NASA CONTRACTOR REPORT

NASA CR-2150

NASA CR-2150

MEASUREMENT OF OPERATOR WORKLOAD IN AN INFORMATION PROCESSING TASK

by Larry L. Jenney, Harry J. Older, and Bernard J. Cameron

Prepared by
BIOTECHNOLOGY, INC.
Falls Church, Va. 22042
for Langley Research Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • DECEMBER 1972

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR-2150		
4. Title and Subtitle: MEASUREMENT OF OPERATOR WORKLOAD IN AN INFORMATION PROCESSING TASK		5. Report Date: December 1972
		6. Performing Organization Code
7. Author(s): Larry L. Jenney, Harry J. Older, and Bernard J. Cameron		8. Performing Organization Report No.
		10. Work Unit No.
9. Performing Organization Name and Address: BioTechnology, Inc. 3027 Rosemary Lane Falls Church, Va. 22042		11. Contract or Grant No. NAS 1-10118
		13. Type of Report and Period Covered: Contractor Report
12. Sponsoring Agency Name and Address: National Aeronautics and Space Administration Washington, D. C. 20546		14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This was an experimental study to develop an improved methodology for measuring workload in an information processing task and to assess the effects of shift length and communication density (rate of information flow) on the ability to process and classify verbal messages. Each of twelve subjects was exposed to combinations of three shift lengths and two communication densities in a counterbalanced, repeated measurements experimental design. In addition to task-specific measures, subjects were administered a battery of perceptual-motor, cognitive and sensory tests on a pre- and post-shift basis. Physiological measures and subjective magnitude estimates of three workload variables were also obtained. Results indicated no systematic variation in task performance measures or in other dependent measures as a function of shift length or communication density. This is attributed to the abence of a secondary loading task, an insufficiently taxing work schedule, and the lack of psychological stress. Subjective magnitude estimates of workload showed fatigue (and to a lesser degree, tension) to be a power function of shift length. Estimates of task difficulty and fatigue were initially lower but increased more sharply over time under low density than under high density conditions. An interpretation of findings and recommendations for future research are included. This research has major implications to human workload problems in information processing of air traffic control verbal data.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
human operator workload information processing voice communications analysis human factors	Unclassified - Unlimited

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	71	$3.00

*For sale by the National Technical Information Service, Springfield, Virginia 22151

TABLE OF CONTENTS

TABLE OF CONTENTS—Continued

LIST OF TABLES

LIST OF FIGURES

MEASUREMENT OF OPERATOR WORKLOAD IN AN INFORMATION PROCESSING TASK

Larry L. Jenney, Harry J. Older, Bernard J. Cameron
BioTechnology, Inc.

CHAPTER 1

BACKGROUND

Problems in Measurement of Operator Workload

Workload may be defined as the level of effort required to perform a given activity or complex of tasks. "Level of effort" is an imprecise term denoting an internal condition or process which, with the exception of purely muscular activities, cannot be measured directly. Therefore, proximate measures must be sought. Either of two approaches may be taken. First, when work results in an objectively measurable product, the quantity or quality of the output can be determined; and from it inferences can be made about the effort required to produce the output. The greater or better the output, the greater the effort or workload. Alternatively, one can seek to measure not the product but a related state of the organism before and after an activity. The change of state (e.g., physiological condition, perceptual-motor capability, cognitive capacity, etc.) is taken to be an index of the amount of work required to perform a given activity.

This is a simplification of the field of ergonomics, but it does serve to bring out two important points. First, workload can be defined and quantified either in terms of a product or a change of state in the working organism. Second, because only indirect access to the process is usually possible, problems of causal relationships and interaction effects frequently arise in the interpretation of experimental results. As the work considered becomes more complex, as the process tends to have more internal and fewer tangible results, and as the interplay of environmental and situational effects grows more subtle, the problems of measuring workload take on increasing difficulty.

This final observation is particularly true when dealing with activities which are largely cognitive, such as information processing. It is commonly understood that virtually every human activity involves sensing, transformation, and storage of information; but studies of workload frequently treat these functions as intermediate steps without determining their specific cost to the individual's energy reserves or the effects which variations in the kind and amount of information processing have on the work output. In part, this may be due to the difficulties of measuring so intangible an activity in meaningful terms and of identifying the impact which environmental and situational factors have on the ability to do mental work.

One major interest in performance measurement has been with environmental stress factors which influence performance (see Trumbull, 1965). By segregating and measuring the effects of elements related to the site and conditions in which work is done, it is, at times, possible to determine the residual performance decrement attributable to the workload per se. The research literature on human performance is replete with reports of such studies dealing with the relation of environment to performance. Among the factors examined have been confinement, sleep deprivation, temperature, changes in partial pressures of ambient gases, noise, lighting, acceleration, and vibration.

Research on environmental effects has been difficult, however, chiefly because of problems in identifying clearly the environmental variables which relate to performance and in isolating them experimentally. In even the most carefully conceived experiments, it is usually impossible for investigators to control fully more than two or three of the independent variables of the environment which influence performance. Typically, the interaction effects are of such complexity that it is almost impossible to identify those factors which are responsible for variations in performance of the tasks under study and to measure accurately the extent of their influence.

To reduce the number of simultaneously varying factors, some investigators have constructed artificial tasks which replicate operational or "real life" work situations. These artificial work situations are highly structured, and the tasks are designed to yield readily quantifiable performance measures. The notion is that by using a standardized task as a baseline, the investigator will be able to identify and measure the environmental and situational factors which also influence performance or contribute to workload. In this way, it is possible to achieve a greater degree of control of the experimental situation and to facilitate the interpretation of cause-effect relationships. However, the use of artificial tasks also has the undesirable consequence of making it difficult to extrapolate from the experimental setting to operational situations of practical significant. This observation is not to deprecate the value of such studies. The methodological contributions and insights into the relationship between environmental and performance which have resulted from such investigations have had a powerful influence on the design of modern systems and on the selection and training of personnel to man them.

Regardless of whether a real or an artificial task is used, the development of performance measures for the task also poses a dilemma. Much of the variance in modern complex systems is attributable to such factors as information processing, decision making, communication, and team interaction. These have proved difficult to quantify and manipulate experimentally. As a result, investigators have been led to choose dependent variables which are amenable to measurement but often somewhat artificial. Typically, measures such as reaction time, vigilance monitoring, tracking, arithmetic computation, physiological effects, and the like have been selected. Again, from a methodological point of view, these measures are attractive due to their relative ease of quantification and their high reliability. However, like artificial tasks, there is the attendant disadvantage of offering limited capacity for generalization to actual operational circumstances.

Faced on the one hand with the difficulties of measuring workload in actual operational circumstances and on the other with using an artificial task which may not be generalizable to "real life" situations, many investigators have attempted to steer a middle course. They have analyzed the work situation to determine its constituent skills and performance capabilities and then constructed an analog complex of experimental tasks.

Typical of this approach was a series of studies by Chiles, Alluisi, and Adams (1968). Extending over an eight-year period, these studies were an important contribution to the field of performance measures. The primary concern was to establish the optimum scheduling of work periods in extended missions, but the aspect of interest here is the methodology employed. Initially, extensive activity and task analyses were conducted to identify task components and associated psychological functions. This formed the basis for constructing experimental tasks which called for similar performance and which could serve as analogs of the actual tasks. The tasks developed included monitoring of static processes (warning lights and auditory vigilance), monitoring of dynamic processes (continuous monitoring of a fluctuating meter pointer), stimulus discrimination, information processing (arithmetic computation) and procedural performance (group problem solving behavior). In most of the experiments, tasks were combined into complexes of activity which were representative of actual mission performance requirements. Typical performance measures or dependent variables included percentage of correct solutions or signal detections, response latency, and rates of information transmission. Through a series of investigations, highly reliable measures of these functions were developed. In many of the studies, correlate physiological data such as skin resistance, skin temperature, heart rate, and respiration rate were also collected.

The results of these investigations provided specific answers to major questions of system design and mission planning. For example,

> "...It was found that two men can handle 24 man-hours of work per day very satisfactorily, even on a long-term basis (30 days or longer). On a shorter-term basis, if the likelihood of an additional stressor is low, three men can handle 48 man-hours of work per day for periods of 15 days or slightly longer."

Such findings have been of demonstrated usefulness, and this series of studies has served as something of a landmark in research of this type.

A study by Morgan and Alluisi (1969) illustrates how the synthetic task approach can be used to assess complex cognitive performance of the sort which was the topic of the present experiment.

They developed a code-transformation task which provided measures of nonverbal mediation to study the effects of workload stress on performance and, conversely, to examine the effects produced by time-sharing the code-transformation task with secondary loading tasks. The methodological implication of their findings is that it is possible to develop synthetic tasks that are sensitive to workload stress and that will serve to expose conditions in which significant performance decrements are to be expected.

Research in Air Traffic Controller Workload

The area of information processing selected for this study was air traffic controller communications. Three reasons underlay this choice. First, the air traffic controller's communication task is typical of the information processing activity which operators must perform in modern, complex systems. Second, the workload of the air traffic controller has been the subject of several important investigations in recent years. Many of these investigations have been directed toward determining the fatigue and stress produced by the length of time spent working and the pace of the job, which are precisely the workload variables which were of interest in the present study. Third, most air traffic controller workload studies have involved measurement of the operator's performance in the actual work situation. This body of data collected *in situ* provides a valuable cross-check on experimental results and a means of establishing the equivalency of the synthetic task with performance demands in the actual operational setting. It must be emphasized, however, that even though the present study dealt with air traffic control communications, no attempt was made to simulate the full array of tasks performed by controllers. The prime concern was to devise an information processing task which was demanding and representative of the performance called for in operators who must monitor voice communications and make decisions based on them.

Most studies in the air traffic control field have relied on measuring changes in the physiological status of controllers as a function of shift length and traffic density rather than on more direct measures of performance in the primary task of the controller. This has been necessary because of the inherent difficulties involved in obtaining reliable, valid, and nonintrusive measures of primary task performance in the actual work situation. The individual controller is a link in a highly complex man-machine communication system; and while the importance of the role of the individual controller is paramount and clearly recognized, the measurement of the quality and quantity of his performance while he is actually controlling traffic presents almost insurmountable methodological difficulties at the present time.

As part of a comprehensive investigation relating to stress and fatigue, the FAA Civil Aeromedical Institute conducted a study at the O'Hare airport tower in Chicago during the summer of 1968. The study was designed (1) to permit a comparison of physiological responses of controllers on different shifts and at different tower positions, (2) to determine the relationships between the stress attendant on air traffic control tasks as compared to those experienced by other populations of workers, and (3) to permit comparisons among the physiological responses of controllers at several terminals where qualitative differences in the work situations were known to exist. Physiological measures taken on the controllers included heart rate, galvanic skin response, blood pressure, and oral temperature. Biochemical measures included the pattern and quantity of phospholipids as well as fibrinogen in blood plasma. Urine samples were analyzed for epinephrine, norepinephrine, 17-OH corticosteroids, sodium, potassium, phosphate, urea, and creatinine. Data were collected from 22 controllers at regular intervals during five, eight-hour work periods on the evening shift (1600–2400) when the density of traffic was heavy, and five days on the morning shift (0000–0800) when the traffic was light.

Results indicated that significantly higher heart rates occurred on the busy evening shift than on the morning shift. On the evening shift, converging, approaching traffic was more arousing than departing, diverging traffic. There was no differential response on the morning shift. Galvanic skin response results indicated that adaptation to the morning shift was incomplete in five days. Blood fibrinogen levels were not significantly elevated above the level expected for controllers within the age group of the sample. On the other hand, controllers had a higher total plasma phospholipid level than populations of normals, schizophrenics, and combat pilots. Phosphatidyl glycerol was significantly higher in controllers' plasma than in the normal population but less than that in combat pilot and schizophrenic populations.

Results of analyses of urine specimens collected at the middle or at the end of each work period and at the end of each postwork period of sleep are reported in much more detail by Hale et al. (1971). Again, these data indicate that "in many respects, the stress of O'Hare tower work exceeded the stress induced in long or difficult flying operations, a 10-hour test in a flight simulator (inexperienced subjects), or prolonged decompression."

An earlier study (Dougherty et al., 1965) investigated the frequency of self-reported stress-related symptoms among air traffic control specialists compared with similar reports from non-air traffic control personnel. The primary conclusion of this study was that "...it is safe to conclude that as an air traffic control specialist progresses through his career, the 'sicker' he thinks himself to be in comparison with non-air traffic control specialists having similar years of experience, occupational status, and location. It is particularly noteworthy that incidence of symptoms is most highly related to experience rather than to age."

A study of substantial relevance to the present research was reported by Grandjean (1968). This study sought to measure the effects of fatigue induced in controllers handling "live" traffic at a major European airport. The dependent measures employed in this investigation included a sensory measure (critical fusion frequency), perceptual-motor performance (a normal tapping and a grid tapping test), and subjective estimates of fatigue on several rating scales. In all cases, there was a marked decrease in functional capability and a significant increase in subjective feelings of fatigue (e.g., weaker, tenser, sadder and less interested, less energetic, less awake) as the shift length grew longer.

The collective import of these studies is that the task of an air traffic controller is a difficult one, producing measurable and significant changes in physiological and subjective indices of stress and fatigue. It was data such as these which led to the selection of an air traffic control communications task for the present study.

The Information Processing Task

The study reported here sought to develop a methodology for dealing with the measurement of workload in an information processing task akin to that required of air traffic controllers and others whose work entails extensive communication activity. The aim was to devise a synthetic situation in which the independent variables were identifiable and manipulable and in which the dependent variables were not only reliable and quantifiable but also representative of an actual and important operational situation. The intent, therefore, was to seek the middle ground between full-scale simulation and task-specific laboratory techniques, and yet retain the advantages of face validity on one hand and generalizability on the other.

From the methodological point of view, the most significant innovation in this study was the development of a synthetic experimental task with a highly cognitive content. The task involved classification of the content of radio messages exchanged between air traffic controllers and pilots approaching four major U.S. airports. Subjects in the experiment were required to monitor tape recordings of these pilot-controller communications and to assign them to appropriate content categories by means of a two-digit code. The task had inherent realism and face validity and called for substantial cognitive activity to generate a correct response. The obtained performance measure had the properties of being quantifiable and reliable. Further, the measure showed promise of being generalizable beyond the immediate experimental context and of being operationally relatable to actual work situations.

Purpose of the Experiment

This experiment was a study of performance in an information processing task. The specific interest was to examine experimentally performance effects resulting from variation of two workload factors: shift length and communication density (i.e., the length of time and the rate of information processing). The primary purpose of this research was to seek improved methods for defining and quantifying the workload imposed by processing verbal communications and for assessing the effect of environmental and situational variables. An additional purpose of the research was to examine the feasibility of using verbal information exchanges, recorded in an actual operating environment, as stimulus material for experimental studies. Such authentic communications, if amenable to experimental control and manipulation, would be of value due to their inherent realism and their close correspondence to actual work situations.

CHAPTER 2
METHOD

Primary Experimental Task

The first concern in this study was development of an appropriate experimental task. The principles which guided the formulation of the task were:

1. The task must entail primarily information processing and as few other skills as possible.
2. The task must be representative of information processing tasks typically performed by human operators.
3. Task performance must be controllable and variable.
4. The task must have a performance index which is quantifiable and directly observable.

Additional features considered desirable but not essential for the purposes of the study were intrinsic difficulty and the capacity to engage and hold the subject's interest.

The task which was devised to meet these criteria was classification of air traffic control messages. In performing the task, the subject was required to monitor recordings of actual radio transmissions between pilots and air traffic controllers, to analyze the transmissions to determine the constituent parts, and to categorize these parts in terms of their information content. This task was considered "pure" in that it called almost exclusively for the ability to abstract and classify, i.e., the ability to "process" information. Further, it entailed no complex motor skills or perceptual capacities not directly related to the classification process. The task was also judged to place demands on the subjects which were representative of those called for in information processing activities. A description of the stimulus materials and the way in which the task was constructed will serve to clarify the nature of the task and performance requirements for the subjects.

The air traffic control messages used as stimulus materials were originally recorded by the Federal Aviation Administration National Aviation Facilities Experimentation Center (NAFEC) as part of an ongoing program of research in air traffic control communications. The materials consisted of eight two-hour recordings of air-ground voice communications with aircraft arriving at four major northeastern airports (John F. Kennedy, LaGuardia, Newark, and Philadelphia). The tape recordings contained nearly 8,000 messages between pilots and controllers dealing with live traffic at four very busy commercial airports.

The NAFEC system for classifying air traffic communications involves analysis of the material to three levels. These are:

Transaction – This is the first level, which is defined as the entire, *uninterrupted* information exchange between a controller and the pilot of a given aircraft. A new transaction begins each time the controller addresses, or is addressed by, a different aircraft. Thus, if the controller dealt with Aircraft A and then Aircraft B, there would be two transactions. If the controller dealt with Aircraft A, Aircraft B, and then Aircraft A again, this would be counted as three transactions because the exchange of information with B had intervened between the two exchanges with A. In the NAFEC classification system, transactions are numbered serially within each two-hour recording.

