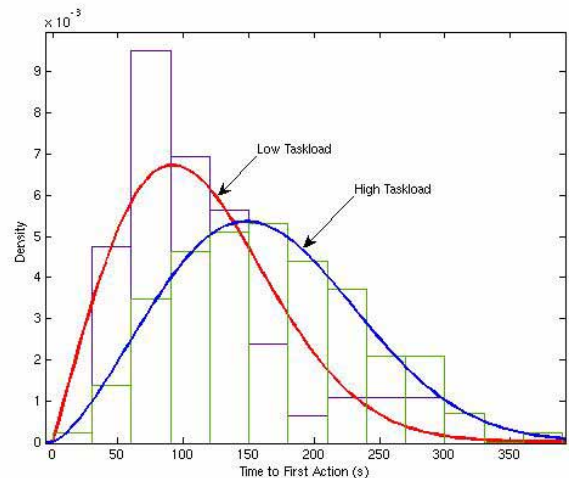


Figure 6. A very similar pattern to Experiment 1 is evident in the time to first action distributions from Experiment 2 as well, although the tasks were substantially different between the experiments. Shift from low to high taskload resulted in a shift in mean TFA (from 113.87 to 163.93 s), an increase in variance (from 3464.38 to 5144.75), and shifts in the scale and shape parameters of the Weibull distributions (from 128.51 to 184.43 and from 2.02 to 2.43, respectively)

DISCUSSION

Time has a long history as a means to investigate cognitive processes, timing data are relatively easy to obtain under both experimental and naturalistic conditions, and time is a variable that is common to the human, the task, and the environment. Time offers thus a common unit of measurement of human performance in the context of the task, and can be used to infer the goodness of the temporal dimension of the operator’s mental model of the task or system being controlled. Measurement of taskload and performance in temporal terms

may also allow for bracketing and making inferences about mental workload, which is not directly measurable.

The results of experiment 1 demonstrate that the differences in performance resulting from high and low taskload—which was verified in the actual conditions as the tasks were performed—can be represented by specific changes in performance distributions. Here, an increase in taskload was shown to increase both the mean and variance of the TFA distribution. The results of Experiment 2 showed patterns that were remarkably similar to those from Experiment 1, despite the very different experimental tasks. The observation that an increase in taskload has a robust effect across two experiments with different degrees of ecological validity serves to make a case for proceeding from highly controllable experimental settings to high-fidelity simulations and finally data obtained from operational settings. Development of a cognitive model of operators' timing performance requires that all available task performance data are meticulously recorded, and this is only possible from controlled experiments such as those above. After such a model is developed, however, we may look to more complex settings (e.g. high-fidelity simulations) for the parameters and patterns observed in these simpler experiments.

ACKNOWLEDGMENTS

This work was supported by the FAA (Coll. Agreement No. 02-G-019; Dr. Carol Manning, tech. monitor) and NASA and Micro Analysis and Design (Contract No. 7400.005.01, Parimal Kopardekar from NASA and Ken Leiden from MAAD, tech. monitors). Views expressed herein are those of the authors and do not necessarily represent official NASA, FAA, or MAAD positions. We express our appreciation to Ben Willems and Pamela Della-Rocco at the FAA William J. Hughes Technical Center for their invaluable assistance in running Experiment 2. Special thanks are due to Sharon Yeakel for programming the experimental simulator and data processing tools.

REFERENCES

- De Keyser, V. (1995). Time in ergonomics research. *Ergonomics*, 38(8), 1639-1660.
- Decortis, F., De Keyser, V., Cacciabue, P. C., & Volta, G. (1991). The temporal dimension of man-machine interaction. In R. S. Weir & J. L. Alty (Eds.), *Human-Computer Interaction and Complex Systems* (pp. 51-72). Academic Press.
- Gawron, V. J. (2000). *Human Performance Measures Handbook*. Lawrence Erlbaum.
- Gunning, D., & Manning, M. (1980). The measurement of aircrew task loading during operational flights. *Proc. 24th HFES Mtg.* (pp. 249-252). HFES.
- Hart, S.G. (1986) Theory and Measurement of Human Workload. In J. Zeidner, (Ed.), *Human Productivity Enhancement: Training and Human Factors in Sys. Design*, v.1. Praeger.
- Hollnagel, E. (1993). *Human reliability analysis, context and control*. Academic Press.
- Hollnagel, E. (1998). *Cognitive reliability and error analysis method: CREAM*. Elsevier.
- Meshkati, N. (1988) Toward Development of a Cohesive Model of Workload. In Hancock, P. A. and Meshkati, N. (Eds.) *Human Mental Workload*. Elsevier,
- Moray, N., Dessouky, M. I., Kijowski, B. A., & Adapathya, R. (1991). Strategic behavior, workload, and performance in task scheduling. *Human Factors*, 33(6), 607-629.
- Ogden, G. O., Levine, J. M., & Eisner, E. J. (1979). Measurement of workload by secondary tasks. *Human Factors*, 21(5), 529-548.
- Raby, M., & Wickens, C. D. (1994). Strategic workload management and decision bias in aviation. *Int'l J. of Av. Psychology*, 4(3), 211-240.
- Rantanen, E. M., & Nunes, A. (In press). Hierarchical conflict detection in air traffic control. *Int'l J. of Av. Psychology*.
- Sperandio, J. C. (1971). Variation of operator's strategies and regulating effects on workload. *Ergonomics*, 14(5), 571-577.
- Sperandio, J. C. (1978). The regulation of working methods as a function of workload among air traffic controllers. *Ergonomics*, 21(3), 195-202.
- Stone, G., Gulilck, R. K., & Gabriel, R. F. (1984). *Use of task/timeline analysis to assess crew workload* (Douglas Paper 7592). Douglas Aircraft Co.
- Tulga, M. K. and Sheridan, T. B. (1980). Dynamic decisions and workload in multitask supervisory control. *IEEE Transactions SMC-10*(5), 217-232.
- Waganaar, W. A., & Sagaria, S. D. (1975). Misperception of exponential growth. *Perception and Psychophysics*, 18, 416-422.
- Wickens, C. D. (2001). Attention to safety and the psychology of surprise. *11th Int'l Symposium of Av. Psych.*, Columbus, OH.
- Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance* (3rd ed.). Prentice-Hall.