POPCORN: A Supervisory Control Simulation for Workload and Performance Research

Sandra G. Hart
NASA-Ames Research Center
Moffett Field, CA

Vernol Battiste and Patrick T. Lester
San Jose State University
San Jose, CA

ABSTRACT

A multi-task simulation of a semi-automatic supervisory control system was developed to provide an environment in which training, operator strategy development, failure detection and resolution, levels of automation, and operator workload can be investigated. The goal was to develop a well-defined, but realistically complex, task that would lend itself to model-based analysis. The name of the task ("POPCORN") reflects the visual display that depicts different task elements milling around waiting to be released and "pop" out to be performed. The operator's task was to complete each of 100 task elements that were represented by different symbols, by selecting a target task and entering the desired a command. The simulated automatic system then completed the selected function automatically. Task difficulty, operator behavior, and experienced workload were varied by manipulating: (1) the number of elements per task; (2) the number of discrete tasks; (3) the penalties for lagging behind the system; (4) task schedule; and (5) payoff structure for performing or failing to perform task elements. Highly significant differences in performance, strategy, and rated workload were found as a function of all experimental manipulations (except reward/penalty). In addition, a proposed technique for reducing the between-subject variability of workload ratings was described and applied successfully. The first simulation conducted with this task defined a range of scenarios that imposed distinctly different levels of workload on operators and resulted in different levels of performance and operator strategies.

INTRODUCTION

The introduction of computer aiding, artificial intelligence, and automation into advanced systems has changed the roles of human operators. Their primary functions have become scheduling, monitoring, decision making, and planning rather than direct mechanical control. Furthermore, the interfaces between the operators and the systems that they control have become indirect, periodic, and discrete rather than direct and continuous as computers are placed between the human operator and the mechanical system.

Automation is a generic term for replacing human actions by human decisions executed by machines and for accomplishing clusters of related tasks by simple executive commands (refs. 1, 2). Often, the decision to automate some or many system functions stems from a desire to enhance system capabilities without overloading operators. Alternatively, it is introduced to allow existing crewmembers (or a reduced number of them) to perform additional tasks or operate in environments in which they could not function without aiding. In the past, automation has been provided to reduce the physical workload of

activities, a goal that has been accomplished with great success. However, a potential consequence of adding automation could be a substantial increase in mental workload to replace the reduced physical workload, due, in part, to the added burden of supervising or monitoring the automation itself. Such a tradeoff between physical and mental workload has been inferred rather than proven, however, because mental processes are difficult to observe or quantify directly. Thus, there is an increasing need to monitor, measure, define, and control whatever "mental workload" is in order to keep it within the capabilities of human operators.

In order to develop valid and sensitive measures of mental workload and performance, standardized primary tasks are needed to test candidate measures. These tasks must impose controlled levels of load with the dynamic decision-making and task-selection activities typical of current and future man-machine systems. Procedures for performing combinations of subtasks under normal and failure conditions should simulate the complexities and alternative solutions typical of advanced systems and computer aiding might be provided to assist operators with specific functions. Manual control issues may receive less emphasis, as the focus of the research will be on activities that are more typical of automated systems. The interface between man and machine will continue to be an important issue, however. With such tasks, theoretically and practically interesting topics, such as training, development of performance strategies, and the subjective experience of workload, could be investigated and models of human performance appropriate for multi-task, supervisory control systems developed.

Laboratory tasks that impose controlled levels of load across a range of functions typical of advanced systems (refs. 3, 4) have been used in many research efforts. These tasks may be manipulated and controlled with precision and predictions about performance may be made from a sound theoretical point of view. One disadvantage, however, is that the workload imposed by a realistically complex combination of such tasks may be substantially different than the sum of the workloads imposed by the components individually. For example, depending on the strategies selected and the degree to which groups of related subtasks are performed automatically, subjective experiences and objective performance might be significantly different than would be predicted from single-task performance.

A multi-task dynamic simulation was developed to represent the environment in which decision makers responsible for semi-automatic systems work (ref. 5). It involved a computer display of tasks (represented by boxes) which appeared according to different random schedules and moved toward a deadline. Operators could perform only one task at a time and were required to develop different performance strategies to accomplish specific experimental scenarios. Interarrival rates, the time until tasks reached the deadline and the time required to perform them, the number of tasks, and the "values" assigned to them were manipulated. The goals of the research were to develop an objective index of task load and to model subject's behavior. In a later study, (ref. 6) three task variables (interarrival rate, task duration, and number of tasks) were manipulated to determine their relative contributions to the subjective experiences of workload. It was found that the number of tasks to be processed per unit of time was the dominant factor.

A similar simulation was developed to extend the optimal control model methodology to characterize human monitoring, information processing, and task

selection in a dynamic multi-task environment (ref. 7). Five stylized tasks that varied in value, processing time, and velocity competed for the operator's attention. The decision process was dynamic, as new tasks with different characteristics continued to arrive, the opportunity window to perform available tasks shrank, and unperformed tasks reached the deadline.

The design of the current simulation was derived, philosophically, from the earlier simulations (refs. 5, 6, 7), however it expanded on them by increasing task complexity, incorporating dependencies among task elements, varying task attributes as a function of human decisions, and providing an extensive procedural structure. Its name, "POPCORN", reflects the appearance of the task elements waiting to be performed (they mill around and then "pop" out of the computer-displayed containers). The operator's job is to decide which tasks to do and which procedures to follow based on an assessment of the current and projected situation, the urgency and difficulty of the tasks, and the reward or penalty for performing or failing to perform them. The system is controlled by operators who select functions to be performed by automatic subsystems (barring preprogrammed "hardware" failures or operator error).

The first study conducted with this simulation was designed to examine the effects of a variety of phenomena typical of supervisory control tasks on operator strategies, performance, and the workload they experience. The goal was to establish task scenarios that would present operators with predictable variations in imposed workload (by varying scheduling, the number of elements per task, time pressure, and availability of tasks for performance) and to provide opportunities for operators to adopt different strategies (depending on whether they were leading, lagging, or level with system demands). A variety of control functions were simulated to provide alternative solutions to different combinations of circumstances. Different penalties for procrastination were invoked whenever an operator failed to meet task schedules and deadlines: (1) Imposition of additional operations to perform on delayed tasks, (2) Loss of points for performing deferred tasks, and (3) Transfer of delayed task elements to a penalty box where immediate performance was required. In addition, the longer a task element remained unperformed, the faster it moved in half of the scenarios, so that less and less time was available for its performance when the operator did attend to it. Interarrival rates were varied so that each task could be completed by a trained operator before another was scheduled. Because the acceleration function made tasks available for performance more quickly, the scheduled arrival times between accelerated tasks was less than it was between fixed-rate tasks to maintain a steady flow of activities. The interval of time during which a task element could be performed (its "opportunity window") was, therefore, influenced by the presence or absence of acceleration and the number of elements per task. The minimum time to perform a task was fixed by the speed at which elements exited from the boxes and the number of elements per task. The maximum time to perform a task was defined by the scheduled interval between successive tasks per box.

Performance on the primary task was evaluated by examining the scores obtained under each experimental condition, to complete it, and the number of errors. Strategies were evaluated by analyzing the functions that were selected. The effect of experimental manipulations, operator strategy, and performance on the subjective experiences of the operators was assessed by responses to rating scales presented immediately after each scenario.

The workload imposed by the tasks was determined by a weighted combination of operators' evaluations of 10 relevant factors. These evaluations were related not only to the experimental manipulations, but to the operators' strategies, performance and pre-existing biases about what aspects of a situation contribute to variations in experienced workload, as well. Ratings on many different scales were obtained because workload is thought to be a multi-dimensional construct (refs. 1, 8, 9). Factors such as the difficulty of the task imposed on the operator, the physical or emotional stress experienced, time pressure, and the amount of effort exerted have been suggested as potential components. In addition, there may be individual differences in which aspects of a task are considered to be relevant to the level of workload experienced (refs. 10, 11). For some individuals, the difficulty of a task may completely define the workload experienced. For others, the physical or mental effort exerted may create the conscious experience of workload. For yet others, feelings of stress, frustration, or fatigue that accompany task performance may affect the conscious experience of workload. Tasks that are performed successfully may be experienced as having low workload whereas those that are performed poorly may be equated with high workload (regardless of the level of effort applied in either case).

A technique for combining ratings on different workload-related dimensions (each weighted to reflect its subjective importance to individual operators) was developed and tested in this and other recent studies (refs. 12, 13). Nine factors that have been found to provide the most complete description of operators' experiences were the basis for the weighting procedure. Unlike other methods of extracting subjective biases from workload ratings, such as the Subjective Workload Assessment Technique (SWAT) (refs. 11, 14), this technique allows a weight of zero to be given to a dimension that is considered to be irrelevant and incorporates a sufficiently broad range of dimensions to characterize the biases of most individuals. In addition, it does not require an abstract prediction of the possible effects of complex combinations of different levels of different dimensions as does the SWAT technique.

METHOD

Subjects

Eight male general aviation pilots served as paid participants in the experiment. They ranged in age from 22 to 35 years. Two additional male subjects participated in a pilot study.

Equipment

The simulation was programmed on a Digital Equipment Corporation PDP-11/40 computer and an Evans and Sutherland Picture System I. The display was presented in a 25.60 cm square area on a Xytron black and white monitor. Operators interacted with the system by positioning a stylus on a magnetic response pad and entering selections by depressing the tip of the stylus. The 25.6 cm display area was projected onto a 5.1 cm area on the response pad (an area approximately equivalent to the dimensions of the display depicted in Figure 1). The operators rested their right arms on the response pad and were able to reach every function with minimal hand movements. The response area was delimited by cut-out area of a 0.6 cm thick plexiglass overlay on the pad.