Transmission – A transaction normally consists of two or more transmissions. A transmission is defined as that segment of the transaction spoken by either party at one time. Thus, if the controller gave instructions to the pilot and the pilot replied, there would be two transmissions. If the controller gave instructions to the pilot who replied and then was addressed again by the controller, this would be counted as three transmissions. In other words, a transmission occurs each time the controller or pilot participates in the transaction. The NAFEC system uses a letter designator (P for pilot, G for ground) to identify the originator of the transmission.

Message – Each transmission consists of one or more messages. A message is a single item of air traffic control information. The NAFEC classification system categorizes messages according to content into forty mutually exclusive types, each identified by a two-digit number. A simplified version of the message classification scheme, as used in this experiment, is shown in Table I.*

The NAFEC system of analysis and classification is represented graphically in Table II. The illustration shows how a portion of air traffic control communications is divided successively into transactions, transmissions, and messages and how the messages are then classified by content. The two-digit descriptors for messages are thôse given in Table I.

The tape recorded air traffic control messages and the content analysis of these materials prepared by NAFEC were used as the basis for constructing the primary experimental task. In the original NAFEC tapes, transactions occurred irregularly, as a function of the flow of traffic approaching the airport at the time the recording was made. In order to control the rate at which messages were presented to the subjects in the experiment, the pauses in the original tapes were adjusted to obtain uniform time intervals between transactions. By lengthening or shortening the

*The NAFEC message content classification was designed to cover all types of air traffic control. Since the tape recordings used in this experiment were concerned only with approach control, the forty message types were reduced to twenty-seven by eliminating those not related to approach control (e.g., taxi instructions) and those which were relevant to approach control but did not happen to appear in the tape recordings used (e.g., ground equipment outage or breakdown).

TABLE I

Air Traffic Control Message Types[1,2]

11. Heading Assignment	41. Air Traffic Advisory
13. Altitude Assignment	42. Aircraft Equipment Status[3]
14. Speed Assignment	43. Weather
15. Clearance (for final approach)	46. Altimeter Setting
16. Holding	48. General Approach Information
	49. Visibility
21. Call-up Message	
22. Transponder Code (general)	51. No Reply[4]
23. Hand-off or Frequency Change	52. Request for Repeat
24. Transponder Code (discrete beacon)	54. Commo Check
	55. Garbled Message
31. Position Report	
32. Altitude Report	61. None of the above
33. Heading or Speed Report	62. More than six of above
34. Radar Contact	
35. Facilities and Services Available	81. Controller-to-controller

1. The NAFEC system consists of 40 message types. The version shown here is the simplified form used in the experiment.
2. The two-digit identifier applies to any assignment or report of a given type, any request for an assignment or report, and any acknowledgement thereof.
3. This refers to all aircraft equipment except radios. Radio status is classified as a 54 message.
4. Any transmission which receives no reply is classified as 51 regardless of the message(s) actually transmitted.

uniform intervals it was possible to vary the pace at which subjects were required to process blocks of information. Only the interval between transactions was manipulated; no adjustment was made in the pauses between transmissions or in the pauses between individual messages.

This rate of information flow, or "communication density," became one of the independent variables in the experiment. Communication density was defined as the proportion of the time the communication channel was used for voice messages. Two different densities were used – 55 percent voice and 45 percent silence (low density) and 70 percent voice and 30 percent silence (high density). In terms of the length of the interval between transactions, this translated to approximately 10 seconds between transactions at low density and five seconds between transactions at high density.

TABLE II

Air Traffic Control Message Classification System

UNEDITED VOICE TRANSCRIPTION	TRANSACTIONS IDENTIFIED	TRANSMISSIONS IDENTIFIED	MESSAGES IDENTIFIED	MESSAGES CLASSIFIED
United 495 descend and maintain three thousand feet. . .United 495 down to three. . .Roger. . . Allegheny 814 turn right to heading one three zero. You're cleared for the approach. . .Right to one three zero, 814 is cleared for approach. . .Roger. . . Philadelphia Approach, American 17 is with you at five thousand. . .Roger, American 17, say your airspeed please. . .We're at two forty. . Okay, American 17, you can reduce to two ten when feasible and descend to three thousand. The altimeter is three zero zero four. . .17 is out of five for three and down to two ten. Altimeter thirty oh four. . .Yes, sir.	1. United 495 descend and maintain three thousand feet. . .United 495 down to three. . .Roger.	1. G. United 495 descend and maintain three thousand feet.	1. G. United 495 [descend and maintain three thousand feet]	1. G. 13
		P. United 495 down to three.	P. United 495 [down to three]	P. 13
		G. Roger.	G. [Roger]	G. 13
	2. Allegheny 814 turn right to heading one three zero. You're cleared for the approach. . .Right to one three zero, 814 is cleared for approach. . . Roger.	2. G. Allegheny 814 turn right to heading one three zero. You're cleared for the approach.	2. G. Allegheny 814 [turn right to heading one three zero] [You're cleared for the approach]	2. G. 11 15
		P. Right to one three zero. 814 is cleared for approach.	P. [Right to one three zero.] 814 [is cleared for approach]	P. 11 15
		G. Roger.	G. [Roger]	G. 11 15
	3. Philadelphia Approach, American 17 is with you at five thousand. . . *Roger, American 17, say* your airspeed please. . . We're at two forty. . . Okay, American 17, you can reduce to two ten when feasible and descend to three thousand. The altimeter is three zero zero four. . .17 is out of five for three and down to two ten. Altimeter thirty oh four. . .Yes, sir.	3. P. Philadelphia Approach, American 17 is with you at five thousand.	3. P. Philadelphia Approach, American 17 [is with you at five thousand]	3. P. 32
		G. Roger, American 17, say your airspeed please.	G. [Roger] American 17 [say your airspeed, please]	G. 32 33
		P. We're at two forty.	P. [We're at two forty]	P. 33
		G. Okay, American 17, you can reduce to two ten when feasible and descend to three thousand. The altimeter is three zero zero four.	G. [Okay] American 17 [you can reduce to two ten when feasible] [and descend to three thousand] [The altimeter is three zero zero four]	G. 33 14 13 46
		P. 17 is out of five for three and down to two ten. Altimeter thirty oh four.	P. 17 [is out of five] [for three] [and down to two ten] [Altimeter thirty oh four]	P. 32 13 14 46
		G. Yes, sir.	G. [Yes, sir]	G. 32 13 14 46

It must be noted that the rate of information flow was not perfectly uniform at either density. Both the length of the individual transactions and the number of messages per transaction varied. The variation in the number of transactions within the 45-minute samples was ±10 percent at low density and ±13 percent at high density. That is, the number of transactions per 45 minutes ranged from 81 to 100 at low density (Mean = 90.7) and from 105 to 137 at high density (Mean = 122.4). The mean number of messages per transaction for 45-minute samples varied between 3.2 and 4.5, with a grand mean of 3.8 messages per transaction for all samples. However, these variations tended to balance out, in that the 45-minute samples with the fewer transactions tended to contain more messages per transaction, and conversely. As a result, the mean duration of transactions in each 45-minute sample remained nearly constant (Mean = 7.9 seconds, Range: 7.7-8.2).

The re-recorded high and low density tapes were further edited to eliminate portions which were considered of unacceptably low intelligibility because of poor voice rendition, low signal-to-noise ratio, or severe mismatch between the volume of pilot and controller voices. This was done to remove, insofar as possible, intelligibility as an independent variable in the experiment.

The remainder, approximately 12 hours of material of roughly uniform intelligibility, was used to produce 12 density tapes and 12 high density tapes, each 45 minutes in length. Of necessity, there was similarity between the high and low density materials since both were extracted from the same pool of original recordings. However, because of the number of 45-minute tapes (24) and the inherent sameness of all the material, it was felt that the overlap between the high and low density samples would not be of significant advantage to subjects in performing the experimental task.*

A final step in the editing of the stimulus materials was to add a voice identification of the transaction number and an end-of-transaction identifier. A female voice was used to provide a clear contrast with the predominantly male voices of the air traffic control recordings. Thus, where the original recording contained a transaction such as:

> "TWA 21, descend and maintain four thousand. . .TWA 21 out of six for four. . .Roger,"

the edited version consisted of:

> "(NUMBER 38) TWA 21 descend and maintain four thousand. . .TWA 21 out of six for four. . .Roger. (END),"

*In fact, during the experiment, some subjects were able to recognize that they had heard certain portions previously, largely because of unusual incidents which came up in the pilot-controller exchanges. However, there was no evidence to indicate that prior exposure enabled subjects to memorize particular message sequences or to perform better on subsequent exposures.

where the words in parentheses were those added in a female "voice-over." The beginning and end identifiers were added to assist the subjects in writing their responses in the proper place on the answer sheets, as explained below, and to facilitate identification of the transaction as a whole. These identifiers consumed a total of about 4 seconds for each transaction. Since they constituted "information," i.e., since they were aural inputs useful in classifying the transaction, the 4 seconds were counted as part of the total time of the transaction. Thus, the channel utilization percentages for low and high density given earlier were derived as follows.

Low Density:

Transaction number (2.5 seconds) + transaction (≈8 seconds) + end-of-transaction identifier (1.5 seconds) = 12 seconds.

Pause between transaction = 10 seconds

% channel utilization = 12/(12 + 10) = 55%.

High Density:

Transaction number (2.5 seconds) + transaction (≈8 seconds) + end-of-transaction identifier (1.5 seconds) = 12 seconds

Pause between transactions = 5 seconds

% channel utilization = 12/(12 + 5) = 70%.

Subjects recorded their individual responses to stimulus materials on preprinted answer sheets. The answer sheets were derived from the NAFEC communications analysis, which had been recorded on punch cards containing transaction numbers, transmission identifiers (P or G) and the individual message classifications. A computer printout of these cards, with blanks substituted for the message classification numbers, was used as the subject's answer sheets. Figure 1 is a sample answer sheet. The NAFEC analysis also served a second purpose. A complete printout of the cards with message classification numbers instead of blanks was used as the key for scoring subjects' responses.

Thus, the experimental task was to listen to recorded air traffic control messages and to classify them according to content on preprinted answer sheets. A classification guide, similar to that shown in Table I above, was available to the subjects for reference. In a sense, the experimental task was to duplicate the original NAFEC analysis and classification, which served as the objective standard of performance. The completeness and correctness of the subjects' responses compared to the NAFEC classification constituted the basis for measuring the subjects' performance on each 45-minute message sample. The experimental task represented a simplified version of the original NAFEC

307	P
307	G
307	P
307	G
308	G
308	P
308	G
308	P
309	P
309	G
310	G
310	P
310	G
310	P
311	P
311	G

Figure 1. Sample Answer Sheet

analysis in that the breakdown of the pilot-controller exchanges into transaction and transmissions was presented on the answer sheets. Further, the answer sheets also indicated the number of messages to be classified in each transmission. Subjects, therefore, were provided with important structural cues to facilitate the classification process. A schematic representation of the experimental task is shown in Figure 2.

Experimental Design

The two workload variables studied were shift length and communication density. The experimental design selected was a counterbalanced design in which shift length and communication density were varied systematically across two groups of six subjects each. A schematic representation of the experimental design is shown in Figure 3.

Basically, this was a repeated-measurements design in which each subject served as his own control. Each of the 12 subjects performed the primary encoding task (classifying air traffic control messages) for three different shift lengths (4, 8, and 12 hours) under two conditions of communication density (low—55 percent and high—70 percent channel utilization). Counterbalancing occurred within the basic design with respect to the order of exposure to shift lengths, the order of exposure to communication densities, and the sequence of stimulus material presentation.

The subjects were divided into two groups of six each (Group I — Subjects A-F, Group II — Subjects G-L). Each group participated in experimental runs for six consecutive days, one at each combination of shift length and density. Each subject thus performed for a total of 48 hours in the experiment. The two groups followed the same schedule with respect to shift lengths, but at different sequences of density. That is, Group I was exposed to densities in the order low, high, low, high, etc., while Group II followed the reverse order.

The primary dependent variable was performance on the basic coding task. In addition, at the beginning and end of each shift, subjects were administered a battery of other dependent, non-task-specific measures (perceuptual-motor, cognitive, and sensory). Physiological measures and subjective estimates of factors related to workload were taken at the end of each hour of work. A description of these measures is given in the following section.

Thus, the experimental plan was to evaluate the effects of shift length and communication density on coding performance and on secondary dependent measures (perceptual-motor, physiological, sensory, and cognitive). In addition, the effects of shift length and density on subjective estimates of fatigue, tension, and workload were to be examined.

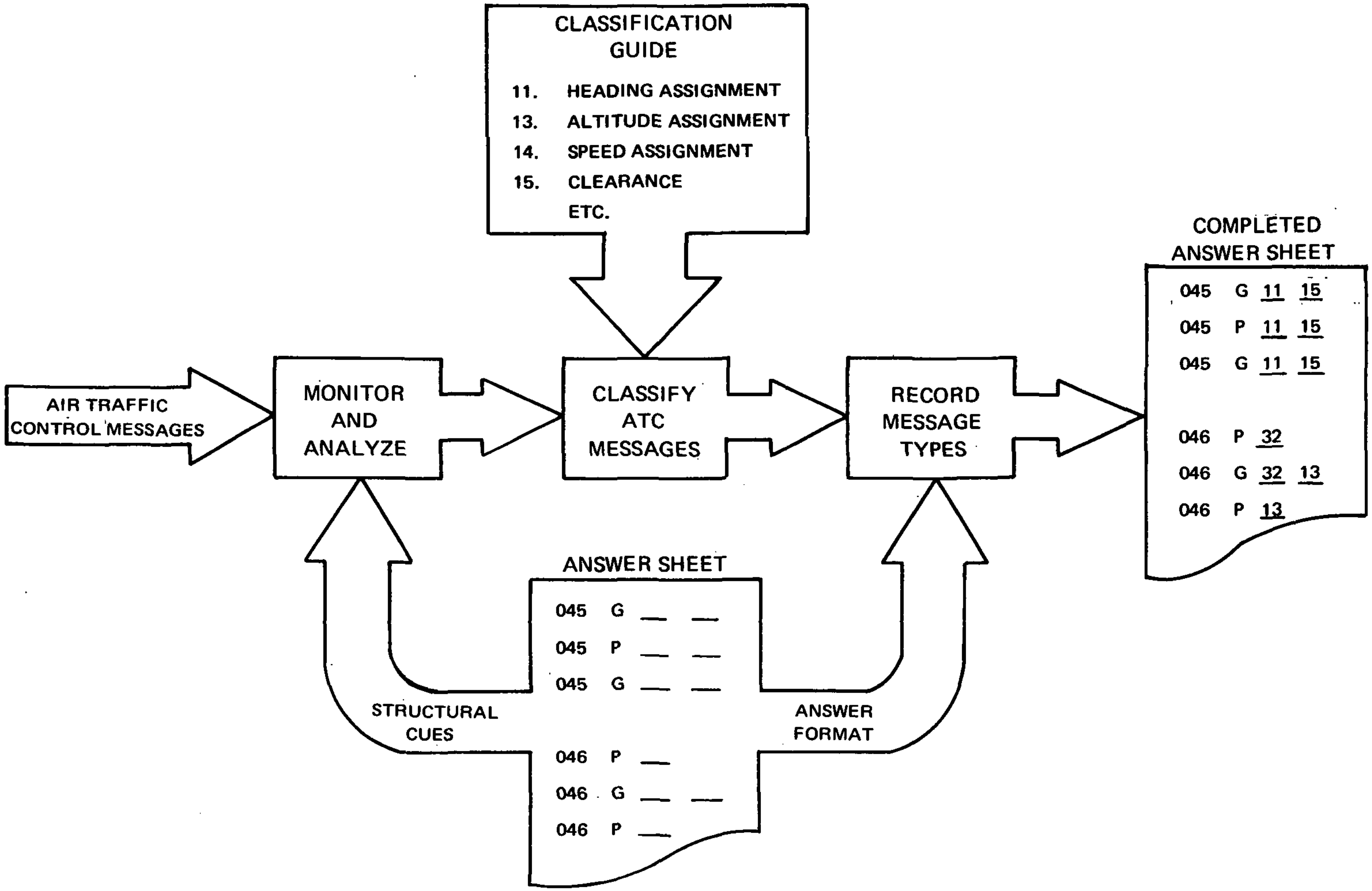


Figure 2. Schematic Diagram of Experimental Task

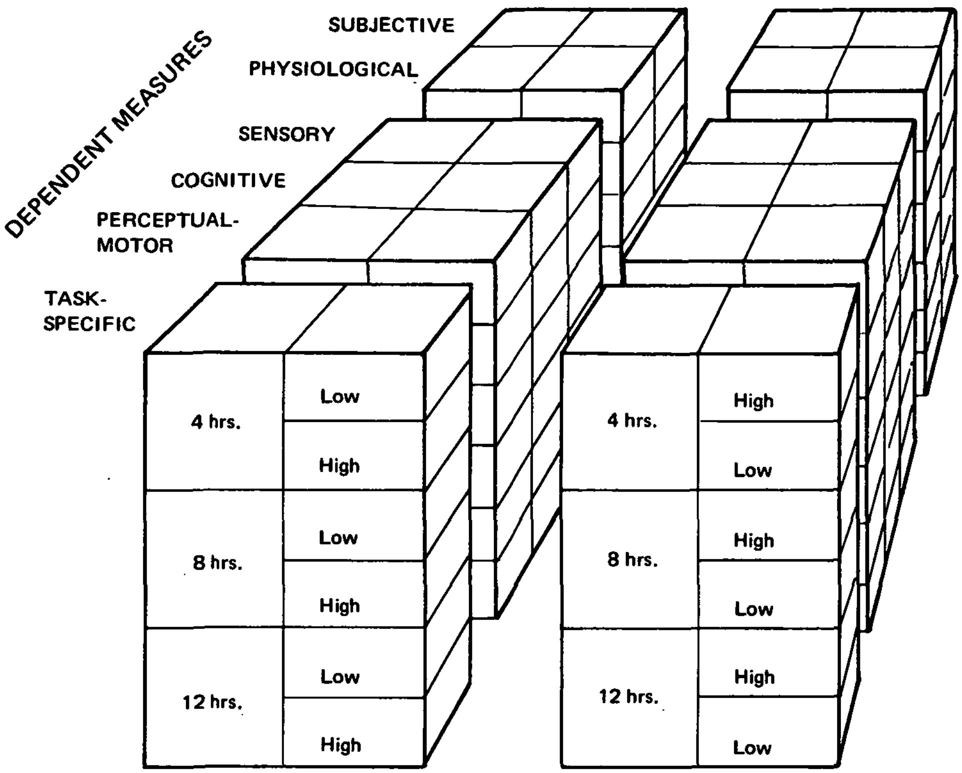


Figure 3. Experimental Design

Classes of Dependent Measures

Task Specific Measure

The task-specific measure was the subject's performance in classifying air traffic control messages. The standard of comparison was the classification of the same material by air traffic communications experts at NAFEC. The subject's score was the percentage of transmissions encoded correctly, based on:

a. Completeness – each message in the transmission had to be assigned a classification number.

