DOT/F AA/EM-81/13
DOT/FAA/CT-82/66

The Relationship Between Effort Rating and Performance in a Critical Tracking Task

Bruce Rosenberg
Jacqueline Rehmann
Earl Stein

October 1982

Final Report

This document is available to the U.S. public through the National Technical Information Service, Springfield, Virginia 22161.

US Department of Transportation
Federal Aviation Administration
Office of Systems Engineering Management
Washington, D.C. 20590

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof.

The United States Government does not endorse products or manufacturers. Trade or manufacturer's names appear herein solely because they are considered essential to the object of this report.

PREFACE

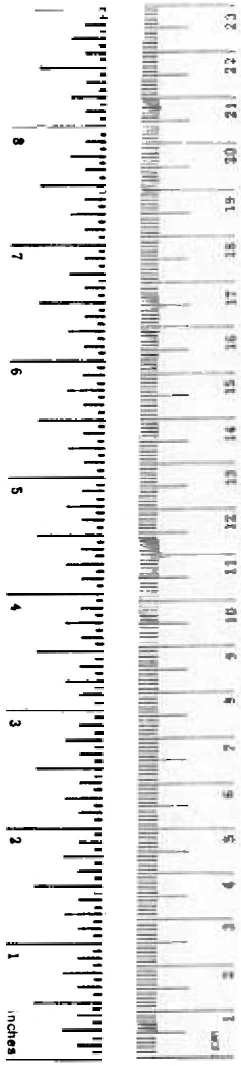
The authors gratefully acknowledge the technial assistance of Douglas Elliott, of the Systems Simulation Branch at the Technical Center, in setting up the analog computer for the experiment described herein. Mr. Elliott also contributed the section found in appendix C of this report entitled "Tracking Task" which describes and illustrates the critical tracking task.

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

Symbol	When You Know	Multiply by	To Find	ymbol
		LENGTH		
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
		AREA		
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
		MASS (weight)		
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
		VOLUME		
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	
pt	pints	0.47	liters	
qt	quarts	0.95	liters	
gal	gallons	3.8	liters	
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
		TEMPERATURE (exact)		
	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	

*1 in = 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price $2.25, SD Catalog No. C13.10:286.



Approximate Conversions from Metric Measures

Syn	When You Know	Multiply by	To Find	
		LENGTH		
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
		AREA		
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
		MASS (weight)		
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
		VOLUME		
	milliliters	0.03	fluid ounces	fl oz
	liters	2.1	pints	pt
	liters	1.06	quarts	qt
	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
		TEMPERATURE (exact)		
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	

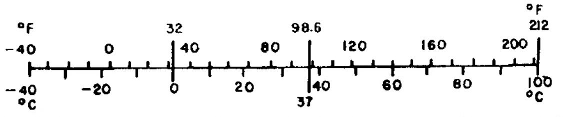


TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

EXECUTIVE SUMMARY

This report documents the results of a preliminary evaluation of a Pilot Objective/Subjective Workload Assessment Technique (POSWAT). The study employed a critical tracking task, in which 24 subjects (pilots and nonpilots) viewed an analog display of the error between operator input and system output, while correcting with opposite pressure on a joystick. The purpose was to determine if there was a relationship between participant responses on a 10-point scale administered during task performance and tracking task difficulty. Eight measures were used in the data analysis and results were verified statistically. The eight measures were critical lambda (degree of system instability), operating lambda, effort rating (a subjective measure), rating response delay, mean tracking error, mean log tracking error, mean stick deflection, and mean log stick deflection. Following a brief review of the workload literature, the experimental methodology is described. The data analysis section includes the questionnaire used.

It is generally concluded that POSWAT used for measuring effort rating and rating delay on a regular basis during this experiment is minimally intrusive, is informative, and merits further evaluation in a cockpit environment. More specifically:

1. Subjects were able to discriminate levels of effort involved in controlling a critical tracking task at four distinct difficulty levels using the POSWAT technique.

2. Nonpilot subjects obtained significantly lower critical lambda values (divergence rates) and reported significantly higher effort than pilot subjects.

3. Response delays did not vary as a function of difficulty level.

4. Subjects were unable to identify difficulty level presentation order in the post-test debriefing session. This contrasts with their discrimination obtained from the minute-by-minute effort rating using the 10-level keyboard.

5. Effort rating varied as a function of the log of stick deflection or tracking error more closely than with difficulty level as defined by proportion of critical lambda.

INTRODUCTION

PURPOSE.