The experiment was conducted in a secluded area of a computer room with dim lighting levels and no distractions. The operators were seated at a small table that contained the stylus and response pad and the operations manual. The display was located immediately in front of the subjects at a distance of approximately 1.0 m.

Experimental Task

Basic Functions

The information, control functions, and displays for the simulated system were presented on a computer display. (Figure 1) The five task types were each represented by a unique symbol (*, +, -, #, =), consistently mapped so that only one symbol appeared in each box. Five types of tasks that occurred several times each were included so that operators had to shift their attention from one to another, as they do in operational environments. Each task served as an abstract representation of a different type of function (e.g., communications, navigation, monitoring, checklists, and autopilot control) that might be performed in a complex system, such as a modern aircraft. In the current experiment, the values assigned to elements from each box, the functions and time required for performance, and element rates were identical for all tasks within each scenario, however these variable are under experimental control and different levels and combinations of levels could be selected for subsequent simulations.

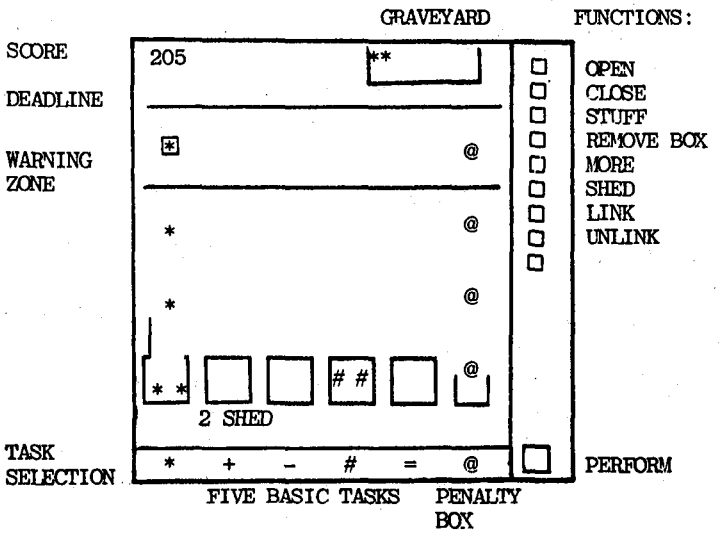


Figure 1. The POPCORN simulation display.

If a new task was scheduled to enter an occupied box, elements from the existing task were transferred to a "penalty box". This marked the end of the window of opportunity to perform the remaining task elements for score points. The operator's goal was to perform as many tasks as possible, maximizing the score and minimizing the time per scenario.

The initial decision to ready a task for performance was made by touching the symbol located immediately below the selected box with the stylus (the SELECT function). Task selections remained in force until a different task was selected; only one task could be operated upon at a time in the basic system. The functions that could be performed on any task were displayed on the right side of the display. Functions were generally momentary; each actuation caused the selected function to be applied one time to the current task. The operator's job was to decide which functions to apply to which tasks. Their actions prompted automatic subsystems to effect the selected functions, much as when a pilot selects a new altitude or navigational point, enters it into a navigation computer, and an autopilot achieves the desired change.

Task elements arrived at scheduled times and milled around in their boxes until they were SELECTed. Once the lid of a box was removed (by the OPEN function), task elements streamed out in a vertical line at a rate determined by their initial velocity (12.5 cm/sec) and the acceleration function for that scenario (either 0 or 1.52 cm/sec/sec). One box could opened at a time or several could be left open. Elements of the currently selected task were performed by touching the PERFORM key area. Each actuation caused the topmost element in the stream of task elements to disappear and the score to be incremented by five points. The maximum possible score for any scenario was 500 (5 points each for 100 task elements).

Boxes could be closed after each task was completed (in anticipation of the arrival of the next task) or with elements remaining to be performed. If any elements were actively exiting from a box, the operator had to place them back in the box (by actuating the STUFF command) before selecting the CLOSE command. By selectively opening and closing one or more boxes, operators could control the number of task elements available for performance and by rapidly selecting and performing one task then another, several tasks could be completed in parallel. An alternate strategy was complete each task, one at a time, before going on to the next. The optimal strategy differed as a function of the schedule and circumstances for each scenario.

Penalties and Procedures for Lagging Behind

If operators waited too long to perform a task element after it had left its box, the symbol moved into a "warning zone" where each element was surrounded by a square symbol. The task could still be performed with no loss of score, but at the cost of an extra procedure. This represents the additional problems encountered in operational settings when operators wait too long to finish a task once it has been started. In order to perform task elements in this zone, the task must be SELECTed, the warning box removed from the symbol (REMOVE BOX), and the topmost task PERFORMed. This two-stage process had to be repeated for each successive task that entered the warning zone. The most efficient strategy was one that allowed tasks to be completed before they entered this area. If tasks did enter this area, however, the operator could either elect to perform the two-stage REMOVE BOX/PERFORM procedure or STUFF the elements back in their original boxes, in effect resetting that task. If a task element was not performed by the time it reached the "deadline", its symbol was placed in the "graveyard" and no points were scored.

Since more than one task of each type was scheduled per box, operators had to complete each task before the next one arrived or the unperformed elements

from the previous task would be transferred to the penalty box. Once in the penalty box, task elements lost their identity (they were represented by "@"), they had to be performed immediately (the box had no lid), and no score points could be gained by performing them (although a five-point penalty was levied if they were not performed). Thus, once operators had begun to lag behind the system to the point that tasks were being transferred to the penalty box, they had to shift to a reactive strategy in which they could accomplish no more than preventing additional loss of points.

If operators decided that things were out of control, two strategies were available: closing some or all of the boxes or shedding the elements in one or more boxes. If the SHED function was selected, the elements remaining in the selected box could no longer be performed (thereby losing the potential for gaining those points), however the five-piont penalty for unperformed tasks could be avoided. This function was provided to allow the operators to elect a strategy available in operational settings (e.g. the decision to ignore certain tasks when loading levels are perceived as excessive).

Functions that Allow Operators to Lead the System

If operators wished to complete tasks ahead of schedule, they could request MORE tasks. For half of the scenarios, only two tasks of each type (with 10 elements each) were scheduled, limiting the opportunity to use this command. For the remaining scenarios, however, five tasks of each type (with four elements each) were scheduled, providing many opportunities to select it.

One form of automation is the performance of multiple related tasks by a single command. This type of activity was simulated with the LINK and UNLINK functions. If LINK was selected, elements from two of the five basic tasks could be acted upon with a single command; every function applied to one task was applied to the other so that tasks could be completed twice as fast. There were limits to the utility of this function, however. If one task was completed before the other, or if elements from one task entered the warning zone, the tasks had to be UNLINKed to be completed.

Experimental Variables

Two levels of each of four experimental variables were combined to create sixteen scenarios. The variables were: (1) reward and penalty for performing (or failing to perform) subtasks, (2) task schedule, (3) number of elements per task, and (4) the consequences of delaying task performance. The experimental design may be seen in Figure 2. The payoff structure was manipulated to determine the impact of penalties (decrements in score) for failing to perform subtasks on operators' strategies and experienced workload. Five points were given for each task element performed within the appropriate amount of time. In half of the scenarios, there was no additional penalty (other than loss of score) for failing to perform tasks (+5/0). In the other scenarios, an additional five-point penalty was levied for each unperformed task element (+5/-5).

Two task schedules were imposed: (1) **MASSED** (tasks appeared simultaneously in the five boxes whenever new tasks were scheduled to appear); and (2) **STAGGERED** (tasks appeared at different, predetermined times in each box). This manipulation was included to assess the effect of organizational complexity.

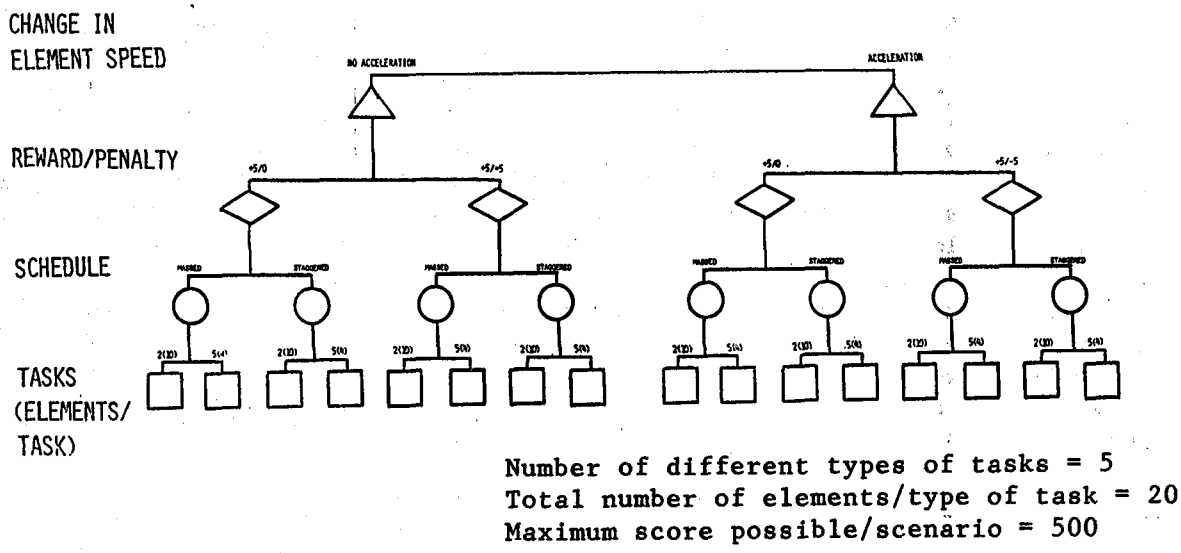


Figure 2: Design of the 16 experimental scenarios.