b. Correspondence – the subject's classifications had to agree with those of the NAFEC experts.

c. Sequence – the order of the two-digit classification numbers had to correspond to the order in which the messages were actually transmittted.

For a subject's response to be counted "correct," he had to classify all messages in the transmission exactly as NAFEC had classified them and in the same temporal sequence. Any response not meeting all three criteria was counted "incorrect."

Laboratory Performance Measures

The pre- and postshift test battery consisting of three types of "laboratory" performance measures (cognitive, perceptual-motor, and sensory) was administered. These measures were selected according to the following criteria:

a. Factorial purity,

b. Experimental evidence supporting identification of the factor,

c. Range of ability levels covered,

d. Sensitivity to the influence of stresses associated with information processing tasks.

A total of nine individual measures were incorporated in the pre- and postshift test battery. Each is described below.

Cognitive Measures.

Number Facility –This was a test of ability to perform simple arithmetic computations. The performance standards were correctness and speed (the number completed in a fixed time interval).

Number Comparison – This was a measure of the ability to judge pairs of numbers as same or different. Performance was measured in terms of the number of items completed in a fixed time interval.

Perceptual-Motor Measures.

Visual Reaction Time – The performance measured was the time interval between the onset of a visual stimulus and the occurrence of a motor response.

Auditory Reaction Time – The performance measured was the time interval between the onset of an auditory stimulus and the occurrence of a motor response.

Response Orientation – The emphasis of this test was on the subject's ability to produce a discrete directional response to a nonspatial (nondirectional) stimulus. Specifically, the subject was required to move a toggle switch in the appropriate direction in response to each of four colored light stimuli. The measured performance was cumulative response time to a sequence of 24 events.

Arm-Hand Steadiness – In this test the subject was required to maintain his fully extended arm and hand in a steady state. Upon signal, the subject extended his arm, inserted a stylus in a hole and attempted to maintain it there without touching the rim of the opening. The performance measure was the number of contacts with the rim accumulated over three 10-second trials.

Perceptual Speed – This test measured the subject's ability to scan a complex display and make a judgment on the information presented. Specifically, the subject was presented with a series of indications on two meters. His task was to determine as rapidly as possible whether the two indications were the same or different and respond by pressing a key appropriate to each category. If the response was correct, the next pair of indications appeared. If the response was incorrect, an error was registered on a counter. The performance measures were time to complete a series of 24 presentations and the number of response errors.

Time Sharing – This was a test of the ability to monitor two displays which could not be viewed simultaneously in an attempt to detect deviation from a standard condition. The subject's task was to scan two separated meters, to detect the onset of movement of the pointer on one or the other, and to respond by pressing a key corresponding to that meter. The measure was cumulative detection time for 24 events over a four-minute period.

Sensory Measure.

Critical Flicker Fusion – CFF was a measure of the ability to discriminate the presence or absence of flicker in a light source. The subject was required to observe a light source with one eye and to indicate the onset of flicker in a steady light source or, alternatively, the point of transition from a flickering to a steady light. The subject was given twelve alternating ascending and descending trials. The measure was the mean CFF frequency for the last ten trials.

Physiological Performance Measures

Measures of heart rate (pulse) and oral temperature were taken near the end of the hourly work unit while the subjects were engaged in the primary coding task. While it was recognized that these measures could by no means present a definitive picture of physiological response to workload, it was hoped that they might provide at least an indication, at a gross level, of any changes in energy expenditure with increasing workload. Heart rate, for example, is known to be a good measure of work performance, particularly when any physical effort is involved. There also is evidence (Dahl & Spence, 1970) that heart rate reflects changes in task demands for cognitive and information processing activities.

Ratings of Subjective Response to Workload

Subjective ratings were taken as the subjects completed each hour's work. The ratings called for subjects to estimate their own state of fatigue and tension and the difficulty of processing the air traffic control communications presented in the previous hour. The latter rating entailed consideration of such factors as the pace of presentation and the intelligibility of the material, but in general it was the overall assessment of task difficulty which was of prime interest.

The subjective estimates were obtained using the cross modality and ratio scaling techniques developed by Stevens (1961, 1962). Stevens originally demonstrated that for many sensory qualities, the subjectively perceived magnitude of the phenomenon being observed is a power function of the physical magnitude of the sensation-producing stimulus. Later work by Stevens (1966a, 1966b) and others (Versace, 1963; Shoenberger & Harris, 1971, for example) has extended the applicability of the technique from simple sensory and psychophysical events to more complex areas of judgment. For instance, the power function has been shown to apply to complex qualities such as heaviness, whole-body vibration, and ride comfort. Its applicability even appears to include nonsensory and purely judgmental areas such as occupational preference and strength of expressed attitude. Experiments with this technique clearly demonstrate that ratio scaling permits quantitative access to very complex human performance.

In the present experiment, three subjective ratings of workload related factors (difficulty, fatigue, and tension) were obtained by the technique of free number matching to estimated magnitude. Three variations of the basic technique were used. For describing the difficulty of the information processing task, subjects were asked to assign a numerical rating. No constraint was placed on the range of values to be used or on the number and size of the increments. However, a direction was imposed on the scale in that the subjects were advised that material of greater difficulty should be assigned a higher number. For estimating fatigue, subjects were requested to draw a line whose length corresponded to their perceived amount of fatigue. Again, no constraint was placed on the scale except the physical limitations of the paper on which the line was drawn. Ratings of tension were obtained by having the subjects adjust the rate of a flashing light to correspond to their degree of "relaxation." A directional sense was indicated by instructing the subjects to let slower flash rates correspond to more relaxed feelings. Since the control knob of the device was not associated with a scale, subjects had no quantitative indication except that provided by the extremes of knob rotation.

CHAPTER 3

DESCRIPTION OF EXPERIMENT

Site

The experimental site was an area approximately 16 x 25 feet, partitioned into four subject's stations, an experimenter's station, and a separately enclosed room for administering perceptual-motor tests. A floor plan of the experimental site is shown in Figure 4.

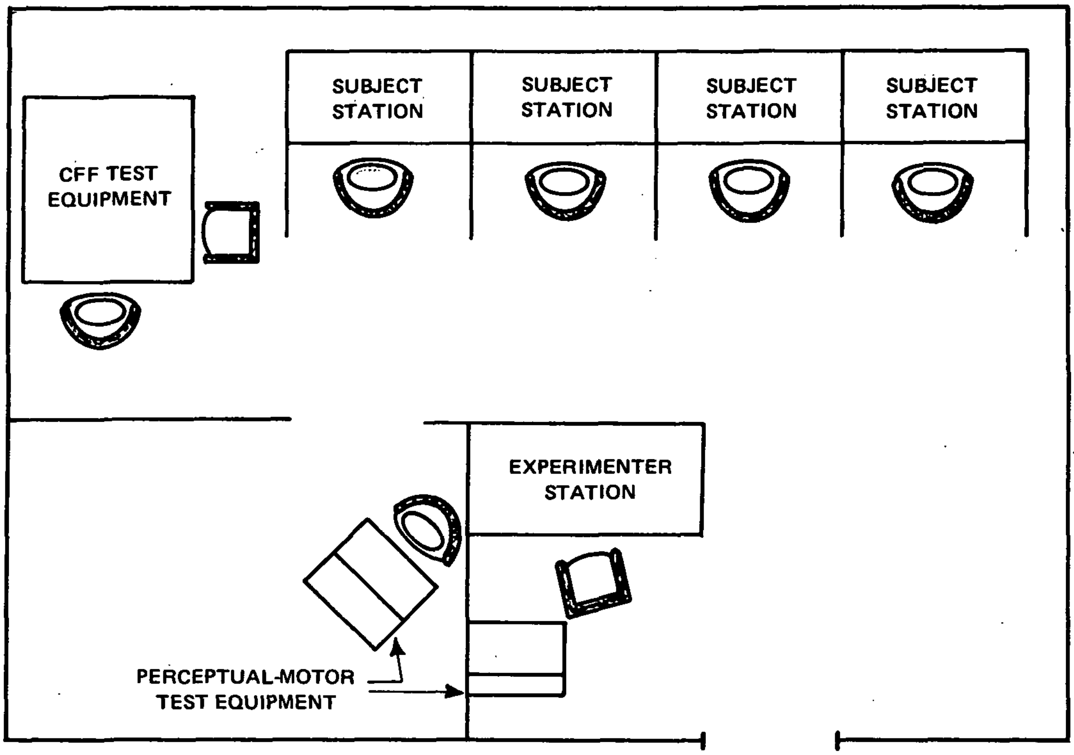


Figure 4. Floor Plan of Experimental Site

Each subject's station was a cubicle (4 ft. x 4 ft.) containing a built-in desk, a chair, and a desk lamp. Each station was equipped with headphones connected to a terminal box, with an individual control for adjusting the audio level of the stimulus material.

In general, the site afforded the subjects a comfortable working environment. The area was not sound-proofed, but extraneous noise was kept to a minimum, and factors which might have disrupted the subjects' concentration on the primary encoding task were controlled.

Experimental Apparatus and Test Instruments

The stimulus materials for the primary coding task were recorded on two-track stereo magnetic tape. One track contained the air traffic control messages, and the other contained the transaction identifiers which had been added to help the subjects keep pace with the stimulus material. By means of a two-channel tape deck and an amplifier, the two tracks were mixed and presented monaurally at the subjects' stations. Except for volume controls to adjust the sound level in individual headphones, operation of the audio equipment was controlled from the experimenter's station.

The six tests comprising the perceptual-motor portion of the pre- and postshift test battery were administered by means of a semi-automatic testing device designed and built by BioTechnology, Inc. under contract to NASA in 1965. The equipment permitted the experimenter to select and initiate the automatic presentation of stimuli for each of the tests in the battery. Readouts of the scores were presented on the experimenter's console. Photographs of the perceptual-motor test apparatus appear in Figure 5. A complete description of the equipment and its operation is contained elsewhere (Reilly & Parker, 1967).

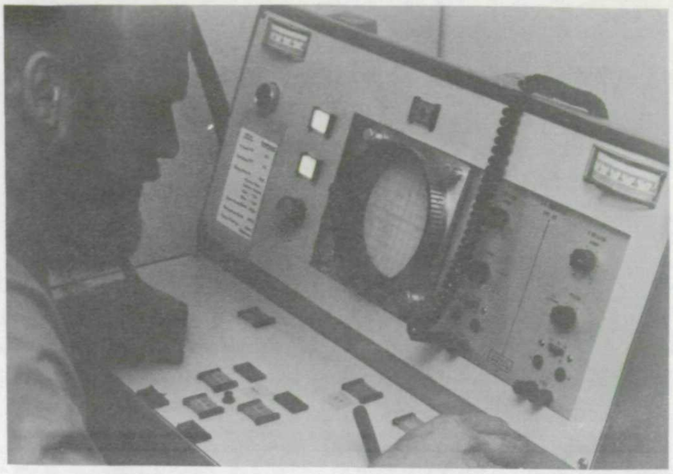
Monocular critical flicker fusion frequency (CFF) was measured by a device manufactured by Lafayette Instrument Company (equipment model 1202). The equipment provided a continuously variable frequency selection from 2 to 128 flashes per second, with equal on and off times in each flash cycle. The light source in the viewing chamber subtended $2^{\circ}10'$ of visual arc on the retina, which assured full foveal stimulation. Figure 6 is a photograph of the apparatus as it was used in the experiment.

The cognitive tests in the pre- and postshift battery (Number Facility and Number Comparison) were paper and pencil tests drawn from the kit of reference tests for cognitive factors compiled by French et al. (1963). Twelve equivalent forms of each test were used, one for each preshift and postshift test session.

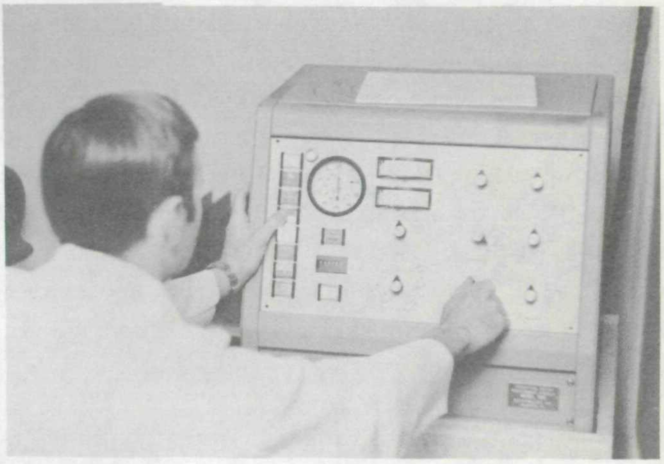
Subjective ratings of fatigue, tension, and workload were obtained through use of a questionnaire. The rationale for this type of rating and the technique of administration were discussed in Chapter 2. Figure 7 is a replica of the questionnaire used in the experiment.

Subjects

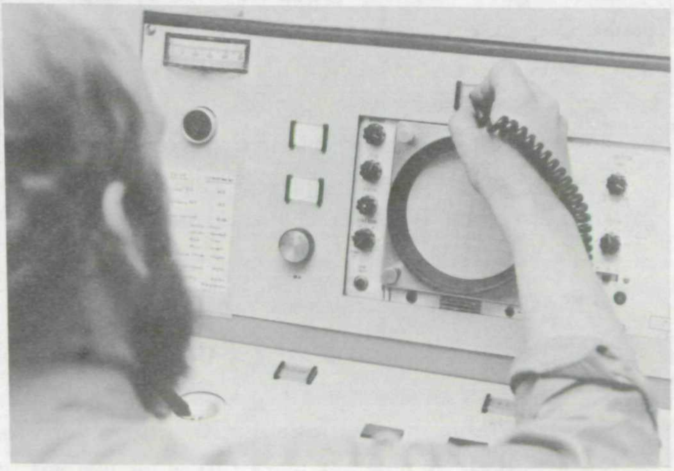
The subjects were 12 males, ranging in age from 19 to 24. Median age for the group was 21.5 years. Median educational level (last grade completed) was 14.8 years. The subjects had no prior experience in air traffic control or in aircraft piloting. None had any familiarity with monitoring radio transmissions (e.g., as a ham radio operator or radio dispatcher). None of the subjects had any reported hearing impairment; and during the study, there were no indications of difficulty in hearing the stimulus material.



SUBJECT'S
CONSOLE



EXPERIMENTER'S
CONSOLE



PERCEPTUAL-MOTOR
TEST IN PROGRESS

Figure 5. Perceptual-Motor Test Apparatus

Figure 6. CFF Measurement Equipment

Training

Because of the subjects' unfamiliarity with air traffic control in general and the specific requirements of the encoding task, a three-day program of training was given to each group of six before the experimental runs were begun. The training period was devoted mainly to instruction and practice in encoding air traffic control messages. However, a brief period of familiarization and practice was also allocated to tasks associated with the other dependent measures.