The scenarios were designed to provide operators with a predictable and reasonable set of requirements; the intervals between successive tasks were sufficiently long for trained operators to complete one task before the next arrived. The number of elements/task (analogous to the time-to-perform tasks in the earlier studies) and the presence or absence of acceleration (accelerated tasks exited more quickly, thus the opportunity for performing them occurred more often, even though the opportunity window for their performance was reduced by speed of their movement toward the deadline) were considered in computing the task schedules. The 16 schedules are depicted in Appendix A.

The number of task elements per scenario was constant (20 elements for each of 5 task types), however, the way they were grouped was varied: (1) Two tasks with 10 elements each [2(10)] per box, or (2) Five tasks with five elements each [5(4)] per box. Each element took the same amount of time to perform, thus, tasks with many elements took longer to complete than those with few elements, however there was less time lost switching among tasks and the schedule was less complicated with the larger tasks. This variable was included so that strategies and performance differences resulting from the tradeoff between task complexity (e.g. elements/task) and number of discrete tasks (10 or 25) could be evaluated.

The longer operators waited to perform tasks, the more urgent they became. In eight of the scenarios (**ACCELERATION**), urgency was simulated by accelerating the movement of task elements in the boxes as long as they remained unperformed. In the other eight scenarios (**no ACCELERATION**), task elements moved at a constant rate that was so leisurely that it inhibited well-trained operators from performing tasks as quickly as they could. The accelerations were 0 and 1.53 cm/sec/sec for the **no ACCELERATION** and **ACCELERATION** conditions, respectively. Although acceleration substantially increased the time pressure under which operators worked, accelerated tasks could be completed more rapidly once a box was opened (a potentially positive factor).

Rating Scales

Operators rated their experiences along 10 workload-related dimensions: task difficulty, time pressure, performance, mental effort, physical effort, frustration, stress, fatigue, type of activity, and overall workload. The scales were presented on the display immediately after each scenario. A stylus was used to position a cursor at the desired scale value. Each scale was a 11.0-cm vertical line labeled with a title (e.g. "MENTAL EFFORT ") and bipolar descriptors (e.g. "EXTREMELY HIGH/EXTREMELY LOW"). Numerical values were assigned to the selected scale positions with a range from 0 to 100 during data analysis.

Two estimates of workload were obtained: a direct rating provided by the operators (with the "OVERALL WORKLOAD" bipolar scale), and a combination of the remaining nine scales weighted to reflect the importance placed on each factor by each subject. The relative importance of the nine factors (e.g. the weights) was determined by a pretest in which the 36 possible pairs of the nine factors were presented one at a time. The member of each pair that was considered to be most relevant to workload by that subject was recorded. The number of times each factor was selected was computed; the possible values each factor might have ranged from 0 (the dimension was not at all relevant) to 8 (it was more important than every other factor), with a total possible sum of 36.

Procedure

A brief introduction that described the purpose of the simulation and the research to be performed with it was read to the participants. An operations manual was given to them to read while the experimental manipulations were described and demonstrated. A one-hour training session was provided to familiarize them with the tasks, equipment, and procedures.

At the end of the training period, the 16 experimental scenarios were presented in a different random order to each subject. A description of the upcoming scenario and a schedule of task arrival times was provided before each scenario and the 10 rating scales were presented following each scenario. At the conclusion of the experiment, the operators rank ordered the four experimental variables with respect to the impact that they felt each had had in influencing the level of workload. The experiment lasted approximately 5 hr, with a long break in the middle and shorter breaks between scenarios.

RESULTS AND DISCUSSION

A three-way analysis of variance for repeated measures was the primary statistical procedure applied to the dependent measures. Analyses were performed on 12 measures of performance (e. g. score, task duration, and inappropriate function selections), 10 measures of operator behavior (e. g. function selections), and 11 subjective ratings (e. g. 10 bipolar scales and the combined weighted workload scale). In addition, the correlations among scores, task durations, selected measures of behavior, and the weighted workload rating were computed. Differences in performance, operator behavior, and subjective experience were examined on a subject-by-subject basis to determine the association between operator strategies and behavior, and the resulting performance and subjective experiences.

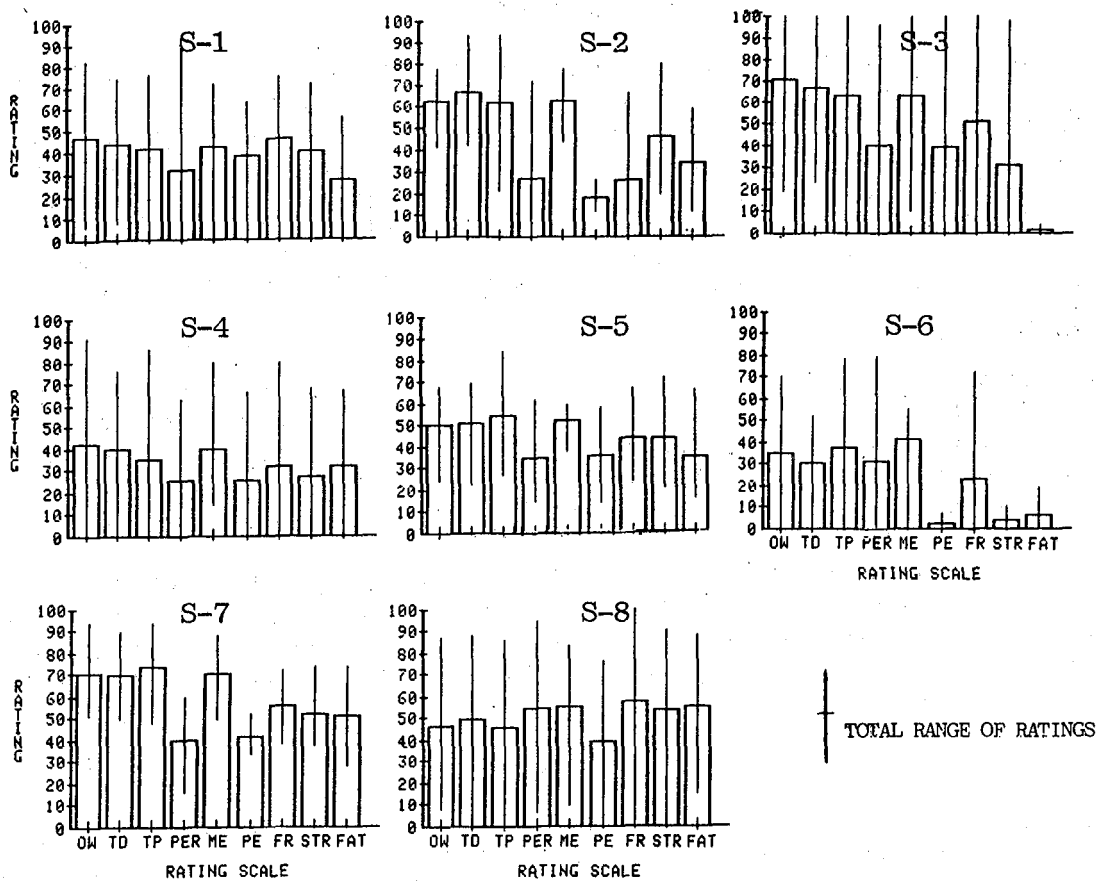


Figure 3: Average and range of bipolar ratings obtained from each of 8 experimental subjects across 16 experimental conditions.

Overview of Dependent Measures

Bipolar Ratings

The average and range of ratings given by each subject across experimental conditions may be seen in Figure 3. It is clear from an examination of the data that individual subjects differed in the magnitudes of ratings given from one scale to the next and also in the range of the rating scales used within and between scales. For example, the between-subject standard deviation (SD) of overall workload ratings across conditions and subjects was 25.5, more than half of the mean value of the rating (52.7).

Workload Weights

The relative importance each subject placed on the nine workload-related factors may be seen in Figure 4. As expected, the subjects disagreed about how much influence the different factors were predicted to have on their experience of workload. It is precisely because of this expected difference of opinion that the preliminary test was conducted, however, to facilitate the statistical removal of this source of between-subject variability from the combined bipolar ratings. In general, Time Pressure, Own Performance, Frustration, and Stress were each selected as more relevant than the other

items more than half of the time. Physical Effort was rarely selected as a relevant variable, and Task Difficulty and Mental Effort (which are usually considered to be important) were just moderately important for this group. For each of the 16 experimental conditions, the nine original bipolar scales, multipled by the appropriate weight, were combined and averaged for each subject. The resulting weighted workload estimate could be conceptualized as the combined area of a bar graph with nine variables; the width of each bar determined by the importance of that factor to the individual (the weight) and the height of each bar determined by the subjective magnitude of the factor in a given experimental condition (the bipolar ratings). (see Appendix B for examples by subject and experimental condition)

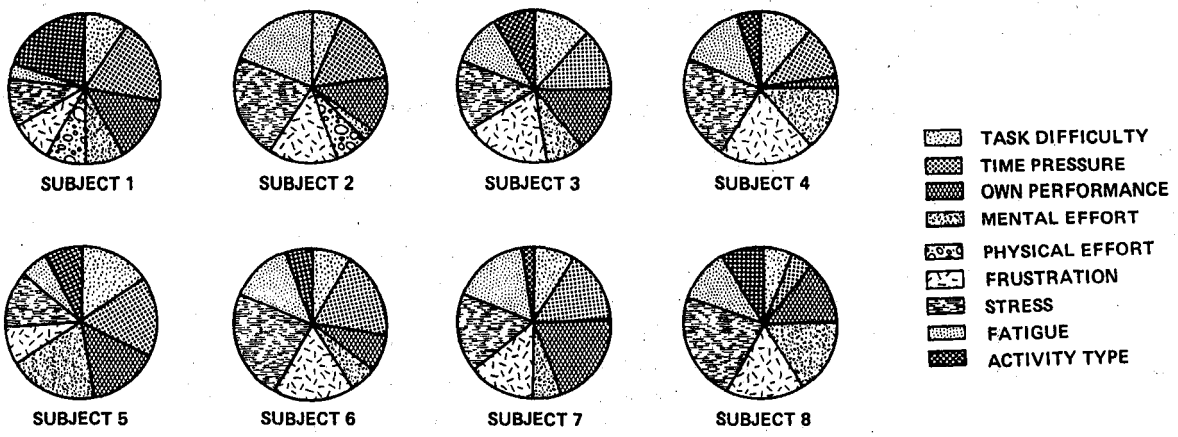


Figure 4: Relative importance to the subjective experience of workload assigned to each of 9 factors by each subject (n = 8)

The magnitude of the weighted workload estimates was less, on the average, (43.4 versus 52.7) than was the overall bipolar rating of workload, however the relationships among the experimental conditions was the same for the two estimates of workload, as illustrated in Figure 5. This reduction in magnitude is expected, as a single rating of overall workload represents the subjective total of whatever factors the individual considered were relevant to an experience of workload, whereas the weighted combination of ratings is statistical average of all of the factors. The benefit of performing the weighting procedure was that the between-subject SD was reduced for every experimental condition taken one at a time. Overall, the reduction 17% (from 25.5 to 21.3). Using a simple linear combination of the nine unweighted ratings also resulted in reduced between-subject variability (with the relationships among experimental conditions maintained), but the reduction was considerably less.

The reduction in between-subject variability achieved with the weighting procedure was less than has been found in other recent applications (see, for example, refs. 12, 13). In other applications, between-subject variablity was reduced by as much as 50% overall. Since the participants in the current study were in greater accord about the relative importance of the different factors than has been found for other groups of subjects, the influence of individual differences in the definition of workload was not as great in the current study as in the others. This weighted workload rating will be used as the primary measure of subjective workload for the remainder of the study.

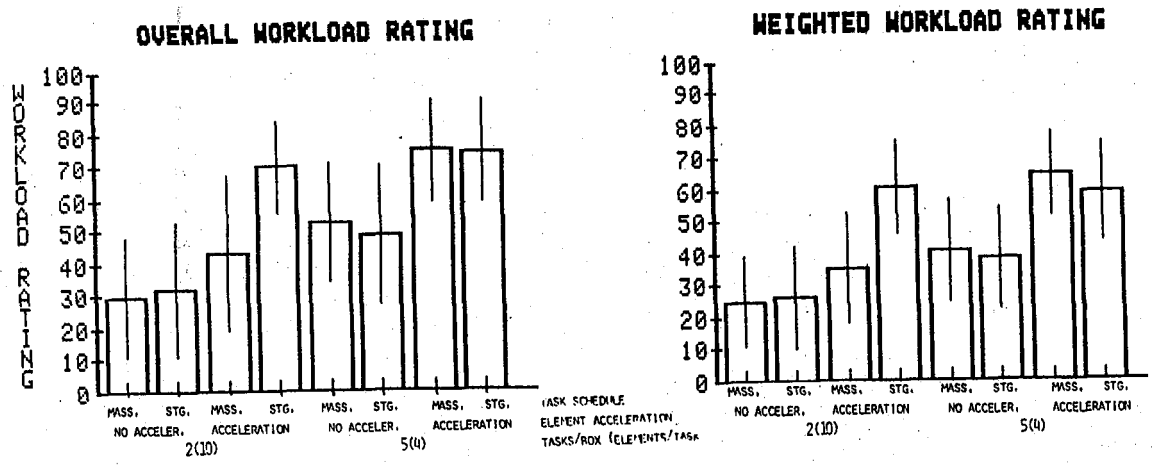


Figure 5: Unweighted and weighted ratings of overall workload by experimental condition (n=8; I depicts +/-1 SD)

Score

The scores ranged from a maximum of 500 to a low of 55. The grand mean was 375 (SD = 125). Thus, the 16 combinations of experimental variables did produce the desired range in performance levels across subjects and scenarios. (Figure 6) On an individual basis, the average scores obtained by individual subjects across experimental conditions ranged from 409 to 321. High scoring subjects performed more consistently than low scoring subjects, and there was a highly significant correlation between score and rated workload (rxy = -0.71), high scores being associated with low workload ratings.

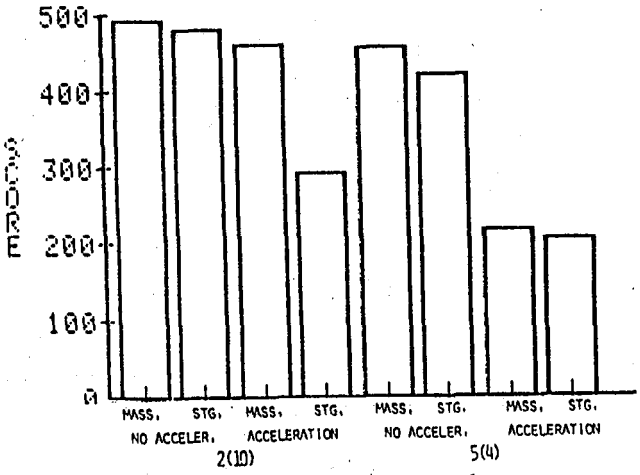


Figure 6: Average scores by experimental condition. (n = 8)

Scores were examined in the order that the scenarios were presented to determine whether or not there was a continuing improvement in performance from the beginning to the end of the experiment across the counterbalanced experimental conditions. No such improvement was found, indicating that the training given was sufficient to achieve stable levels of performance.

Task Duration

The scenario durations ranged from 615 to 216 sec. The average length of time was 383 sec (SD = 117 sec). On an individual basis, the average time taken to perform a scenario ranged from 410 to 369 sec. The subjects with the best scores also had the fastest times, suggesting there that was no speed/accuracy tradeoff, however the overall correlation between score and duration was only +0.49. The correlation between scenario duration and workload was -0.41, shorter sessions being associated with greater workload. The presence of **ACCELERATION** resulted in a sharp decrease in session length, as can be seen in Figure 7, because task elements moved more quickly and were, therefore

available for performance with less delay. In all cases, the obtained session durations were less than the baseline durations used to create the schedules. When the schedules were designed, it was assumed that tasks would be performed one at a time, that LINK, MORE, SHED, etc. would not be used to decrease time-to-completion, and that all tasks would be performed so as to impose schedules that would allow time for an average operator to complete most of the tasks.

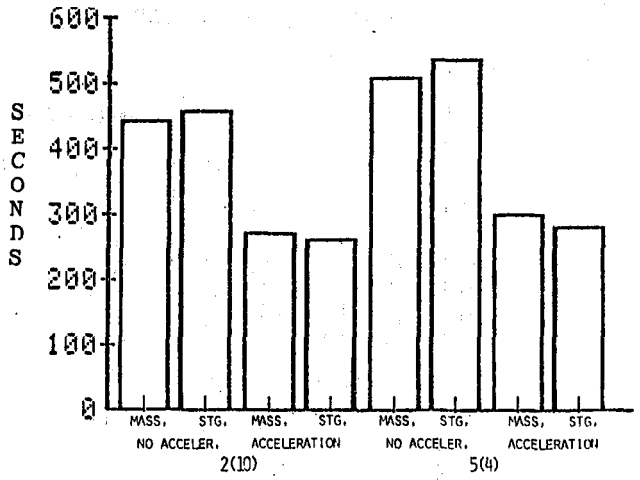


Figure 7: Average durations of the experimental conditions. (n = 8)

Operator Strategies

The relative proportion of function key actuations made be each subject may be seen in Figure 8. There was relatively little difference among subjects in PERFORM key actuations, although high-scoring subjects, obviously, used it more often than low-scoring subjects because they operated on several tasks at the same time, rapidly switching from one open task to another. Thus, the two high-scoring subjects (mean = 409) averaged 80 different task selections per scenario (S-5 and S-7), while the low-scoring subject (mean = 321) averaged 25 selections per scenario (S-8). High-scoring subjects used the OPEN, CLOSE, STUFF, and MORE commands nearly twice as often as low-scoring subjects, thereby controlling the flow of active tasks. Although there was considerable variation in the use of the LINK and SHED commands, their use was not significantly correlated with score, rated workload or task duration.