Instruction and guided practice in the encoding task was accomplished in approximately 15 hours, distributed over a three-day period. This included two hours of orientation to air traffic control and a briefing on the mechanics of the encoding task, 11 hours of group instruction and individual practice in coding the various messages types, and two hours of "dress rehearsal" for the experimental runs. During the "dress rehearsal" on the final day of training, the subjects were exposed to the work/rest cycle and the task requirements which would obtain during the experimental run.

Training in the tasks associated with the nontask-specific dependent measures consumed a total of about one and a half hours for each group of subjects. Instruction consisted of group demonstration and an individual practice trial for each element of the dependent measures battery.

Subject: ____________ Date: ______ Time: _____ 4 8 12 L H

1. Assign a number to the workload imposed by the material you have been working with during the past hour. Take into consideration such factors as how much difficulty you had keeping up with the material, and how intelligible it was, but the number you assign should be your overall assessment of the difficulty of the past hour's material as a whole. The higher the number, the more difficult the material. Write the number here ____________.

2. Draw a straight horizontal line that indicates how tired you feel. Let the length of the line equal the amount of tiredness, i.e., the more tired you feel, the longer the line. Draw the line directly below.

3. Set the light to indicate how relaxed you feel. Let the number of flashes equal the amount of relaxation such that the slower the light blinks, the more relaxed you feel.

Figure 7. Form for Obtaining Subjective Workload Ratings

Schedule

The experimental design called for all subjects to work three different shift lengths at each of the two communication densities. Subjects were divided into two groups of six, which were run separately. Each group participated in six consecutive days of experimental runs. On each day, all six subjects in the group worked at the same density but for varying shift lengths. The working hours were arranged so that only four subjects in the group were on duty at any given time. Figure 8 is a schematic representation of the overall schedule of experimentation.

The subjects' working day began at 0800 hours with the administration of the preshift test battery. At 0900, subjects started on the primary coding task, working through to the completion of their assigned 4-, 8-, or 12-hour shift. After each 45 minutes of work, the subjects made three subjective ratings (workload, fatigue, and tension) and then were given a 10-minute break. After each four hours of work, subjects scheduled to continue into the next 4-hour period were given a 45-minute rest period for food and refreshment. Those who were ending or starting work at that hour were administered the pre- or postshift test battery. The work day ended at approximately 2230 with completion of the postshift test battery for the four subjects on duty at that time. Thus, a subject scheduled for a 12-hour shift was actually on duty for slightly over fourteen hours. Figure 9 is a sample schedule for a subject during the week of experimental runs.

The schedule for pre- and postshift dependent measures was controlled with respect to the order in which the tests were administered and their relation to the subject's work schedule for that day. Figure 10 shows the daily schedule. Figure 11 shows the sequence of testing and the time allotted for the perceptual-motor, sensory, and cognitive portions of the pre- and postshift test battery.

Experimental Procedures and Administration

The only activity required of subjects during their work shift was performance of the basic encoding task. As explained in Chapter 2, the subjects' specific mode of response was to record the classification of messages as a series of two-digit numbers on a specially prepared answer sheet.

About five minutes before the end of each 45-minute work period, the subjects' pulse rate and oral temperature were taken. At the conclusion of the 45-minute work period, the subjects filled out a form asking for three subjective estimates of the workload imposed during the past hour.

The pre- and postshift test battery was administered to two subjects at a time. One subject was tested on the preceptual-motor portion of the battery while the other was given the CFF and cognitive tests. The two subjects then exchanged places to complete the battery.

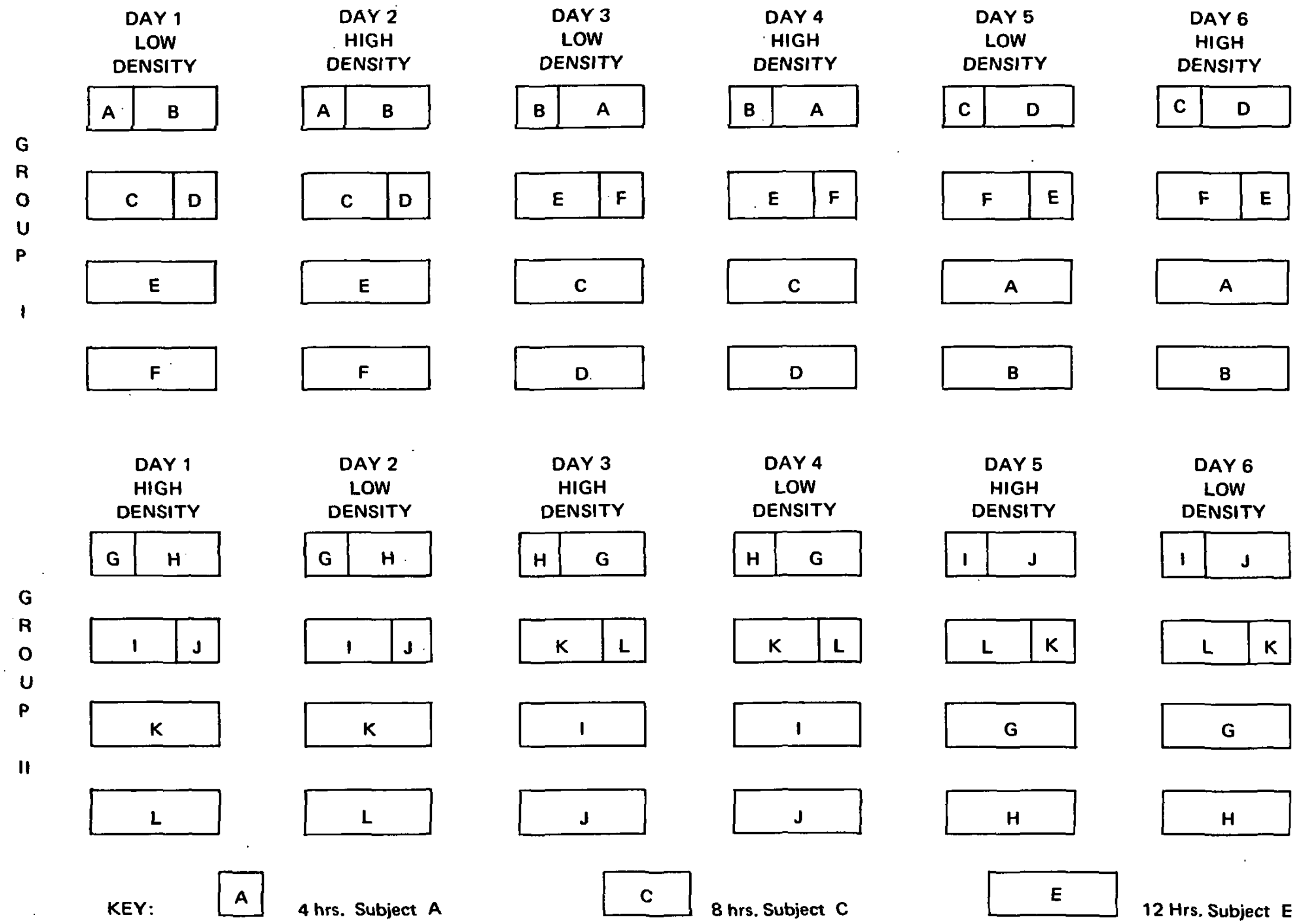


Figure 8. Experimental Schedule

NAME Tom

GROUP I SUBJECT C

	DAY 1	DAY 2	DAY 3	DAY 4	DAY 5	DAY 6
0820 – 0900	PRE TEST	PRE TEST			PRE TEST	PRE TEST
0900 – 0950	EXP. RUN	EXP. RUN			EXP. RUN	EXP. RUN
0950 – 1000	BREAK	BREAK			BREAK	BREAK
1000 – 1050	EXP. RUN	EXP. RUN			EXP. RUN	EXP. RUN
1050 – 1100	BREAK	BREAK			BREAK	BREAK
1100 – 1150	EXP. RUN	EXP. RUN			EXP. RUN	EXP. RUN
1150 – 1200	BREAK	BREAK			BREAK	BREAK
1200 – 1250	EXP. RUN	EXP. RUN			EXP. RUN	EXP. RUN
1250 – 1330	POST TEST	POST TEST	PRE TEST	PRE TEST	LUNCH	LUNCH
1330 – 1420			EXP. RUN	EXP. RUN	EXP. RUN	EXP. RUN
1420 – 1430			BREAK	BREAK	BREAK	BREAK
1430 – 1520			EXP. RUN	EXP. RUN	EXP. RUN	EXP. RUN
1520 – 1530			BREAK	BREAK	BREAK	BREAK
1530 – 1620			EXP. RUN	EXP. RUN	EXP. RUN	EXP. RUN
1620 – 1630			BREAK	BREAK	BREAK	BREAK
1630 – 1720			EXP. RUN	EXP. RUN	EXP. RUN	EXP. RUN
1720 – 1800			SNACK	SNACK	SNACK	SNACK
1800 – 1850			EXP. RUN	EXP. RUN	EXP. RUN	EXP. RUN
1850 – 1900			BREAK	BREAK	BREAK	BREAK
1900 – 1950			EXP. RUN	EXP. RUN	EXP. RUN	EXP. RUN
1950 – 2000			BREAK	BREAK	BREAK	BREAK
2000 – 2050			EXP. RUN	EXP. RUN	EXP. RUN	EXP. RUN
2050 – 2100			BREAK	BREAK	BREAK	BREAK
2100 – 2150			EXP. RUN	EXP. RUN	EXP. RUN	EXP. RUN
2150 – 2230			POST TEST	POST TEST	POST TEST	POST TEST

Figure 9. Sample Schedule for Subject

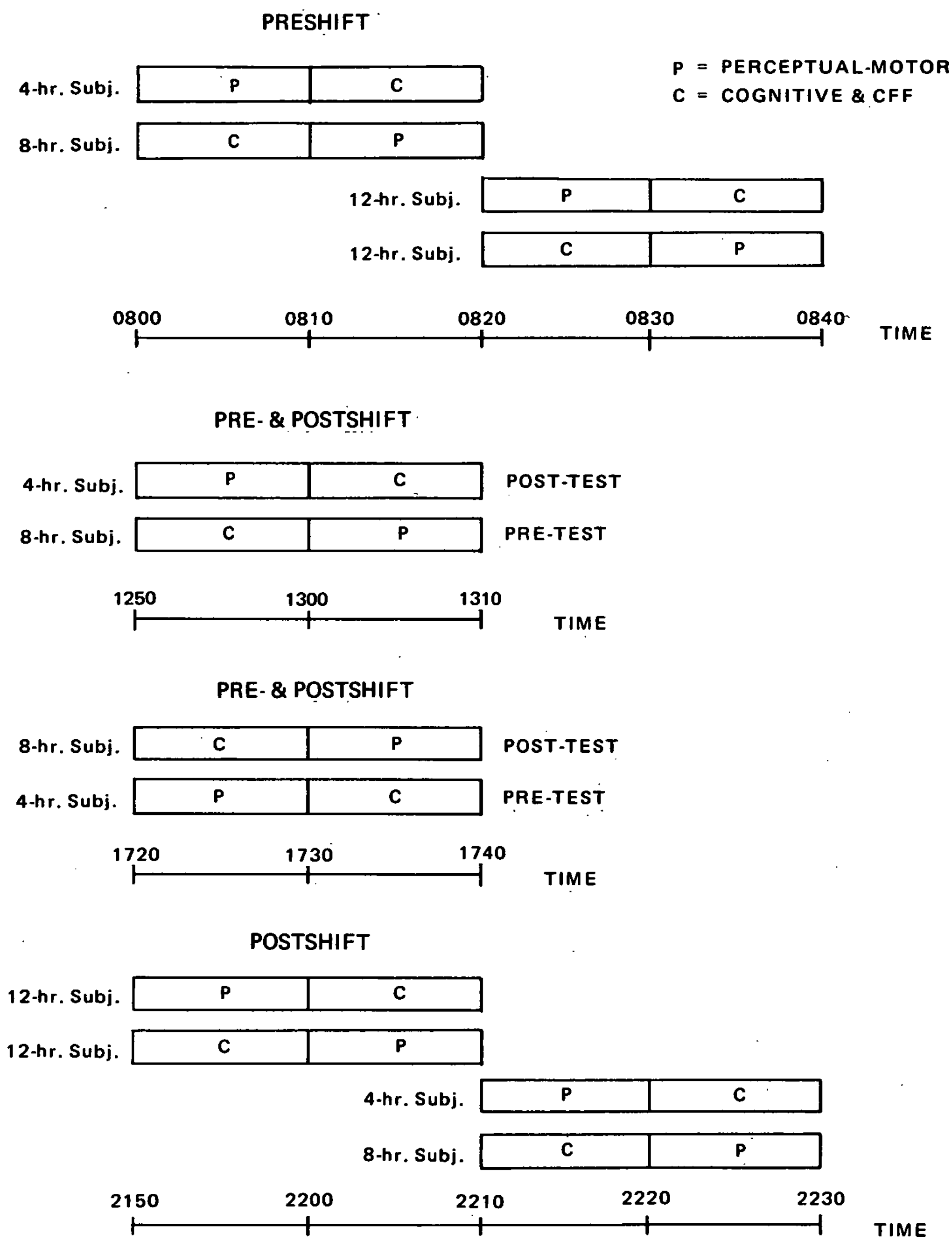
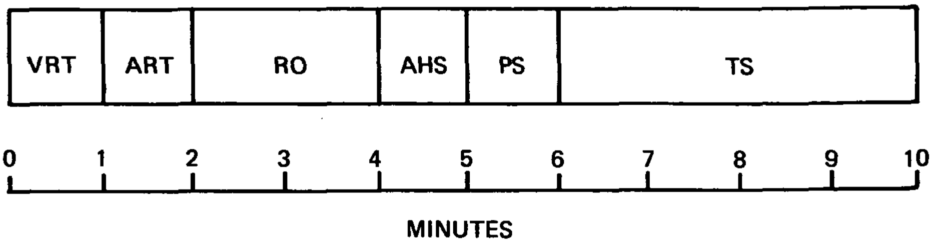


Figure 10. Daily Schedule of Pre- and Postshift Tests

PERCEPTUAL–MOTOR



VRT – VISUAL REACTION TIME

ART – AUDITORY REACTION TIME

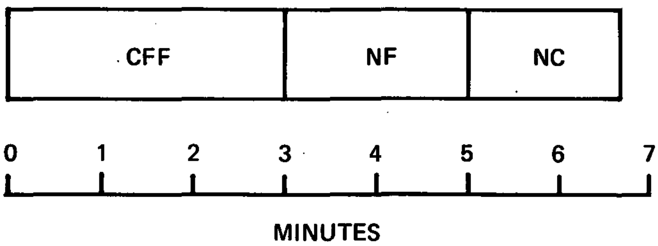
RO – RESPONSE ORIENTATION

AHS – ARM-HAND STEADINESS

PS – PERCEPTUAL SPEED

TS – TIME SHARING

COGNITIVE AND CFF



CFF – CRITICAL FLICKER FUSION

NF – NUMBER FACILITY

NC – NUMBER COMPARISON

Figure 11. Schedule of Pre- and Postshift Test Battery

A statement of the purpose of the experiment and the instructions for the pre- and postshift tests, as they were presented to the subjects, are contained in Appendix B. The subjects were told that the purpose of the experiment was to study the effectiveness of the main coding task as a training technique. No mention was made of any expected performance degradation as a result of shift length or density, and the interest in fatigue effects resulting from workload was not emphasized. This was done to minimize performance effects which might arise through suggestion and to encourage the subjects to exert a full and sincere effort throughout the work day and work week.

During the course of the experiment, the subjects were encouraged to follow their normal behavior patterns, insofar as possible. Smoking was permitted except during the pre- and postshift test battery and for 10 minutes before oral temperature was taken. Subjects were asked to abstain from beverages containing caffein while working and during the 10-minute rest periods. Coffee and tea were permitted at meal times. Subjects were also asked to refrain from discussing with each other the details of the air traffic control messages and the ratings they had assigned to workload, fatigue, and tension.

During the six-day period of experimentation, the subjects were lodged in quarters provided by BioTechnology, Inc. This arrangement was made to provide control of the subjects' activities during off-duty hours and to assist in maintaining the rather strict work schedule called for by the experimental design. Subjects were also provided meals during the six days of experimentation. No major restrictions were imposed on off-duty activities, except abstention from alcohol and curtailment of study or recreational activities which might interfere with adequate sleep. In addition to lodging and meals, subjects were paid a stipend, contingent upon completion of the experiment.

CHAPTER 4

DATA ANALYSIS AND RESULTS

Information Processing Task

Performance on the primary information processing task was scored by comparing the subject's classification of each message with that assigned by NAFEC personnel. Individual message scores were then summed to the transmission level. The subject's classification of the transmission was considered correct only if all messages therein were classified, as NAFEC has classified them, and in the same sequence. Thus, if the NAFEC classification of a transmission was 13 32, only the last of the following examples of subject's responses would be correct:

	32	(omission)
13	33	(wrong classification)
32	13	(improper sequence)
13	32	(correct)

The resulting raw score (number of transmissions correct per 45-minute work period*) was converted to percentage form to permit later comparison among work periods.

A second performance score was obtained by summing the raw transmission-correct score to the transaction level. A subject's classification of a transaction was counted correct only if all transmissions therein were classified completely and correctly. The resulting transactions-correct score was also converted to percentage form. Since scoring at the transaction level represented a more stringent standard of performance than at the transmission level, there was some question initially as to which was the appropriate measure. The correlation coefficient (Pearson product moment) between transmission and transaction scores for all subjects under all conditions was calculated and found to be 0.9479. Therefore, subsequent analyses were based on only a single score, percentage transmissions correct per hour.

Since this study was concerned with the overall effect on performance produced by different shift lengths and communication densities, the individual subject's scores on only the first and last hours for each day were compared. Initial inspection of the data indicated that some variation in performance occurred simply as a result of the experiment having been run on successive days (practice effect). As a next step, therefore, a correction factor (the difference between the daily mean and the grand mean for the six-day experimental period) was applied to individual scores to

*For simplicity, a 45-minute work period is referred to hereafter as an hour.

compensate for the practice effect. It was assumed that the practice effect was part of the error of measurement and did not represent a systematic source of variation resulting from the operation of the independent variables.

Table III is a summary of the means and standard deviations of differences of first and last hour scores for each condition of shift length and communication density. The scores are expressed as percentages, and negative differences indicate a performance decrement from the first to the last hour. It can be seen that performance differences between the first and last hours were small; all were less than two percentage points. Moreover, the overall range of performance was quite narrow, and no systematic pattern of variation was present. These observations were confirmed through an analysis of variance which indicated that none of the differences was significant. (See Table VIII, Appendix C.)

TABLE III

Summary of Scores for the Information Processing Task

	4 HR. SHIFT		8 HR. SHIFT		12 HR. SHIFT	
	FIRST HOUR	LAST HOUR	FIRST HOUR	LAST HOUR	FIRST HOUR	LAST HOUR
LOW DENSITY						
$\overline{X}$	72.7	73.9	74.5	74.6	73.2	73.9
S.D.	9.6	7.4	10.7	5.9	9.6	7.5
Diff.	1.2		0.1		0.7	
HIGH DENSITY						
$\overline{X}$	73.4	75.0	74.6	74.2	75.2	75.4
S.D.	10.9	9.2	7.1	6.8	9.3	7.7
Diff.	1.6		–0.4		0.2	

N = 12 $\overline{X}$ = Mean S.D. = Standard Deviation Diff. = Difference of first and last hour means

Laboratory Performance Measures

Perceptual-Motor

The perceptual-motor measures consisted of six individual tests administered on a pre- and postshift schedule. The required performance, duration, and scoring of these tests are summarized in Table IV.

TABLE IV

Perceptual-Motor Test Battery

TEST	PERFORMANCE	DURATION	SCORE
Visual Reaction Time	Press switch in response to light.	Four trials in one minute	Mean reaction time based on four trials
Auditory Reaction Time	Press switch in response to tone.	Four trials in one minute	Mean reaction time based on four trials
Response Orientation	Move toggle switch in appropriate direction in response to each of four colored light stimuli.	Two minutes	Cumulative response time to sequence of 24 signals
Arm-Hand Steadiness	Hold tip of stylus in aperture with arm and hand fully extended.	Three 10-second trials equally spaced over one minute	Total contacts with rim of aperture during three 10-second trials
Perceptual Speed	Judge pair of meter settings as same or different.	One minute	Time to complete sequence of 24 presentations and number of errors
Time Sharing	Monitor two meters to detect random onset of pointer movement.	Four minutes	Cumulative response time for 24 events

Means and standard deviations of performance scores for each test and each experimental condition are presented in Table V. In all cases, an increase in the score from the preshift to the postshift administration constituted a performance decrement. However, the scores in the difference column of Table V are expressed such that negative values indicate a performance decrement. This reversal was made to facilitate comparison of these results with information processing task measures and with other elements of the pre- and postshift test battery, where a higher score indicated better performance.

Preliminary analysis revealed no apparent pattern in the magnitude or direction of performance changes and no clear relationship between perceptual-motor performance and the independent variables. In general, perceptual-motor test scores were characterized by inconsistency within very narrow limits, i.e., the magnitude of change from condition to condition was small regardless of which measure one examines. Analyses of variance for each of the tests revealed that, with a few negligible exceptions, the differences between preshift and postshift scores were not significant. (See Tables IX through XIV, Appendix C.)

TABLE V

Summary of Perceptual-Motor Test Scores

TEST		LOW DENSITY						HIGH DENSITY					
		4 HR. SHIFT		8 HR. SHIFT		12 HR. SHIFT		4 HR. SHIFT		8 HR. SHIFT		12 HR. SHIFT	
		PRE	POST	PRE	POST	PRE	POST	PRE	POST	PRE	POST	PRE	POST
Visual Reaction Time (Seconds)	$\overline{X}$	0.1750	0.1594	0.1627	0.1471	0.1744	0.1676	0.1615	0.1604	0.1666	0.1692	0.1608	0.1660
	S.D.	0.0380	0.0198	0.0366	0.0178	0.0356	0.0375	0.0259	0.0313	0.0304	0.0313	0.0397	0.0495
	Diff.	0.0156		0.0156		0.0068		0.0011		–0.0026		–0.0052	
Auditory Reaction Time (Seconds)	$\overline{X}$	0.1652	0.1656	0.1675	0.1543	0.1618	0.1729	0.1758	0.1714	0.1645	0.1697	0.1531	0.1847
	S.D.	0.0374	0.0282	0.0346	0.0282	0.0316	0.0424	0.0435	0.0264	0.0282	0.0264	0.0244	0.0344
	Diff.	–0.0004		0.0132		–0.0111		0.0044		–0.0052		–0.0316	
Response Orientation (Seconds)	$\overline{X}$	12.32	11.38	11.01	12.00	11.32	11.01	11.18	11.08	11.71	11.02	10.75	10.62
	S.D.	3.48	3.79	1.89	3.08	1.60	1.91	2.89	2.58	3.60	2.02	2.28	1.55
	Diff.	0.94		–0.99		0.31		0.10		0.69		0.13	
Arm-Hand Steadiness (Contacts with rim)	$\overline{X}$	22.9	18.0	20.7	18.1	23.1	20.7	22.6	16.4	23.1	21.2	22.7	16.7
	S.D.	10.0	9.9	14.5	13.1	18.2	14.2	12.3	8.9	17.7	15.9	16.2	9.3
	Diff.	4.9		2.6		2.4		6.2		1.9		6.0	
Perceptual Speed (Seconds)	$\overline{X}$	40.66	39.79	44.20	39.24	40.87	40.95	40.59	38.42	40.92	39.37	40.53	38.84
	S.D.	7.93	5.71	6.11	6.38	5.87	6.00	5.00	4.83	5.09	3.57	3.75	2.81
	Diff.	0.87		4.96		–0.08		2.17		1.55		1.69	
Time Sharing (Seconds)	$\overline{X}$	22.58	22.99	23.68	23.57	23.24	22.02	22.46	22.79	24.61	23.15	23.63	22.41
	S.D.	2.41	3.43	1.89	2.08	3.73	2.03	2.51	4.00	2.91	3.45	3.57	3.99
	Diff.	–0.41		0.11		1.22		–0.33		1.46		1.22	

N = 12 $\overline{X}$ = Mean S.D. = Standard Deviation Diff. = Difference of Pre and Post Means

Cognitive

The two tests of cognitive capacity employed in the pre- and postshift battery were Number Facility and Number Comparison. The Number Facility test, which required the subjects to perform simple arithmetic problems, was scored in terms of the number of problems correctly completed in a two-minute period. Number Comparison, which involved judging whether pairs of numbers were the same or different, was scored by subtracting incorrect responses from correct responses.

Table VI is a summary of Number Facility and Number Comparison scores for each experimental condition. A decrease in the score from the preshift to the postshift administration indicates a performance decrement and is denoted in the difference column by a negative value.

The pattern of cognitive test scores was one of great stability across all experimental conditions. The range of mean low scores or mean high scores was only about three items for either test, and the difference between preshift and postshift means for any experimental condition was less than two items in all but one case. Analysis of variance indicated that none of the differences was significant. (See Tables XV and XVI, Appendix C.)

Sensory

Critical fusion frequency (CFF) was measured before and after each shift. Measurement consisted of twelve trials, alternatively ascending and descending. The subject's score was the mean CFF for the last 10 trials.

CFF results are summarized in Table VI. A negative difference between preshift and postshift measurements denotes a decrease of CFF and is to be interpreted as a lessened ability to discriminate flicker and a performance decrement. An analysis of variance (Table XVII, Appendix C) indicated that only the difference between preshift and postshift performance as a function of density was significant ($p < 0.05$). The data further indicated that low density conditions resulted in a greater reduction of CFF than high density conditions.

A secondary result, not related to the purpose of this study, was also found. As the six-day experimental period progressed, subjects generally exhibited a rise in CFF. This phenomenon, which was interpreted to be a learning effect, has been authoritatively treated by Knox (1945) who indicated that CFF rises slightly with practice providing that subjects are not given a specific set and providing one sequence of stimulus presentation (e.g., fusion to flicker) is not favored. These latter variables exhibit an interaction effect with practice such that CFF will be higher if the individual's set is for flicker and lower if the set is for fusion. Further, CFF will be higher if the stimulus sequence is fusion to flicker (decreasing flash frequency) and lower if the sequence is flicker to fusion (increasing flash frequency).

TABLE VI

Summary of Cognitive and Sensory Test Scores

COGNITIVE

TEST		LOW DENSITY						HIGH DENSITY					
		4 HR. SHIFT		8 HR. SHIFT		12 HR. SHIFT		4 HR. SHIFT		8 HR. SHIFT		12 HR. SHIFT	
		PRE	POST	PRE	POST	PRE	POST	PRE	POST	PRE	POST	PRE	POST
Number Facility (Problem Correct)	$\overline{X}$	32.2	33.0	31.2	31.7	30.9	32.5	29.7	31.2	32.2	31.5	30.5	30.2
	S.D.	8.1	7.1	8.7	6.7	7.4	8.6	5.7	7.6	8.2	8.5	7.1	5.6
	Diff.	0.8		0.5		1.6		1.5		−0.7		−0.3	
Number Comparison (Correct Minus Incorrect Responses)	$\overline{X}$	24.6	27.9	26.7	25.2	25.6	26.7	27.2	28.3	26.1	26.1	25.2	26.9
	S.D.	5.8	6.1	5.2	4.6	6.7	5.0	5.9	4.1	6.0	6.0	5.9	6.1
	Diff.	3.3		−1.5		1.1		1.1		0.0		1.7	

SENSORY

TEST		4 HR. PRE	4 HR. POST	8 HR. PRE	8 HR. POST	12 HR. PRE	12 HR. POST	4 HR. PRE	4 HR. POST	8 HR. PRE	8 HR. POST	12 HR. PRE	12 HR. POST
Critical Fusion Frequency (Flashes per Second)	$\overline{X}$	48.80	47.34	47.42	47.37	48.65	48.01	48.12	47.86	47.73	48.06	47.92	48.58
	S.D.	1.90	2.08	2.29	2.84	3.24	2.72	2.57	1.89	2.64	2.10	2.14	2.48
	Diff.	−1.46		−0.05		−0.64		−0.26		0.33		0.66	

N = 12 $\overline{X}$ = Mean S.D. = Standard Deviation Diff. = Difference of Pre and Post Means

The experimental technique employed in the present study was selected with these findings in mind. The alternate ascending and descending trials were intended to counteract the effects of experiential inertia. To avoid establishing a set, neither flicker nor fusion were voluntarily stressed over the other in instructions to subjects. For these reasons, the observed rise in CFF during the week (during which each subject was exposed to a total of 144 trials) was concluded to be a residual effect attributable to practice alone. Appropriate adjustments were made in the experimental data before analysis for the effects of the independent variables of shift length and communication density.

Physiological Measures

Oral temperature and pulse rate were measured at one-hour intervals, near the end of each work period, throughout the experiment. Inspection of individual subject records revealed no marked deviation from normal which might be attributed to the experimental conditions. Those variations which did exist showed the characteristic form attributable to the general sleep-wakefulness cycle.

Initial and terminal values of oral temperature and pulse rate for each experimental condition are shown in Table VII, which also presents standard deviations and differences between first and last hour means for each measure.

Figure 12 is a plot of the obtained values of oral temperature set against a smoothed empirical cycle of diurnal variation in oral temperature. The deviations of the data obtained in this study from the generalized curve are attributable to sample size ($n = 12$).

Subjective Measures

Self-estimates of fatigue and tension using free number matching techniques were obtained from subjects before starting work and each hour thereafter. Similar subjective ratings of the difficulty of the stimulus material (task difficulty) were obtained after each hour of work. Geometric means were calculated for the 12 subjects' responses for each hour under each experimental condition. The results are presented in Figures 13 through 15. Least-squares fits were also calculated for the logarithms of the geometric means against each shift length and density combination. The fitted lines are shown in Figures 13 through 15, along with the standard deviation (σ_{yx}) and correlation coefficient (r) for each set of geometric means.

Figure 13 shows that the fatigue data plot as straight-line functions and that the least-squares fit is good. In all but one case (12 hours at high density), $r \approx 0.9$. This supports the hypothesis that the subjective relationship between fatigue and shift length is quite close and is readily perceived by individuals. Or in other words, estimates of fatigue are a power function of the length of time at work. This finding is in general agreement with the results obtained by Grandjean (1968), who collected self-estimates of fatigue and several related factors from air traffic controllers after various work periods.

TABLE VII

Summary of Oral Temperature and Pulse Measurements

MEASURE		LOW DENSITY						HIGH DENSITY					
		4 HR. SHIFT		8 HR. SHIFT		12 HR. SHIFT		4 HR. SHIFT		8 HR. SHIFT		12 HR. SHIFT	
		FIRST HOUR	LAST HOUR	FIRST HOUR	LAST HOUR	FIRST HOUR	LAST HOUR	FIRST HOUR	LAST HOUR	FIRST HOUR	LAST HOUR	FIRST HOUR	LAST HOUR
Oral Temperature	$\overline{X}$	98.5	98.6	98.4	98.7	98.0	98.6	98.5	98.6	98.4	98.8	98.2	98.6
	S.D.	0.51	0.53	0.51	0.50	0.74	0.55	0.59	0.44	0.55	0.32	0.58	0.44
	Diff.	0.1		0.3		0.6		0.1		0.4		0.4	
Pulse Rate	$\overline{X}$	81.8	79.0	82.0	80.2	77.8	75.3	80.8	79.7	81.7	81.3	77.3	76.3
	S.D.	7.9	7.6	10.7	8.8	6.6	9.0	8.4	10.4	9.4	6.9	8.1	7.2
	Diff.	−2.8		−1.8		−2.5		−1.1		−0.4		−1.0	

N = 12 $\overline{X}$ = Mean S.D. = Standard Deviation Diff. = Difference of first and last hour means

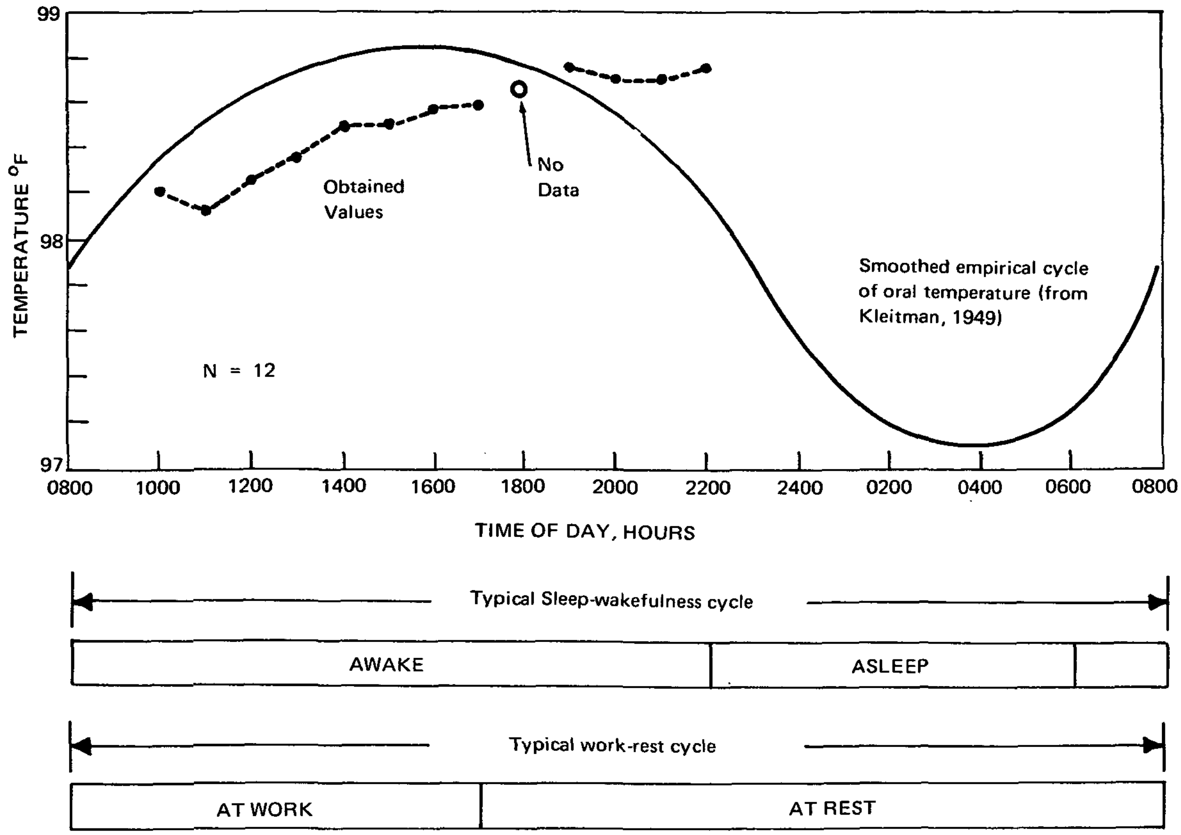


Figure 12. Diurnal Variation of Oral Temperature

Figure 14, which contains plots of perceived tension as a function of shift length, does not show so clear a relationship between the two factors. The correlations were generally weak and in several cases were little better than chance. From this, it may be concluded that tension was not perceived by subjects to be a function of shift length or that it was not recognized as a correlate (or constituent) of workload. A further discussion of this point is reserved for the next chapter.