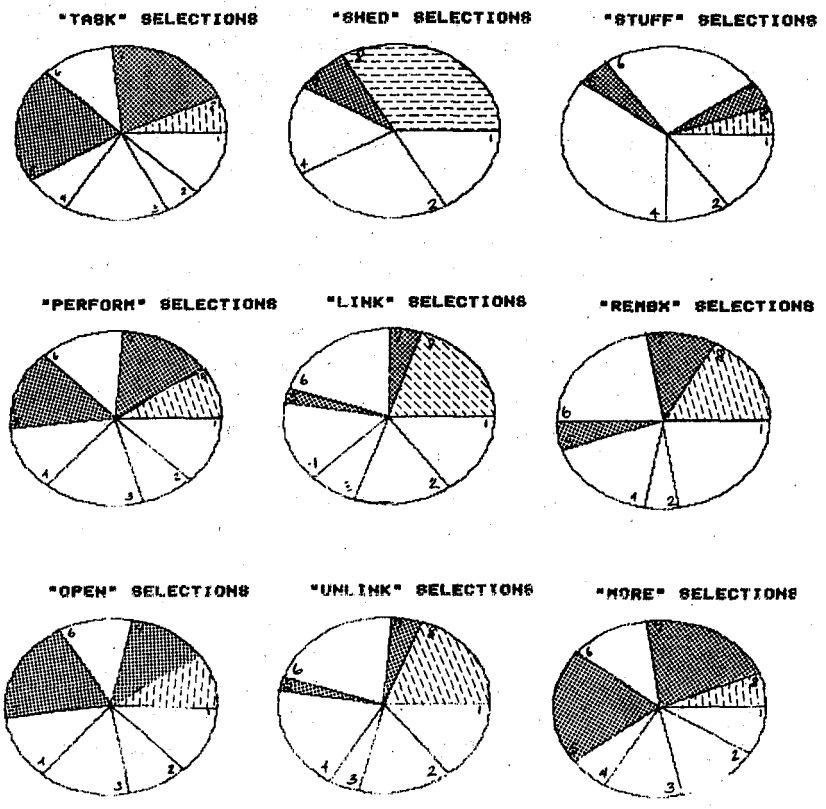


Figure 8: Relative proportion of times each function was selected by subjects whose score was low (▒▒▒), high (▓▓▓), or average ( ).

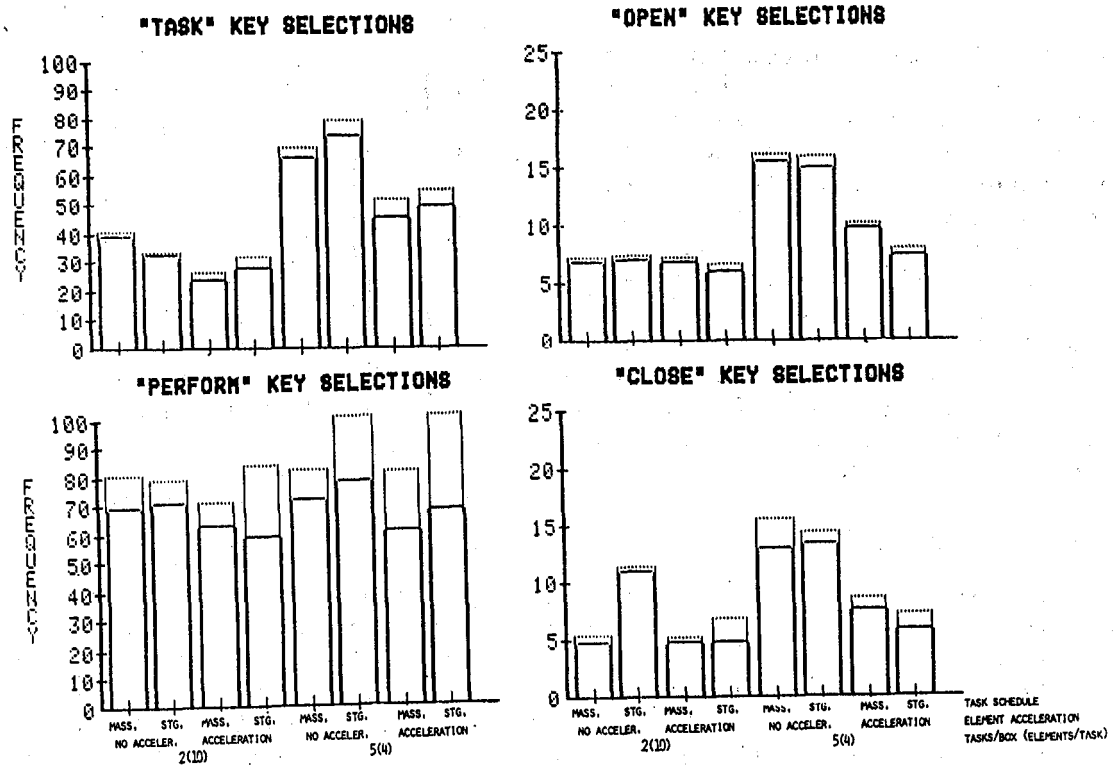


Figure 9: Frequency of appropriate (┌─┐) and inappropriate (┌╌╌┐) selections of the basic functions by experimental condition (n = 8).

Basic Functions

The scenarios were designed so that each task had the same payoff, rate, number of elements/tasks, and schedule, thus, similar performance was anticipated across the five tasks. Individual analyses of variance for repeated measures were performed for each function to determine whether each was selected equally often for the five tasks task across experimental conditions. Since no significant differences were found for any function, subsequent analyses were performed collapsed across task type.

On the average, the four basic functions were selected 152 times per scenario. Of these selections, 85%, were made correctly. The remaining selections were slips (the operator intended to select one function but actuated another instead) or errors (the opearato# selected an inappropriate function). An inappropriate TASK or PERFORM selection occurred when no task elements were available. An inappropriate OPEN or CLOSE selection was one that was made when the selected box was already open or closed. Different TASKS were selected 48 times (96% correctly), PERFORM was selected 85 times (83% correctly), and OPEN and CLOSE were selected 10 and 9 times, respectively (99% correctly), across experimental conditions. (Figure 9)

Problem-solving functions

The four functions that were provided as solutions to lagging behind the system were selected two times each, on the average. Use of the SHED command and the performance of tasks from the penalty box characterized low-scorers. (Figure 8) The REMOVE BOX command was used only once per scenario by high-

Comparison of Experimental Conditions

Measures of performance, behavior, and subjective experience were analyzed to determine the relative impact of the four experimental variables. Significant variations in measures of performance may be seen in Figures 6 and 7. Significant variations in the selections of functions may be seen in Figures 9, 10, and 11. Workload ratings may be seen in Figure 5.

REWARD/PENALTY Conditions

Performance There were no significant variations in scores or duration

Workload Rating There was no significant variation in rated workload

Use of Basic Functions There were no significant variations in the frequency with which any of the basic functions were selected as a function of the REWARD/PENALTY condition.

Use of Problem-Solving Functions There was a significant increase in the use of the SHED command (from 0.6 to 2.25 selections/scenario) when a penalty was levied for failing to perform task elements. If it was clear to the operator that he could not perform one or more elements before they reached the graveyard or were transferred to the penalty box, this was the correct strategy (on +5/-5 trials) to avoid an additional five-point penalty. Since as many as 10 elements could have been shed with a single command (all those remaining in the box), this measure may underestimate the impact of this function on the subsequent structure of the task. None of the other commands were used significantly differently due to the REWARD/PENALTY condition.

Use of Lead-Generating Functions There was no significant change in the of use of these commands as a function of the REWARD/PENALTY condition.

Schedule

Performance There was a significant ($F(1,7) = 19.16$, $p<.01$) decrease in score (from 404 to 347) between the MASSED and STAGGERED schedules. Overall, the length of time taken to complete scenarios was not affected by the schedule, even though the scheduled durations were offset by 5 sec in the STAGGERED condition, potentially increasing the time required to complete a task.

Workload Rating Rated workload increased significantly ($F(1,7) = 19.11$, $p< .01$) from 41 (MASSED) to 46 (STAGGERED) as schedule complexity was increased, reflecting the additional mental processing load imposed by more complex schedules.

Use of Basic Functions There were no significant differences in the use of any of the basic functions due to schedule alone. The same number of boxes were OPENed, CLOSEed, and SELECTed. The PERFORM key was actuated more often with the STAGGERED schedule than with the MASSED schedule, but 25% of the selections were made in error. This resulted in a significant difference in the number of erroneous PERFORM key selections ($F(1,7) = 14.5$, $p<.01$).

Use of Problem-Solving Functions All of the problem-solving functions were used more often with the STAGGERED condition, indicating that the operators were lagging the system. STUFF ($F(1,7) = 13.72$, $p<.01$) and REMOVE BOX

($F(1,7) = 13.75$, $p<.01$) commands increased significantly, and more tasks ended up in the penalty box ($F(1,7) = 9.17$, $p<.05$).

Use of Lead-Generating Functions The LINK and UNLINK commands were selected significantly ($F(1,7) = 9.01$, $p<.05$) less often with the **STAGGERED** condition, as expected, because it was rare that two tasks were at the same stage of performance and were, therefore likely candidates for LINKing. MORE tasks were selected half as often with the **STAGGERED** schedule ($F(1,7) = 11.46$, $p < .05$), another indication that subjects were not able to get ahead of the system in this condition.

Acceleration of elements

Performance There was a highly significant difference in scores due to the presence (460) or absence (291) of **ACCELERATION** ($F(1,7) = 773.2$, $p<.001$), particularly when **ACCELERATION** was combined with a **STAGGERED** schedule and when there were a greater number of different tasks. These synergistic effects were reflected in a significant **SCHEDULE** by **ACCELERATION** interaction ($F(1,7) = 8.43$, $p<.05$) and in a significant **ELEMENTS/TASK** by **ACCELERATION** interaction ($F(1,7) = 20.77$ $p<.01$).