Figure 15, difficulty of stimulus material vs. shift length, shows a somewhat unexpected result. The correlations between estimates of task difficulty and shift length were reasonably good under low density conditions, but very weak under high density conditions. From this, it would appear that density (or density in combination with shift length) is the more powerful determinant of estimated task difficulty. Additional support for this contention can be seen in Figure 16 which is a plot of the fatigue, tension, and task difficulty levels and gradients for each experimental condition.

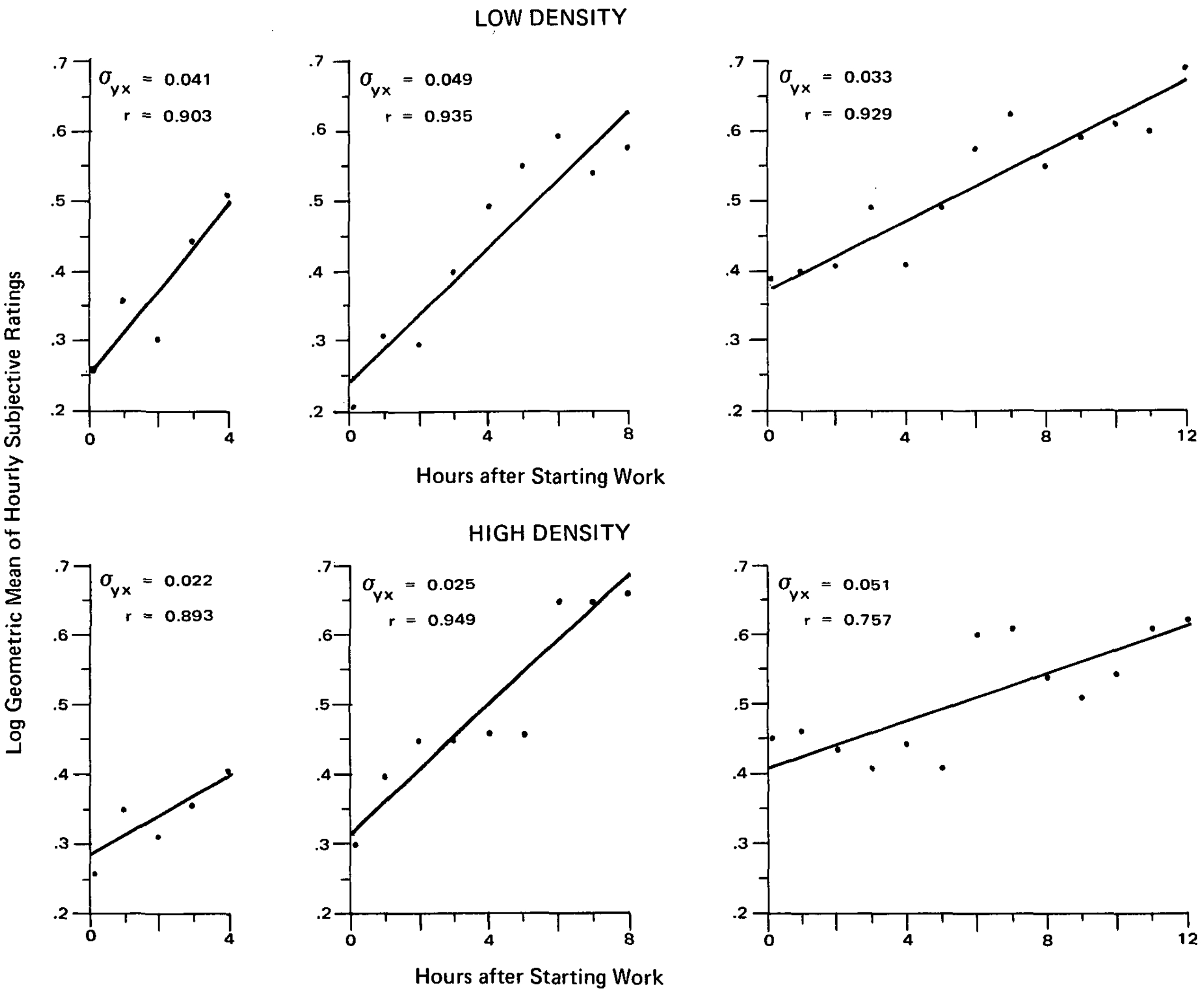


Figure 13. Subjective Estimates of Fatigue as a Function of Shift Length

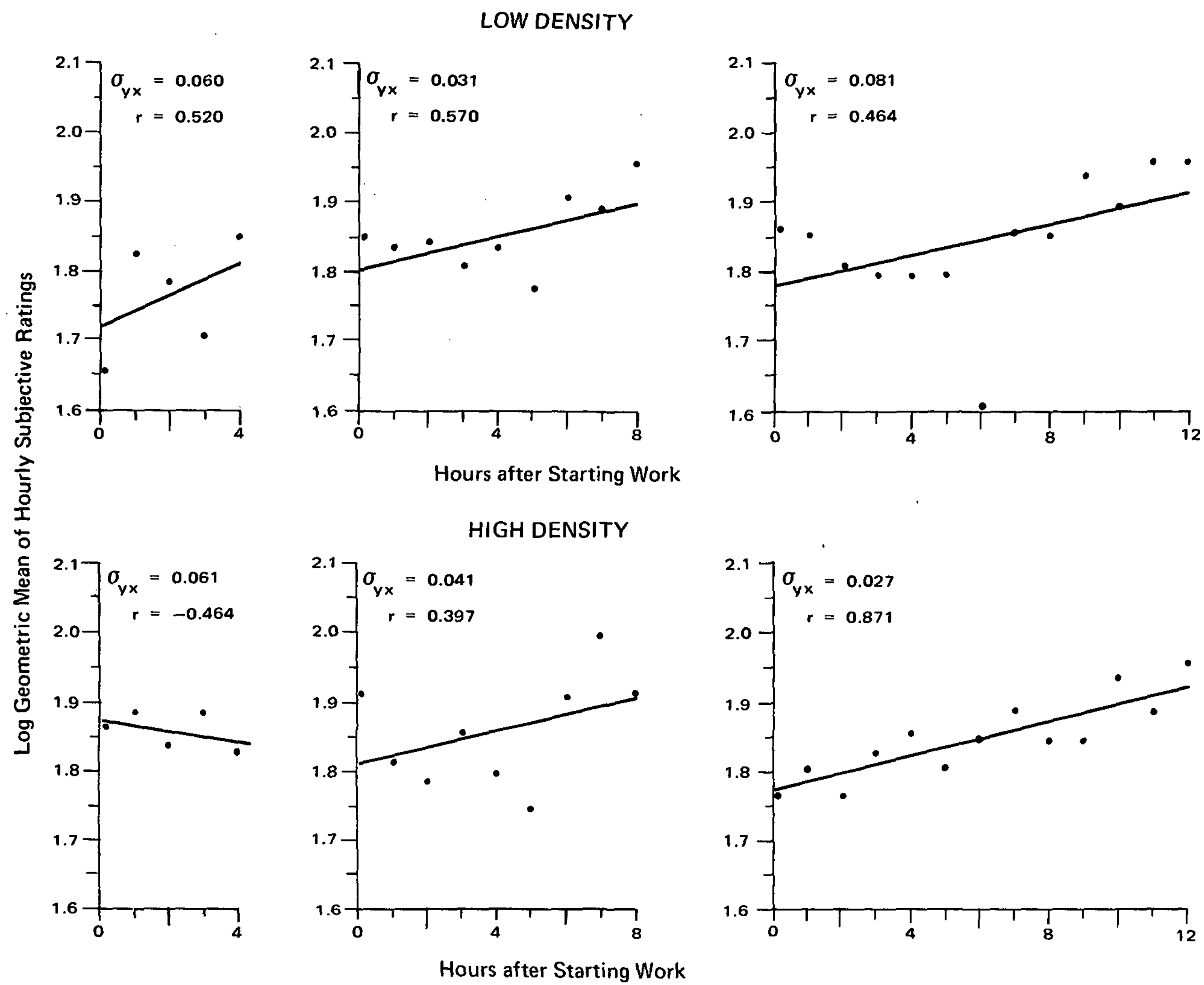


Figure 14. Subjective Estimates of Tension as a Function of Shift Length

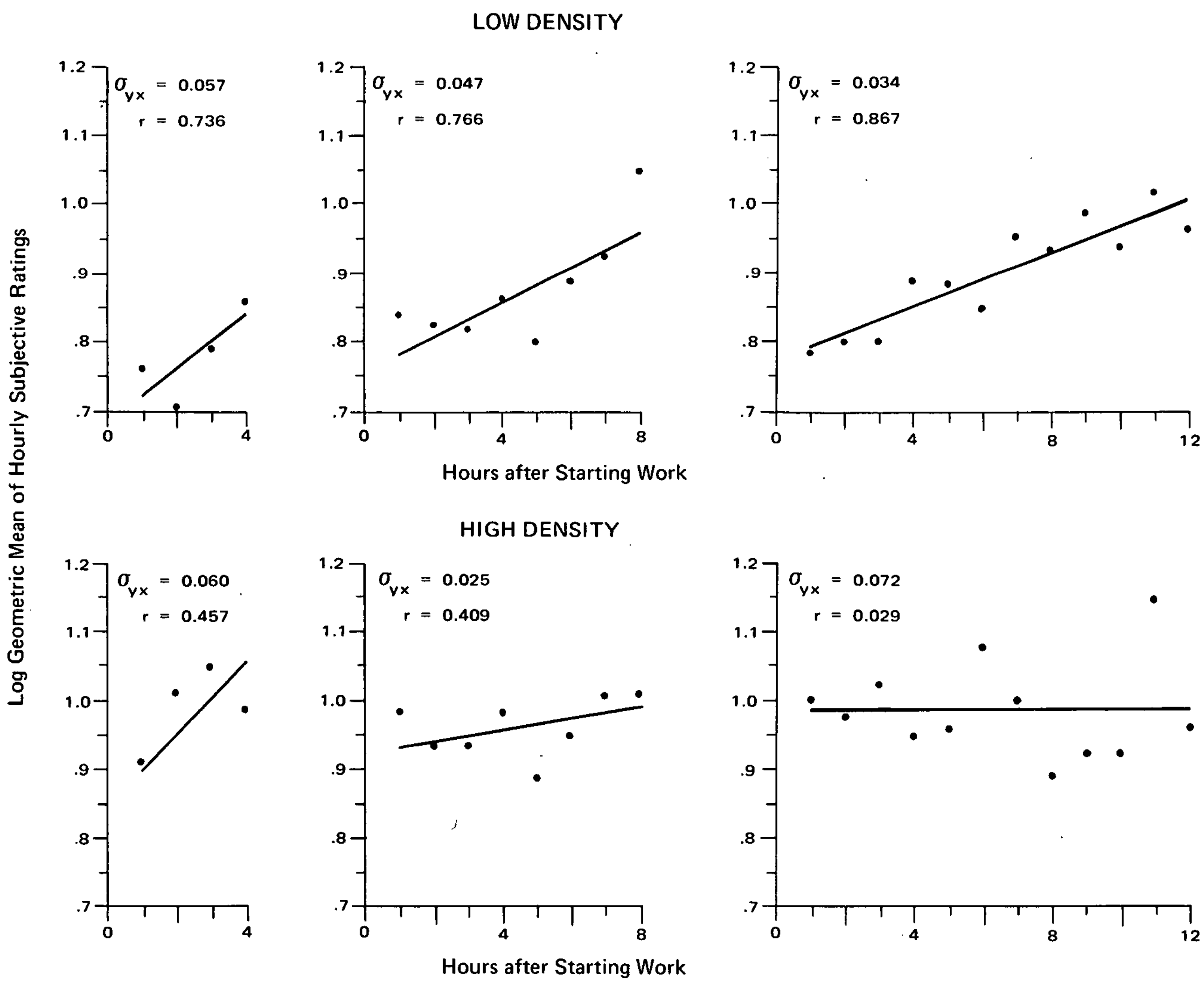


Figure 15. Subjective Estimates of Task Difficulty as a Function of Shift Length

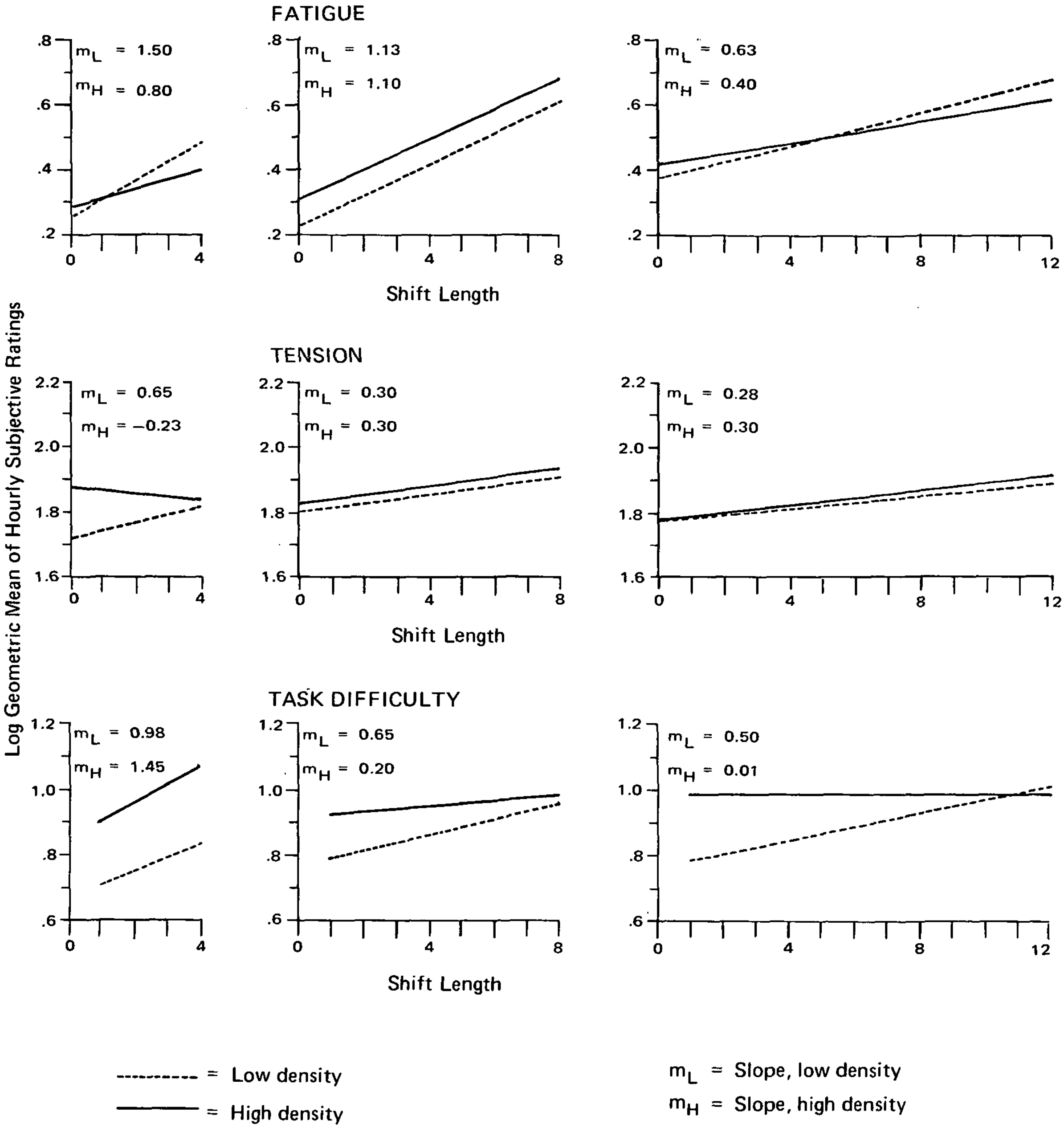


Figure 16. Levels and Gradients of Subjective Workload Estimates

In Figure 16, the plotted lines represent the least-squares fits for each variable under each experimental condition. The ordinate values indicate the level of perceived fatigue, tension, or task difficulty after each hour of work. Since these are all linear functions of the form $y = mx + b$, the slope (m) is the gradient of subjective estimates as a function of time. Thus, the value of m indicates the rate at which estimates of workload increased over time.