Scenarios with **ACCELERATION** were completed significantly more quickly ($F(1,7) = 168.7$, $p<.001$) than those without (488 versus 277 sec, respectively). With **ACCELERATION**, task elements arrived more quickly and were available for performance at a faster rate once in a box, thus, operators were not constrained by system delays in completing tasks.

With **ACCELERATION**, the number of times that functions were selected inappropriately was increased, possibly because operators were under greater time pressure. Significantly more tasks ended up in the graveyard ($F(1,7) = 83.41$ $p<.001$) and penalty box ($F(1,7) = 31.29$, $p<.001$) with **ACCELERATION**.

Workload Rating There was a significant ($F(1,7) = 30.56$, $p<.001$) increase in workload ratings with **ACCELERATION** (from 32 to 55). The influence of **ACCELERATION** on experienced workload was particularly great when it was combined with a **STAGGERED** schedule with many elements to be performed per task. This was reflected in a significant three-way interaction among **SCHEDULE**, **ACCELERATION**, and **ELEMENTS/TASK** ($F(1,7) = 14.44$, $p<.01$). Rated workload may have been highest in the 2(10), **ACCELERATED** scenarios because tasks with many elements took longer to complete and were thus subject to the effects of acceleration for a longer time.

Use of Basic Functions There was no significant change in the use of the TASK select or the PERFORM functions due to **ACCELERATION**. There was, however, a significant decrease in the number of times that the OPEN ($F(1,7) = 15.29$, $p<.01$) and CLOSE ($F(1,7) = 9.07$, $p<.01$) functions were used, particularly when there were many different tasks per box. There was a significant three-way interaction among **ACCELERATION**, **SCHEDULE**, and **ELEMENTS/TASK** for the OPEN function ($F(1,7) = 119.62$, $p<.001$). Boxes were OPENed 6.5 times per scenario, on the average, in the 2(10) condition regardless of **SCHEDULE** or **ACCELERATION**, whereas they were OPENed as often as 16 times per scenario without **ACCELERATION** in the 5(4) condition and 10 times with **ACCELERATION**. In the easier conditions, and when only two tasks with 10 elements each were scheduled, subjects OPENed each box one time and left it that way. They did not OPEN and CLOSE boxes as a management strategy. When five tasks were scheduled per box,

however, they did close the boxes occasionally between different tasks, but considerably less often than once for every one of the 25, four-element tasks.

Use of Problem-Solving Functions All of these functions were used more often with **ACCELERATION** than without. Significantly more tasks had to be SHED ($F(1,7) = 9.36$, $p<.01$), STUFFed $F(1,7) = 5.67$, $p<.05$) and performed with the additional REMOVE BOX procedure ($F(1,7) = 6.12$, $p<.05$). These differences indicate that subjects were more likely to lag behind the system with **ACCELERATION** than without.

Use of Lead-Generating Functions A related finding was that there were fewer requests for MORE tasks ahead of schedule with **ACCELERATION** than without ($F(1,7) = 19.24$, $p<.01$). The difference was particularly great when more tasks were actually available (in the 5(4) condition). There was a significant interaction between **SCHEDULE** and **ELEMENTS/TASK** ($F(1,7) = 12.13$, $p<.01$). **ACCELERATION** did not affect the use of LINK and unLINK.

Number of ELEMENTS/TASK

Performance There was a significant ($F(1,7) = 114.1$, $p<.001$) decrease in score (from 430 to 322) when there were more different tasks with fewer elements each. This decrease was accentuated by **ACCELERATION** ($F(1,7) = 20.77$, $p<.01$) and by a **STAGGERED** schedule ($F(1,7) = 15.13$, $p<.01$).

The time taken to complete a scenario was significantly longer when there were more discrete tasks to be performed ($F(1,7) = 43.3$. $p<.001$) than when the same number of elements were grouped into fewer (albeit more complex) tasks. To some extent, this increase in time occured because four-element tasks did not remain in the boxes as long as ten-element tasks and thus never developed the same rates of speed due to **ACCELERATION** .

More functions were selected inappropriately as the number of discrete tasks increased (18 versus 23%). The decrease in score, increase in time-to-complete a scenario, and increase in errors in the 5(4) condition may reflect the cost of shifting attention among 25 smaller tasks, even though each was individually less complex.

Workload ratings The greatest increase in rated workload was found between the 5(4) and 2(10) conditions. This significant increase ($F(1,7) = 51.2$, $p<.001$) reflected the operators' perceptions that an increase in the number of different tasks that they were required to do (even if the total number of subtask elements remained the same) imposed a substantial increase in their workload.

Use of Basic Functions Not surprisingly, there were significantly ($F(1,7) = 50.8$, $p<.001$) more TASK selections with the 5(4) condition than with the 2(10) condition, because subjects had to shift their attention among many discrete tasks. The difference (33 versus 64) was not as great, however, as the 250% increase in the actual number of different tasks scheduled for each box. Although the OPEN function was selected significantly ($F(1,7) = 42.6$, $p<.001$) more often in the 5(4) conditions than in the 2(10) conditions, the increase (from 7 to 13 times per scenario) was proportionally less than would be expected from the actual increase in number of different tasks per scenario (from 10 to 25). Relatively speaking, subjects shifted their attention from one task to the next less often as the number of discrete tasks was increased.

Use of Problem-Solving Functions There was no significant change in the use of the REMOVE BOX and STUFF commands as a consequence of the number of tasks per box. There was, however, a significant ($F(1,7) = 9.23$, $p<.05$) increase in the use of the SHED command when there were fewer elements, but more tasks. This might have occurred because the SHED command had a less dramatic effect on reducing the number of tasks remaining to be performed for score points when there were only 4 elements per task rather than 10. There were significantly more tasks transferred to the penalty box ($F(1,7) = 29.6$, $p<.001$) in the 5(4) scenarios than in the 10(2) scenarios. Seven times as many elements were performed from the penalty box (but with no increase in score) in the 5(4) scenarios than in the 2(10) scenarios. This occurred because there were five separate arrivals of tasks in each box, thereby increasing the chance (by 250%) that a new task would enter an box still occupied by an existing task. Since there was no significant difference in the number of times the PERFORM function was selected, the lower scores obtained with the 5(4) conditions occurred because more tasks were SHED and more ended up in the penalty box (thus no points were gained for them even if they were performed), not because they selected the PERFORM function less often.

Use of Lead-Generating Functions Although more tasks were requested ahead of schedule in the 5(4) condition than in the 2(10) condition (5.3 times per scenario versus 3.1), the difference was not significant. In addition, the increase was considerably less than would be expected by the increased opportunities to request tasks ahead of schedule provided by the 5(4) scenario (4 times per box) than the 2(10) scenario (once per box). The LINK and unLINK commands were used considerably less often than they could have been in the 5(4) scenarios. The difference in usage between the two conditions was not significant.

Relative Importance of Experimental Conditions

The relative impact of the different experimental manipulations was analyzed by examining the amount of variance accounted for by each of them in the statistical analyses performed on the scores, workload ratings, and function selections. In addition, each subject was asked to rank order the four factors with respect to the impact that they felt each had had on workload. The **REWARD/PENALTY** conditions contributed little to variations in performance, behavior or opinion. The **ELEMENTS/TASK** had the greatest impact on the frequency of basic function selections. Presence or absence of **ACCELERATION** and **MASSED** versus **STAGGERED** schedules, particularly when they covaried, had the greatest impact on problem-solving behavior, lead-generating responses, and score. Although the number of **ELEMENTS/TASK** contributed most to the variance of workload ratings, the factor selected as most influential by the subjects at the end of the experiment was the **SCHEDULE** (a relatively less important influence on measures obtained during and immediatly after the scenarios).

CONCLUSIONS

All of the experimental manipulations, alone and in combination, generated highly significant differences in operator behavior, performance, and experienced workload with the exception of the **REWARD/PENALTY** condition. Each variable had slightly different influences on individual measures, however,

scorers. Instead of performing the additional REMOVE BOX step for tasks in the warning zone, they selected the STUFF option, using this strategy six times more often per scenario than did the low-scoring subjects. The STUFF and REMOVE BOX commands were selected in error at least once per scenario, although the SHED command was never selected erroneously. (Figure 10)

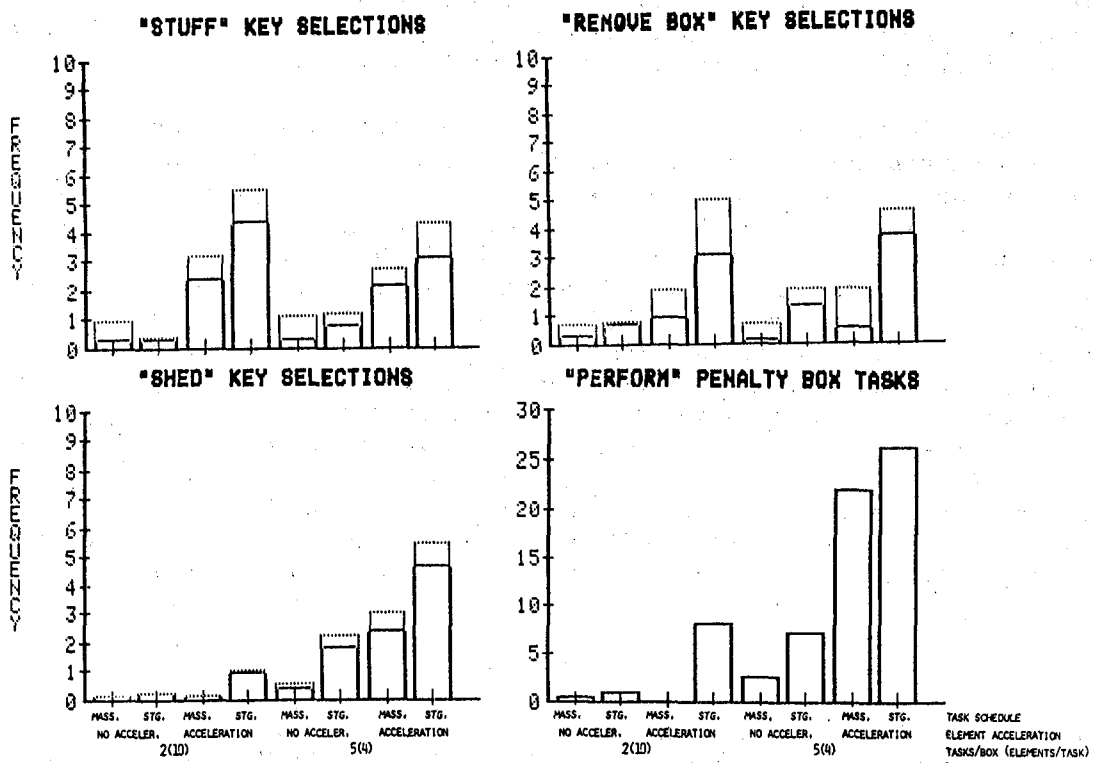


Figure 10: Frequency of appropriate (⌐¬) and inappropriate (¬⌐) selections of problem-solving functions by experimental condition (n = 8)

Lead-generating Functions

The LINK command was used rarely by high-scoring subjects (once or twice per scenario), but relatively often by the others (seven times per scenario). It was usually selected appropriately. (Figure 11) On the average, the MORE command was selected four or five times per scenario, however more than half of the time more tasks were requested none were available for the selected box. High-scorers used the MORE command three times more often than low-scorers because they were able to complete tasks ahead of schedule.

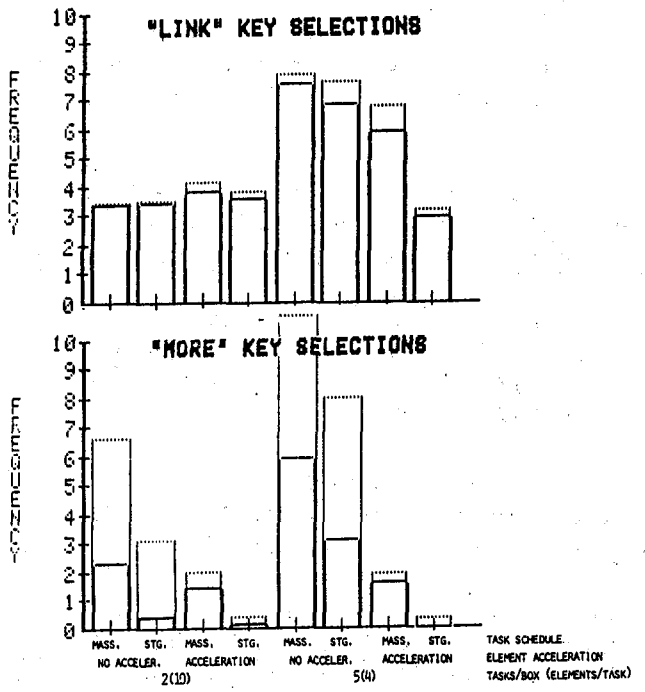


Figure 11: Frequency of appropriate (⌐¬) and inappropriate (¬⌐) lead-generating function selections (n=8)

allowing a detailed and informative analysis of the effect of experimental manipulations of imposed workload on different aspects of operators' behavior and performance. The initial objective of developing a set of scenarios that imposed a range of variations in performance and workload and investigating the impact of different penalties for procrastination was satisfied. In addition, the efficacy of the proposed weighted combination of workload components in reducing between-subject variability was demonstrated.

In future experiments, the effects of training must be determined to establish optimal procedures and asymptotic levels. In addition, the influence of task rate and value should be manipulated so as to replicate the critical variables in the earlier studies (refs. 5, 6, 7) using the current paradigm. In future research, particular attention should be given to the impact of machine-aiding, automation, and system failures on performance, behavior, and workload. Given the success of this simulation in generating significant variations in performance and workload, this paradigm should continue to provide a useful environment in which measures of workload and performance can be developed, tested and calibrated once standardized levels of imposed task load have been established. This experiment was designed to evaluate utility of the POPCORN simulation as an experimental task. It remains to future researchers to apply the different theoretical and mathematical models (depending on their experimental goals) to use this simulation as a prototype of multi-task, automated and semi-automated supervisory control systems.

REFERENCES

1. Hart, S. G. and Sheridan, T. S. Pilot workload, performance, and aircraft control automation. Proceedings of the AGARD Symposium on Human Factors Considerations in High Performance Aircraft.. Williamsburg, VA, April 1984. (in Press)

2. National Research Council. Automation in Combat Aircraft. Committee on Automation in Combat Aircraft. Washington, D. C.: National Academy Press, 1982.

3. Shingledecker, C. A., Crabtree, M. S. and Acton, W. H. Standardized tests for the evaluation and classification of workload metrics. Proceedings of the Human Factors Society - 26th Annual Meeting. Seattle, WA, 1982, 648-651.

4. Derrick, W. L. and Wickens, C. D. A Multiple Processing Resource Explanation of the Subjective Dimensions of Operator Workload. Urbana-Champaign, IL. Engineering Psychology Research Laboratory Technical Report EPL-84-2/ONR-84-1, February 1984.

5. Tulga, K. M. and Sheridan, T. S. Dynamic decisions and workload in multitask supervisory control. IEEE Transactions on Systems, Man and Cybernetics, 1980, SMC-10, 217-231.

6. Daryanian, B. Subjective scaling of mental workload in a multi-task environment. Proceedings of the 16th Annual Conference on Manual Control. Massachusetts Institute of Technology, 1980, 172-188.

7. Pattipati, K. R., Kleinman, D. L., and Ephrath, A. R. A dynamic decision model of human task selection performance. IEEE Transactions on Systems, Man,

and Cybernetics, 1983, SMC-13(3), 145-156.

8. Sheridan, T. S. & Stassen, H. Definitions, models, and measures of human workload. In N. Moray (Ed.) Mental Workload: Its Theory and Measurement. New York: Plenum Press, 1979, 219-234.

9. Moray, N. Subjective mental workload. Human Factors, 1982, 24(1), 25-40.

10. Hart, S. G., Childress, M. E. & Hauser, J. R. Individual definitions of the term "workload". In the Proceedings of the 1982 Psychology in the DOD Symposium. U. S. Air Force Academy, 1982.

11. Reid, G. B., Eggemeier, F. T. and Nygren, T. E. An individual differences approach to SWAT scale development. Proceedings of the Human Factors Society - 26th Annual Meeting. Seattle, WA, 1982, 639-62.

12. Kantowitz, B. H., Hart, S. G., Bortolussi, M. R., Shively, R. J., and Kantowitz, S. C. Measuring pilot workload in a moving-base simulator: II. Building levels of workload. Proceedings of the 20th Annual Conference on Manual Control., 1984 (in press)

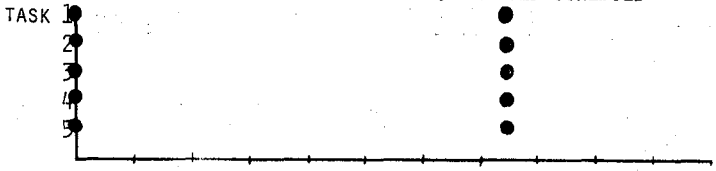
13. Miller, R. C. and Hart, S. G. Assessing the subjective workload of directional orientation tasks. Proceedings of the 20th Annual Conference on Manual Control., 1984 (in press)

APPENDIX A: Scheduled arrival times of tasks

SCHEDULED ARRIVAL TIME OF TASKS (BY SCENARIO)

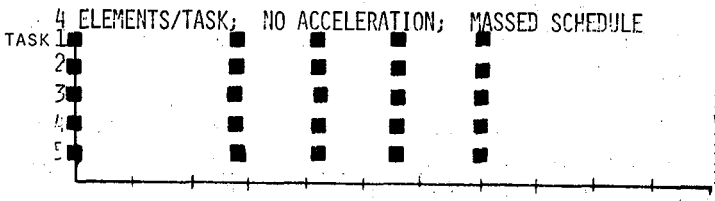
SUMMARY OF RESULTS (BY SCENARIO)

10 ELEMENTS/TASK; NO ACCELERATION; MASSED SCHEDULE



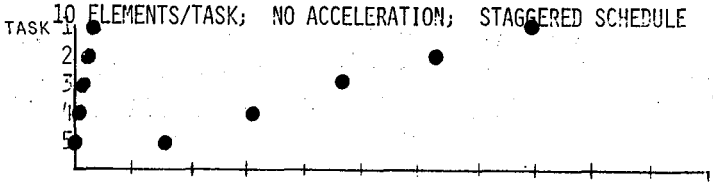
DIFFICULTY RANK: 1 (least)
SCORE: 454
DURATION: 512
WORKLOAD RATING: 40
BASIC FUNCTION SELECTIONS: 70/82/16/16=186
PROBLEM SOLVING FUNCTION SELECTIONS: 1/1/1/3=6
LEAD-GENERATION FUNCTION SELECTIONS: 11/8=19