It can be seen that in all cases, subjective estimates of the three workload variables were initially lower under low density conditions than under high density conditions. However, the slope of the low density plots was greater (steeper) for task difficulty and fatigue, except for task difficulty in the 4-hour shift. (This exception can probably be attributed to the length of the work period which was so brief that a clear pattern did not have time to develop.) Thus, it would appear that, on the subjective level, communication density had a strong effect on workload, at least in terms of estimates of fatigue and task difficulty. The data on the third subjective variable, tension, were inconclusive.

CHAPTER 5

INTERPRETATION OF RESULTS

Discussion

This study attempted to develop an improved methodology for measurement of information processing workload. The feasibility of using communications recorded in an actual air traffic control operating environment as stimulus material for experimental purposes also was examined. It was hoped that the synthetic task developed here would yield reliable (repeatable) measures in a representative work situation and that performance on this task could be shown to be sensitive to the effects of shift length and communication density.

Task Performance

The data indicated that neither shift length nor communication density had a systematic effect on the ability to encode stimulus material correctly. This was true at both the transmission and transaction levels. From these findings it may be concluded that the task was within the subjects' performance capability even for sustained periods of work (i.e., up to 12 hours) and regardless of the rate at which information had to be processed.

The inherent difficulty in the imposed task apparently was not sufficient to produce a performance decrement even over prolonged periods of work. The subjects may have experienced fatigue (subjective estimates clearly suggest that they did), but they were able to mobilize their energy reserves to sustain their established levels of performance. These findings are consistent with other research (Hartman, 1965) in which Air Force pilots were measured in a 24-hour simulator flight broken into 11 two-hour legs terminated by an Instrument Landing System (ILS) landing. Hartman's results indicated that performance in this type of task can be maintained at initial levels for approximately 20 hours, but that loss in proficiency, when it does occur, can be sudden and precipitous.

It should be noted that in the present study, as opposed to a simulator flight task in which a subject monitors a number of displays, subjects were allowed to concentrate on the primary task to the exclusion of all else. The addition of a secondary loading task would serve to burden the subject additionally and might function as an indicator of the magnitude of the energy expended in maintaining primary task performance. Previous research in workload measurement (Alluisi, 1967) indicates that when subjects are forced to time share among tasks, decrement in the time sharing aspect is one of the earliest features of performance loss. In future studies of workload in an information processing task, use of a secondary task to compensate for the ease of the part-task work situation and to provide a possible means for the measurement of reserve capacity appears warranted.

Another factor which may have served to preclude measurable performance decrements was the work/rest schedule. The ten-minute breaks each hour and the 45-minute rest periods between four-hour sessions probably afforded sufficient time for subjects to recoup whatever energy losses may have occurred as a result of shift length or message density. In retrospect, it is evident that the experimental conditions were less demanding than necessary to produce performance decrement. In future experiments, a more taxing schedule (either with less frequent breaks, longer shift lengths, or both) should be imposed if performance decrement is desired when using a single-variable task such as was the case in this study.

The information processing task used in this study did provide a consistent and repeatable measure which might serve as an index of general performance capability when combined with other tasks to form a representative complex of information processing activities. In the present study it was noted that, after a subject had established a performance level at the end of training, he tended to remain at that level throughout the experimental period. There was a practice effect manifested during the week but the improvement in scores was not large. There also was little fluctuation in the subject's relative performance level. Subjects who scored high during training tended to be the best performers throughout the experiment. Also, those who scored low in training tended to remain at the bottom of the group. The task therefore appears to be sensitive to individual differences in basic performance capacity. It did not, however, demonstrate a comparable sensitivity to the nominal stress effects imposed in this investigation.

Laboratory Performance Measures

The foregoing comments on the task performance measures apply equally to the dependent measures which made up the pre- and postshift test battery. None of the cognitive, perceptual-motor, or sensory measures exhibited systematic patterns of change as a function of the independent variables, even though several have a previously well established sensitivity to fatigue and workload. Again, the lack of positive findings is traceable to factors such as insufficient inherent task difficulty, the absence of a secondary task, high motivation, a nonstrenuous schedule, and the lack of realistic psychological stressors. The results of this experiment appear to be consistent with other studies of stress and fatigue in which it has been often found that individuals can, when being evaluated, mobilize themselves to perform quite well on measured tasks even when they are manifestly greatly fatigued.

Physiological Measures

Heart rate and oral temperature were not measured on a continuous basis in this study, and there was little expectation that hourly readings would provide a sufficiently detailed profile of physiological correlates of workload, as proved to be the case. In future experiments, continuous measurements should be made, but even so, there is little likelihood that physiological variables will exhibit marked effects without the introduction of real psychological stress, substantially beyond that which could reasonably be employed in the laboratory situation.

Subjective Measures

Perhaps the most interesting set of findings were those relating to subjective magnitude estimates of fatigue, tension, and task difficulty. With respect to fatigue, it is clear that estimates obtained at the end of each hour were, in fact, power functions of the elapsed time at work. Correlation coefficients of the individual data points with the least-squares fits are 0.89 or better in all but one case.

The relationship between shift length and subjective estimates of tension is inconsistent, but generally weak. In one case, the correlation coefficient is 0.87, but most are in the vicinity of 0.50, and one condition (4 hours at high density) exhibited a negative correlation. The most probable explanation is that suggested earlier–subjects did not perceive tension to be a result of time at work. This seems plausible since the common conception of tension does not attribute it to duration of work but to the pressures of the activity. The hypothesis in this experiment stemmed largely from the study by Grandjean (1968), who found that air traffic controllers reported greater feelings of tension as a function of shift length. However, it must be remembered that Grandjean's subjects were actually handling air traffic at a major airport, and presumably they were reacting more to the cumulative responsibility (psychological stress) of the job rather than simply to the number of hours spent working. Since subjects in the present experiment had no such responsibility and were not subjected to psychological stress, it is not surprising that increasing feelings of tension were not reported.

The rather mixed reports on task difficulty as a function of shift length probably result from the operation of two extraneous factors. First, while there were some differences in the inherent difficulty of the stimulus material from hour to hour (e.g., number of messages per transmission, clarity of voice rendition, controller and pilot speech mannerisms or accent), these differences were essentially randomly distributed across the work periods. This may have had a confounding effect on any underlying tendency of subjects to perceive a direct relationship between task difficulty and shift length. Second, as observed earlier, the body of experimental evidence indicated that the information processing task was not of sufficient inherent difficulty to produce significant performance decrements. Thus, with a generally "easy" task, it is not surprising that differential ratings of task demands from hour to hour did not show a pronounced trend. Actually, the lack of pattern in task difficulty ratings reflects favorably on the integrity and potential value of the technique. If subjective magnitude estimates of task demands are to be useful indicators of actual task difficulty, they should be independent of fatigue and tension and should not reflect changes over time.

The observed relationship between task difficulty estimates and communication density ran counter to the original hypothesis. It had been assumed that the highest values of m (steepest slopes) on task difficulty, as well as on fatigue and tension, would be associated with high communication density. By and large, the reverse was found to be true. The steeper gradients of the three subjective workload estimates tended to occur in low density sessions. Hence, it would appear that the low density condition represented a situation in which subjects might have been "underworked" and that the waiting time between transactions (which was about equal to the length of the average transaction) acted as an irritant or as a general contributor to perceived workload. This was borne out by unsolicited comments from subjects who stated that the low density sessions were "boring" or "too slow" and that the pace of high density was "about right." From this, it can be concluded that the ability of the individual was miscalculated. It was originally supposed that high density (70% channel utilization) was a near saturation condition. In effect, it appears that the saturation threshold is somewhat higher.

Summary of Conclusions

The following conclusions are drawn from this study:

1. The methodology and synthetic task developed here are promising for future workload research. The feasibility of using communications recorded in an actual operating environment as stimulus material for experimental purposes has been demonstrated. The task performance measure exhibited high intertrial and intersubject stability and could serve as a reliable baseline indicator for future studies of information processing workload.

2. The information processing task was not sensitive to effects of shift length and communication density in the experimental conditions studied here. Subjects apparently were able to call upon energy reserves to maintain consistent levels of performance even during the longest shift length (12 hours). Perceptual-motor, cognitive, sensory, and physiological measures also did not vary systematically as a function of shift length or communication density. The factors which contributed to the lack of a systematic decrement in task performance and in other dependent measures may have been low inherent task difficulty, absence of a secondary loading task, too frequent rest periods, absence of realistic psychological stressors, or a combination of all.

3. Hourly subjective estimates of fatigue increased as a function of shift length. Subjective estimates of task difficulty and fatigue increased more sharply over time under low density than under high density conditions. Subjective estimates of task difficulty were essentially independent of shift length and of the other two workload estimates and appeared to reflect perceived differences in the inherent difficulty of the stimulus material.

Recommendations for Future Research

The findings of this study indicate that future research on information processing activities should incorporate three basic modifications in the experimental design:

1. A secondary loading task involving information processing either aurally, visually, or both should be added.

2. A more demanding experimental routine should be established (e.g., shorter or less frequent rest periods and longer shift length).
3. Higher communication densities should be investigated.

It is felt that these steps would result in a task sufficiently sensitive to expose subtle performance decrements produced by fatigue and workload at varying conditions of shift length and communication density.

The encouraging findings as to the usefulness and validity of subjective magnitude estimates suggest that the technique should be applied in further studies of this type. The technique should also be studied for its applicability as an indicator of psychological state in other work situations.

The differences between findings reported here and those obtained by other investigators who have studied workload in actual operating environments should be explored more thoroughly. The absence of significant change in physiological and perceptual-motor/sensory variables in the present study may well be a result of the lack of true psychological stress in the laboratory situation. Further studies of this factor could have important implications for generalization of findings from simulated to actual work environments.

APPENDIX A

REFERENCES

Alluisi, E. A. Methodology in the use of synthetic tasks to assess complex performance. *Human Factors*. 1967, *9*(4), 375–384.

Chiles, W. D., Alluisi, E. A., & Adams, O. S. Work schedules and performance during confinement. *Human Factors*, 1968, *10*(2), 143–195.

Dahl, H., & Spence, D. Mean heart rate predicted by task demand characteristics. *Psychophysiology*, 1970, *7*, 369–376.

Dougherty, J. D., Trites, D. K., & Dille, J. R. Self-reported stress-related symptoms among air traffic control specialists (ATCS) and non-ATCS personnel. *Aerospace Medicine*, 1965, *36*(10), 956–960.

French, J. W., Ekstrom, R. B., & Price, L. A. Manual for kit of reference tests for cognitive factors. Educational Testing Service, Princeton, New Jersey, 1963.

Grandjean, E. Fatigue: Its physiological and psychological significance. *Ergonomics*, 1968, *11*(5), 427–436.

Hale, H. B., Williams, E. W., Smith, B. N., & Melton, C. E., Jr. Excretion patterns of air traffic controllers. *Aerospace Medicine*, 1971, *42*(2), 127–138.

Hartman, B. Fatigue effects in 24-hour simulated transport flight. USAF School of Aerospace Medicine Report SAM-TR-65-16, Brooks AFB, Texas, April 1965.

Kleitman, N. The sleep-wakefulness cycle of submarine personnel, in panel on psychology and physiology—A survey report on human factors in undersea warfare. National Research Council Committee on Undersea Warfare, Washington, D.C., 1949, 329–341.

Knox, G. W. Investigations of flicker and fusion: I. The effect of practice, under the influence of various attitudes, on the CFF. *Journal of General Psychology*, 1945, *33*, 121–129.

Morgan, B. B., Jr., & Alluisi, E. A. Effects of work-load stress on skilled performance with a code transformation task. Proceedings of 77th Annual Convention of the American Psychological Association, 1969, 753–754.

Reilly, R. E., & Parker, J. F., Jr. Development of an improved perceptual-motor performance measurement system. Prepared by BioTechnology, Inc., under Contract NASW-1329, for National Aeronautics and Space Administration, Washington, D.C., August 1967.

Shoenberger, R. W., & Harris, C. S. Psychophysical assessment of whole body vibration. *Human Factors*, 1971, *13*(1), 41–50.

Stevens, S. S. The psychophysics of sensory function. In W. A. Rosenblith (Ed.) *Sensory Communication*. Boston: MIT Press, 1961.

Stevens, S. S. The surprising simplicity of sensory metrics. *American Psychologist*, 1962, *17*, 29–39.

Stevens, S. S. On the operation known as judgment. *American Scientist*, 1966a, *54*, 385–401.

Stevens, S. S. A metric for social consensus. *Science*, 1966b, *151*, 530–541.

Trumbull, R. Environment modification for human performance. Report ACR-105, Office of Naval Research, Washington, D.C., July 1965. (AD 620 232)

Versace, J. Measurement of ride comfort. SAE Paper 638A, Society of Automotive Engineers, New York, January 1963.

APPENDIX B

INSTRUCTIONS TO SUBJECTS

Purpose of the Experiment

This study has to do with the way people learn to perform an information processing task. We are concerned with such factors as the long-term effects of training in this kind of task, what changes occur in the learning process over a particular time span, and other aspects of the learning pattern. We are also interested in the implications that these learning patterns have for training people in similar kinds of tasks.

The material to which you will be listening throughout the course of the study consists of messages that pass between air traffic controllers and airplane pilots. These messages vary with respect to their content, and their intelligibility. They are actual messages recorded at different airports in the eastern corridor, and you will be trained to classify them according to content.

In point of fact, air traffic controller trainees typically sit at consoles within a terminal tower or center and listen to large numbers of interactions between pilots and other controllers before they are allowed to control actual flights.

The terms of the study require that twelve male volunteers, in two groups of six each, undergo a six-day period of testing. In preparation for this testing, volunteers will also be required to complete a three-day period of training in the tasks which they will be expected to perform during the experimental test period.

The purpose of this experiment, described in detail above, is to study performance in an information processing task. The conditions of this experiment do not require the taking of dangerous drugs or intoxicants, the loss or deprivation of sleep, exposure to hazardous or potentially injurious circumstances, or performance of physically arduous tasks.

Perceptual-Motor Tests

Visual Reaction Time

Now we are going to measure your reaction time. Find the switch labeled RT at the back and just to the right of center on the lower portion of your console. This switch will light up as your signal to respond. Place your fingers *very lightly* on the switch. When the light appears, press the switch as rapidly as possible. There will be four trials with about ten seconds between signals.

Auditory Reaction Time

Now we will measure your auditory reaction time. The signal will be a tone which is produced by the small speaker located just beneath your left-hand meter. You will use the same response switch. There will be four trials with about ten seconds between signals. When you hear the tone, press your switch as rapidly as possible.

Response Orientation

This test measures your ability to make a directional control movement in response to a nondirectional signal. Your control is the black lever switch located in back of your right-hand control stick. Notice that it may be moved in four directions: left, right, forward, and back, and it returns to center when released. Just to the left of the lever switch is an unlabeled display module. During the test, this module will present a series of colored lights: green, red, white, and blue, in random order. Each color corresponds to a position on the switch. When a light appears, you are to move the switch as quickly as you can to the appropriate position and extinguish the light. The relationships are as follows

Green	=	Back (toward subject)
White	=	Forward
Red	=	Left
Blue	=	Right

When each light comes on, the clock starts running. When you make the correct response and extinguish the light, you also stop the clock. Your score is the total amount of time accumulated over the entire sequence. If your first response to a light is not correct, simply continue responding until you extinguish the light. Once the test has begun, do not assume it is finished for any reason until you are told to stop.

Arm-Hand Steadiness

This test measures the amount of tremor in your arm and hand while held *fully extended without locking your elbow*. Plug the stylus into the blue terminal on the CRT. Hold the stylus as you would hold a pencil. Now extend your arm and insert the tip of the stylus into the aperture directly above the CRT. Do not lock your elbow. Do not jam the collar of the stylus against the rim of the aperture. Your task is to hold the tip of the stylus inside of the aperture without contacting the rim. There will be three 10-second trials with about five seconds break between trials. When the amber warning light at the top of your console appears, insert the stylus. Within a few seconds the green light will appear indicating that you are being scored. *Continue holding the stylus until the green light goes out*, then rest your arm on the console until the next amber warning light appears.

Perceptual Speed

This is a test of your ability to make rapid comparisons between two displays. The displays are the meters located at the top of your console. Notice that directly in front of you there are two switches–the switch on the left is labeled "S;" the switch on the right is labeled "D." Your task will be to compare the indications on your meters to determine whether they are the same or different. If both meters show the same value, press the switch labeled "S;" if the meter indications are different, press the switch labeled "D." If you make a correct response, that is, if you press "S" when the meter indications are actually the same or if you press "D" when the indications are, in fact, different, the next pair of meter indications will appear. If you make an incorrect response, the meters will not change and an error will be recorded on the counter. As soon as you realize that you have made an incorrect response, immediately press the other switch and continue with the test. Both speed and accuracy are important. Your score is the amount of time that you take to process the entire sequence of values and the number of errors made.

Time Sharing

This test measures how well you can divide your attention between two displays to detect the occurrence of certain events. In this case, you will be required to monitor the two meters at the top of your console in order to detect the movement of either of the pointers. Directly in front of you are two switches labeled T–S: the switch on the left corresponds to the meter on the left; the switch on the right corresponds to the meter on the right.

When the test is started, you are to scan back and forth visually between the two meters. As soon as you notice that a pointer has begun to move, press the appropriate switch as quickly as you can. Whenever a pointer begins to move, the clock starts; when you make the correct response, you will stop the pointer and the clock. There is no relationship between any value shown on the meter and which meter might begin to move next. Whenever one or both meters reads beyond 30, press your RESET switch. This is simply a precaution to keep the meters from running off scale in the event that you do not detect movement of the pointer soon enough.

Cognitive Tests

Number Facility

This is a test to see how quickly and accurately you can add, subtract, multiply, and divide. It is not expected that you will finish all the problems in the time allowed.

Write your answers in the space provided. Your score on this test will be the number of problems that are done correctly. Work as rapidly as you can without sacrificing accuracy.

You will have two minutes for the test.

Number Comparison

This is a test to find out how quickly you can compare two numbers and decide whether or not they are the same. If the numbers are the same, go on to the next pair, making no mark on the page. If the numbers are *not* the same, put an X on the line between them.

Your score will be the number marked correctly minus the number marked incorrectly. Therefore, it will not be to your advantage to guess unless you have some idea whether or not the numbers are the same.

You will have 1½ minutes for this test.

APPENDIX C

ANALYSES OF VARIANCE

Tables VIII–XVII on the following pages are analyses of variance for the primary information processing task and each of the elements of the pre- and postshift test battery. The key to abbreviations used in column headings is:

SS – Sum of Squares

df – Degrees of Freedom

MS – Mean Squared

et – Error Term

F – F-Test

TABLE VIII

Analysis of Variance – Information Processing Task[1]

Source		SS	df	MS	et	$F^{2,3}$
Total		8433.15	143			
Between Ss		5109.35	11			
Within Ss		3323.80	132			
Density (D)		25.25	1	25.25	(1)	2.68
Shift Length (L)		15.30	2	7.65	(2)	0.34
Position (P)		12.43	1	12.43	(3)	1.08
D X L		22.44	2	11.22	(4)	0.90
D X P		0.41	1	0.41	(5)	0.29
L X P		15.13	2	7.56	(6)	0.01
D X L X P		1.35	2	0.67	(7)	0.02
Error Terms						
Ss/D	(1)	103.73	11	9.43		
Ss/L	(2)	501.31	22	22.78		
Ss/P	(3)	126.16	11	11.46		
Ss/D X L	(4)	274.72	22	12.48		
Ss/D X P	(5)	844.36	11	76.76		
Ss/L X P	(6)	577.77	22	26.26		
Ss/D X L X P	(7)	804.79	22	36.58		

1. Scores corrected for practice effect.
2. p .05 = 4.84, p .01 = 9.65 (df 1, 11)
3. p .05 = 3.44, p .01 = 5.72 (df 2, 22)

TABLE IX

Analysis of Variance – Visual Reaction Time

Source		SS	df	MS	et	$F^{1,2}$
Total		0.15866	143			
Between Ss		0.05186	11			
Within Ss		0.10680	132			
Density (D)		0.00000	1	0.00000	(1)	0
Shift Length (L)		0.00083	2	0.00042	(2)	0.28767
Position (P)		0.00083	1	0.00083	(3)	4.88235[1]
D X L		0.00302	2	0.00151	(4)	1.91139
D X P		0.00179	1	0.00179	(5)	4.58974
L X P		0.00029	2	0.00015	(6)	0.21429
D X L X P		0.00007	2	0.00004	(7)	0.03846
Error Terms						
Ss/D	(1)	0.00601	11	0.00055		
Ss/L	(2)	0.03220	22	0.00146		
Ss/P	(3)	0.00187	11	0.00017		
Ss/D X L	(4)	0.01739	22	0.00079		
Ss/D X P	(5)	0.00426	11	0.00039		
Ss/L X P	(6)	0.01539	22	0.00070		
Ss/D X L X P	(7)	0.02285	22	0.00104		

1. p .05 = 4.84, p .01 = 9.65 (df 1, 11)
2. p .05 = 3.44, p .01 = 5.72 (df 2, 22)

TABLE X

Analysis of Variance – Auditory Reaction Time

Source		SS	df	MS	et	F[1,2]
Total		0.20804	143			
Between Ss		0.07410	11			
Within Ss		0.13394	132			
Density (D)		0.00103	1	0.00103	(1)	1.58462
Shift Length (L)		0.00077	2	0.00039	(2)	0.31452
Position (P)		0.00082	1	0.00082	(3)	1.07895
D X L		0.00029	2	0.00015	(4)	0.13393
D X P		0.00130	1	0.00130	(5)	1.19266
L X P		0.00504	2	0.00252	(6)	2.68085
D X L X P		0.00090	2	0.00045	(7)	0.41667
Error Terms						
Ss/D	(1)	0.00710	11	0.00065		
Ss/L	(2)	0.02735	22	0.00124		
Ss/P	(3)	0.00832	11	0.00076		
Ss/D X L	(4)	0.02459	22	0.00112		
Ss/D X P	(5)	0.01204	11	0.00109		
Ss/L X P	(6)	0.02072	22	0.00094		
Ss/D X L X P	(7)	0.02367	22	0.00108		

1. p .05 = 4.84, p .01 = 9.65 (df 1, 11)
2. p .05 = 3.44, p .01 = 5.72 (df 2, 22)

TABLE XI

Analysis of Variance – Response Orientation

Source		SS	df	MS	et	$F^{1,2}$
Total		834.69	143			
Between Ss		479.17	11			
Within Ss		355.52	132			
Density (D)		7.19	1	7.19	(1)	8.66[1]
Shift Length (L)		9.28	2	4.64	(2)	0.62
Position (P)		1.41	1	1.41	(3)	0.66
D X L		2.00	2	1.00	(4)	0.65
D X P		0.41	1	0.41	(5)	0.17
L X P		2.73	2	1.36	(6)	1.92
D X L X P		10.42	2	5.21	(7)	2.28
Error Terms						
Ss/D	(1)	9.17	11	0.83		
Ss/L	(2)	163.52	22	7.43		
SS/P	(3)	23.59	11	2.14		
Ss/D X L	(4)	33.83	22	1.54		
Ss/D X P	(5)	25.94	11	2.36		
Ss/L X P	(6)	15.54	22	0.71		
Ss/D X L X P	(7)	50.49	22	2.29		

1. p .05 = 4.84, p .01 = 9.65 (df 1, 11)
2. p .05 = 3.44, p .01 = 5.72 (df 2, 22)

TABLE XII

Analysis of Variance – Arm-Hand Steadiness

Source		SS	df	MS	et	F[1,2]
Total		26,463.38	143			
Between Ss		16,503.08	11			
Within Ss		9,960.30	132			
Density (D)		0.01	1	0.01	(1)	0.000139
Shift Length (L)		27.18	2	13.59	(2)	0.15
Position (P)		540.56	1	540.56	(3)	4.71
D X L		138.51	2	69.26	(4)	0.99
D X P		11.68	1	11.68	(5)	0.10
L X P		65.05	2	32.53	(6)	0.72
D X L X P		18.42	2	9.21	(7)	0.16
Error Terms						
Ss/D	(1)	790.30	11	71.85		
Ss/L	(2)	2,032.38	22	92.38		
Ss/P	(3)	1,263.02	11	114.82		
Ss/D X L	(4)	1,531.42	22	69.61		
Ss/D X P	(5)	1,316.91	11	119.72		
Ss/L X P	(6)	996.62	22	45.30		
Ss/D X L X P	(7)	1,228.24	22	55.83		

1. p .05 = 4.84, p .01 = 9.65 (df 1, 11)
2. p .05 = 3.44, p .01 = 5.72 (df 2, 22)

TABLE XIII

Analysis of Variance – Perceptual Speed

Source		SS	df	MS	et	$F^{1,2}$
Total		4405.05	143			
Between Ss		2978.46	11			
Within Ss		1426.59	132			
Density (D)		49.53	1	49.53	(1)	4.59
Shift Length (L)		27.77	2	13.89	(2)	0.83
Position (P)		82.37	1	82.37	(3)	9.89[1]
D X L		4.37	2	2.19	(4)	0.31
D X P		42.61	1	42.61	(5)	8.54[1]
L X P		80.39	2	40.20	(6)	5.67[2]
D X L X P		6.80	2	3.40	(7)	0.39
Error Terms						
Ss/D	(1)	118.62	11	10.78		
Ss/L	(2)	366.86	22	16.68		
Ss/P	(3)	91.67	11	8.33		
Ss/D X L	(4)	154.46	22	7.02		
Ss/D X P	(5)	54.87	11	4.99		
Ss/L X P	(6)	156.07	22	7.09		
Ss/D X L X P	(7)	190.20	22	8.65		

1. p .05 = 4.84, p .01 = 9.65 (df 1, 11)
2. p .05 = 3.44, p .01 = 5.72 (df 2, 22)

TABLE XIV

Analysis of Variance – Time Sharing

Source		SS	df	MS	et	$F^{1,2}$
Total		1302.58	143			
Between Ss		419.35	11			
Within Ss		883.23	132			
Density (D)		0.92	1	0.92	(1)	0.18
Shift Length (L)		31.56	2	15.78	(2)	1.66
Position (P)		10.78	1	10.78	(3)	3.75
D X L		1.99	2	0.99	(4)	0.09
D X P		2.05	1	2.05	(5)	0.71
L X P		16.28	2	8.14	(6)	1.82
D X L X P		3.44	2	1.72	(7)	0.24
Error Terms						
Ss/D	(1)	56.59	11	5.14		
Ss/L	(2)	208.74	22	9.49		
Ss/P	(3)	31.59	11	2.87		
Ss/D X L	(4)	233.27	22	10.60		
Ss/D X P	(5)	31.41	11	2.85		
Ss/L X P	(6)	98.40	22	4.47		
Ss/D X L X P	(7)	156.21	22	7.10		

1. p .05 = 4.84, p . 01 = 9.65 (df 1, 11)
2. p .05 = 3.44, p .01 = 5.72 (df 2, 22)

TABLE XV

Analysis of Variance – Number Facility

Source		SS	df	MS	et	F[1,2]
Total		8146.45	143			
Between Ss		6931.97	11			
Within Ss		1214.48	132			
Density (D)		44.44	1	44.44	(1)	2.73
Shift Length (L)		13.35	2	6.67	(2)	0.80
Position (P)		9.00	1	9.00	(3)	2.05
D X L		32.77	2	16.38	(4)	1.62
D X P		4.70	1	4.70	(5)	0.24
L X P		11.62	2	5.81	(6)	1.89
D X·L X P		9.43	2	4.71	(7)	0.59
Error Terms						
Ss/D	(1)	178.79	11	16.25		
Ss/L	(2)	182.46	22	8.29		
Ss/P	(3)	48.34	11	4.39		
Ss/D X L	(4)	222.67	22	10.12		
Ss/D X P	(5)	216.04	11	19.64		
Ss/L X P	(6)	67.63	22	3.07		
Ss/D X L X P	(7)	173.24	22	7.87		

1. p .05 = 4.84, p .01 = 9.65 (df 1, 11)
2. p .05 = 3.44, p .01 = 5.72 (df 2, 22)

TABLE XVI

Analysis of Variance – Number Comparison

Source		SS	df	MS	et	F[1,2]
Total		4525.21	143			
Between Ss		3090.81	11			
Within Ss		1434.40	132			
Density (D)		11.11	1	11.11	(1)	1.28
Shift Length (L)		26.06	2	13.03	(2)	0.80
Position (P)		36.00	1	36.00	(3)	2.71
D X L		16.05	2	8.02	(4)	1.07
D X P		0.03	1	0.03	(5)	0.00
L X P		58.50	2	29.25	(6)	3.35
D X L X P		22.89	2	11.44	(7)	1.22
Error Terms						
Ss/D	(1)	95.36	11	8.67		
Ss/L	(2)	358.06	22	16.28		
Ss/P	(3)	145.83	11	13.26		
Ss/D X L	(4)	163.76	22	7.44		
Ss/D X P	(5)	103.48	11	9.41		
Ss/L X P	(6)	191.67	22	8.71		
Ss/D X L X P	(7)	205.60	22	9.35		

1. p .05 = 4.84, p .01 = 9.65 (df 1, 11)
2. p .05 = 3.44, p .01 = 5.72 (df 2, 22)

TABLE XVII

Analysis of Variance – Critical Fusion Frequency

Source		SS	df	MS	et	F[1,2]
Total		753.47	143			
Between Ss		508.12	11			
Within Ss		245.35	132			
Density (D)		0.60	1	0.60	(1)	0.27
Shift Length (L)		9.40	2	4.70	(2)	2.47
Position (P)		2.25	1	2.25	(3)	0.58
D X L		3.07	2	1.53	(4)	0.83
D X P		7.79	1	7.79	(5)	8.20[1]
L X P		6.80	2	3.40	(6)	2.72
D X L X P		1.80	2	0.90	(7)	0.73
Error Terms						
Ss/D	(1)	24.27	11	2.21		
Ss/L	(2)	41.71	22	1.90		
Ss/P	(3)	42.56	11	3.87		
Ss/D X L	(4)	40.49	22	1.84		
Ss/D X P	(5)	10.40	11	0.95		
Ss/L X P	(6)	27.39	22	1.25		
Ss/D X L X P	(7)	26.82	22	1.22		

1. p .05 = 4.84, p .01 = 9.65 (df 1, 11)
2. p .05 = 3.44, p .01 = 5.72 (df 2, 22)

NATIONAL AERONAUTICS AND SPACE ADMISTRATION
WASHINGTON, D.C. 20546

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE $300

FIRST CLASS MAIL

POSTAGE AND FEES PAID
NATIONAL AERONAUTICS AND
SPACE ADMINISTRATION

U.S.MAIL

NASA 451

POSTMASTER: If Undeliverable (Section 158 Postal Manual) Do Not Return

---

"The aeronautical and space activities of the United States shall be conducted so as to contribute . . . to the expansion of human knowledge of phenomena in the atmosphere and space. The Administration shall provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof."

—NATIONAL AERONAUTICS AND SPACE ACT OF 1958

NASA SCIENTIFIC AND TECHNICAL PUBLICATIONS

TECHNICAL REPORTS: Scientific and technical information considered important, complete, and a lasting contribution to existing knowledge.

TECHNICAL NOTES: Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

TECHNICAL MEMORANDUMS: Information receiving limited distribution because of preliminary data, security classification, or other reasons.

CONTRACTOR REPORTS: Scientific and technical information generated under a NASA contract or grant and considered an important contribution to existing knowledge.

TECHNICAL TRANSLATIONS: Information published in a foreign language considered to merit NASA distribution in English.

SPECIAL PUBLICATIONS: Information derived from or of value to NASA activities. Publications include conference proceedings, monographs, data compilations, handbooks, sourcebooks, and special bibliographies.

TECHNOLOGY UTILIZATION PUBLICATIONS: Information on technology used by NASA that may be of particular interest in commercial and other non-aerospace applications. Publications include Tech Briefs, Technology Utilization Reports and Technology Surveys.

Details on the availability of these publications may be obtained from:

SCIENTIFIC AND TECHNICAL INFORMATION OFFICE

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Washington, D.C. 20546