4 ELEMENTS/TASK; NO ACCELERATION; MASSED SCHEDULE



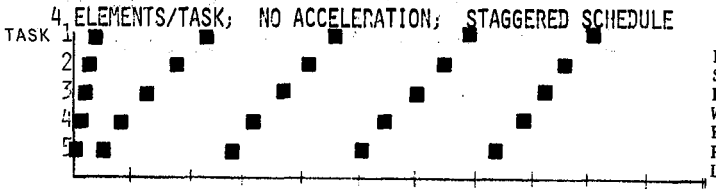
DIFFICULTY RANK: 2
SCORE: 491
DURATION: 444
WORKLOAD RATING: 25
BASIC FUNCTION SELECTIONS: 40/81/7/5=133
PROBLEM SOLVING FUNCTION SELECTIONS: 1/0/1/0=2
LEAD-GENERATION FUNCTION SELECTIONS: 7/3=10

10 ELEMENTS/TASK; NO ACCELERATION; STAGGERED SCHEDULE



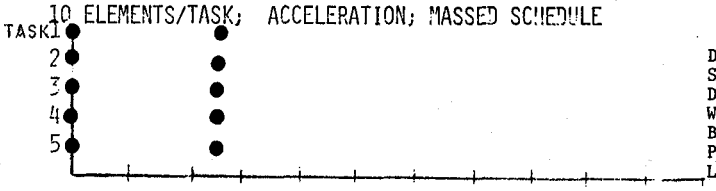
DIFFICULTY RANK: 3
SCORE: 417
DURATION: 539
WORKLOAD RATING: 38
BASIC FUNCTION SELECTIONS: 79/101/16/14=210
PROBLEM SOLVING FUNCTION SELECTIONS: 1/2/2/7=12
LEAD-GENERATION FUNCTION SELECTIONS: 8/8=16

4 ELEMENTS/TASK; NO ACCELERATION; STAGGERED SCHEDULE



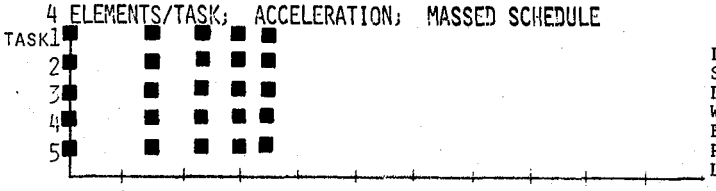
DIFFICULTY RANK: 4
SCORE: 477
DURATION: 456
WORKLOAD RATING: 26
BASIC FUNCTION SELECTIONS: 33/79/7/11=130
PROBLEM SOLVING FUNCTION SELECTIONS: 0/0/1/1=2
LEAD-GENERATION FUNCTION SELECTIONS: 3/3=6

10 ELEMENTS/TASK; ACCELERATION; MASSED SCHEDULE



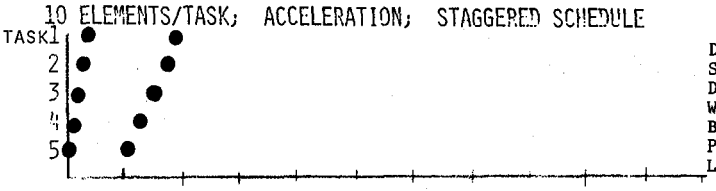
DIFFICULTY RANK: 5
SCORE: 460
DURATION: 272
WORKLOAD RATING: 35
BASIC FUNCTION SELECTIONS: 26/71/7/5=109
PROBLEM SOLVING FUNCTION SELECTIONS: 3/0/2/0=5
LEAD-GENERATION FUNCTION SELECTIONS: 2/4=6

4 ELEMENTS/TASK; ACCELERATION; MASSED SCHEDULE



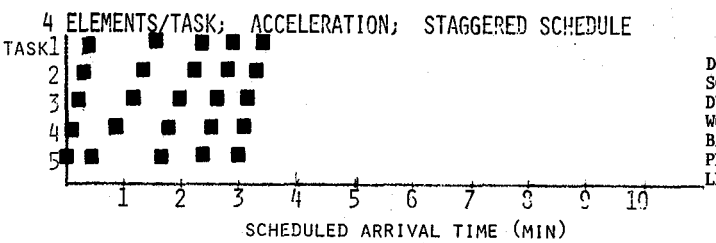
DIFFICULTY RANK: 6
SCORE: 213
DURATION: 300
WORKLOAD RATING: 64
BASIC FUNCTION SELECTIONS: 52/81/10/9=152
PROBLEM SOLVING FUNCTION SELECTIONS: 3/3/2/22=30
LEAD-GENERATION FUNCTION SELECTIONS: 2/7=9

10 ELEMENTS/TASK; ACCELERATION; STAGGERED SCHEDULE



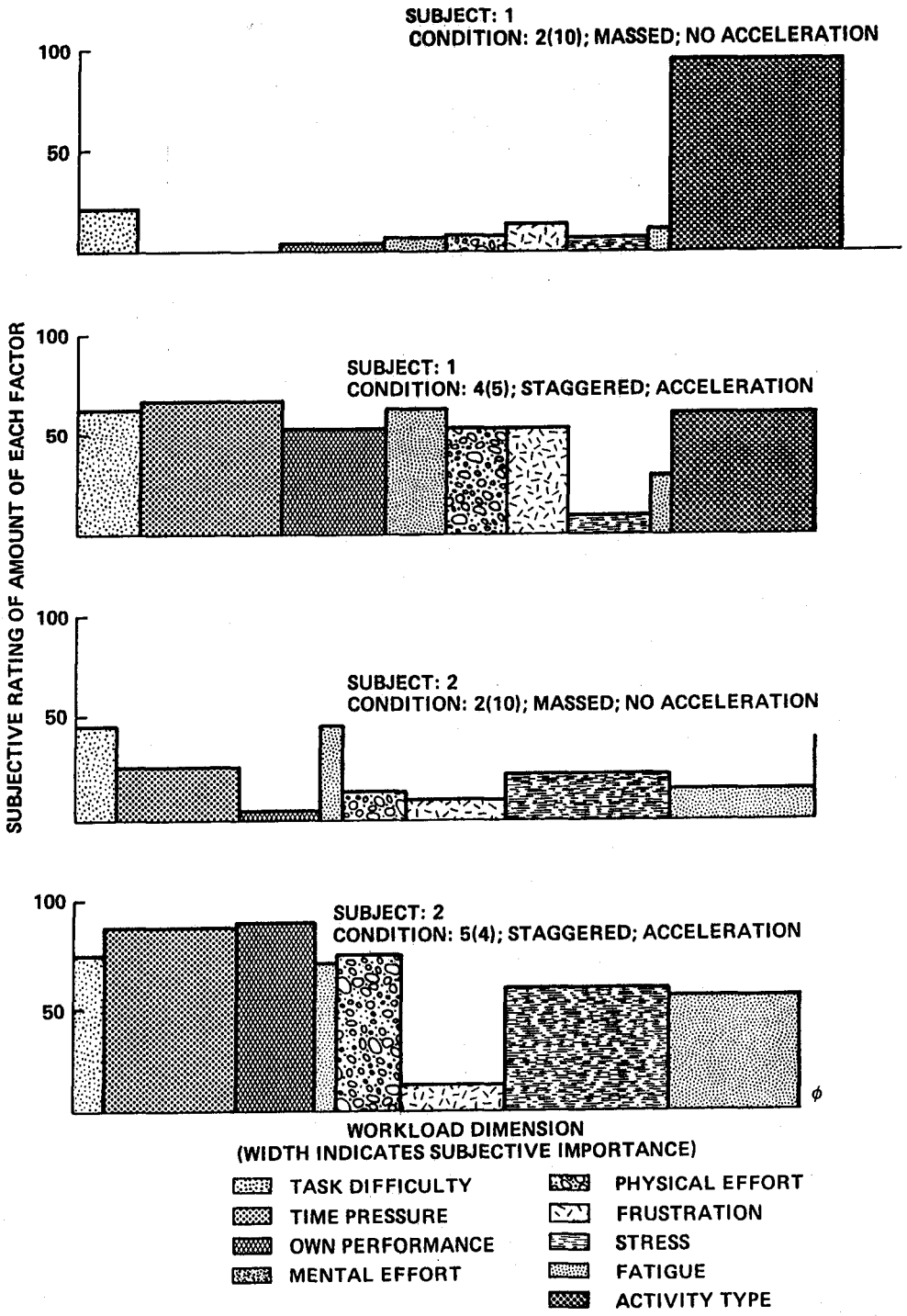
DIFFICULTY RANK: 7
SCORE: 203
DURATION: 278
WORKLOAD RATING: 59
BASIC FUNCTION SELECTIONS: 55/100/8/7=170
PROBLEM SOLVING FUNCTION SELECTIONS: 4/5/5/27=41
LEAD-GENERATION FUNCTION SELECTIONS: 0/3=3

4 ELEMENTS/TASK; ACCELERATION; STAGGERED SCHEDULE



DIFFICULTY RANK: 8 (most)
SCORE: 290
DURATION: 260
WORKLOAD RATING: 60
BASIC FUNCTION SELECTIONS: 31/84/7/7=129
PROBLEM SOLVING FUNCTION SELECTIONS: 6/1/5/8=20
LEAD-GENERATION FUNCTION SELECTIONS: 0/4=4

● TASKS WITH 10 ELEMENTS EACH
■ TASKS WITH 4 ELEMENTS EACH



Appendix B: Eaxmple of weighting procedure applied to the bipolar ratings obtained from each of two different operators after performing a relatively easy scenario and a relatively difficult scenario. (See Figure 4 for the importance placed on each of the factors by these two subjects).