The purpose of this report is to document the results of a preliminary evaluation of a Pilot Objective/Subjective Workload Assessment Technique (POSWAT). The technique is intended for use in evaluating the potential impact associated with changes in cockpit procedure and instrumentation, such as those resulting from the introduction of a Cockpit Display of Traffic Information (CDTI). The technique would serve as an appropriate workload measurement method that would provide a common basis for assessing the results of many individual experiments. This study employed a two-axis, compensatory critical tracking task, in which 24 subjects viewed an analog display of the error between operator input and system output, while correcting with opposite pressure on a joystick. The purpose is to determine if there is a relationship between participant responses on a subjective 10-point scale administered during task performance and objectively predetermined tracking task difficulty. If participant responses reliably change as a function of task difficulty, then the workload assessment tool has application in future simulation and operations in which pilot workload is measured. The research was conducted at the Federal Aviation Administration (FAA) Technical Center, Atlantic City, New Jersey, as part of a joint NASA/FAA program. The issue that the current research wishes to address concerns the effect of CDTI on pilot workload. Pilot workload imposed by equipment design and/or operational procedures is a major concern. To date, it has not been possible to develop stable measures which are useful and can reliably predict workload in varying flight situations. Further, the growing number of system errors, the anticipated growth of traffic, the necessary increases in automating the current air traffic control (ATC) system, probable changes in the traditional roles of the controller and pilot, and the evolution to the more flight efficient aircraft designs, make a comprehensive workload research program imperative (Albrecht, 1981). To this end, a series of general aviation simulation and operational flight studies will be carried out at the FAA Technical Center to evaluate the CDTI concept and its effect on the level of pilot workload. However, before these studies can be accomplished, measurement methods must be established and pretested to confirm both empirical and face validity.

BACKGROUND.

Since the advent of a scientific concern for man-machine relationships, investigators have been trying to evaluate workload as an indicator of how well equipment design interfaces with the needs and limitations of human operators. Prior to undertaking the current research, a comprehensive review of the workload literature was completed (Rehmann, 1982). Results indicate a relative consensus among investigators that measurement of workload is no simple affair. At best, workload is viewed as a multidimensional construct (Eggemeir, 1980; Chiles, 1979). In the realm of such complexity, it is unlikely that any simple technique will suffice to account for all the variance (Williges and Wierville, 1979). Given the wide variety of contexts in which attempts have been made to specify the nature of workload (i.e., personnel selection, job selection, man-machine design in industry, laboratory research), there has been only marginal success in measuring, specifying, and predicting workload (Chiles, 1979). It is, therefore, not surprising that similar problems exist in aviation.

Attempts to define workload have been as diverse as the measurement techniques employed. Various types of pilot workload have been defined including mental, perceptual, physical, and emotional. Goerres (1977) uses the term "psychophysical workload" to encompass all the load factors on the pilot, and his reaction to them. He states, "psychophysical workload...comprises the effects of the grand total workload on the human organism, human behavior, and subjective feeling." Further, workload depends on the duration and intensity of the activity, intra-individual factors in the subject, such as an individual's present state of health, and job-related knowledge and skill.

Katz (1980) views the concept of pilot workload using the following formula:

Total Workload = Physical Workload + Mental Workload

He says that although physical loads cannot be ignored in research, mental workload has become complex to the point where an understanding of it is crucial to understanding pilot workload. Physical workload is readily quantifiable, whereas mental workload has been described as an "intervening" variable, and is not directly observable (Sheridan and Simpson, 1979). In their research, Sheridan and Simpson refer to mental workload in terms of a "sense of mental effort," or how hard one feels one is working. One person may indicate a feeling of great mental effort, while another individual may claim to be exerting almost no mental effort, while both perform equally. For this reason, the researchers feel that mental workload is not performance per se, and it is not task demand, but rather a term that implies a combination of mental effort, information processing, and emotion in response to task demands.

From the discussion, it becomes apparent that no generally accepted definition of workload exists, and each investigator is tasked to develop his/her own model of construct which best fits the situation (Rehmann, 1981; Chiles 1979). In a general sense, workload is viewed as a combination of input to the operator, information processing, task demand, and operator performance. One is then faced with the task of accurate measurement techniques that will measure one or all of the workload components listed.

The general class of behavioral measures is discussed and includes subjective measures, spare mental capacity, and primary task measures. Results of workload studies using these methods have shown some favorable results.

BEHAVIORAL MEASURES.

SUBJECTIVE RESPONSES. The use of subjective responses made by participants is a common method of assessing workload, and includes psychometrically defined rating scales, structured questionnaires, open-ended questionnaires, and structured and unstructured interviews. Surprisingly, research on the results of subjective measures indicates that they are often the most sensitive and provide meaningful data to the investigator.

This is attributed to pilot acceptance, which is generally favorable, and also to the fact that opinion ratings are not intrusive and can be administered following laboratory or field testing. No special provisions of physical space, portability, data transmission, or integration into the aircraft system are required. In most cases, the subjective rating is used with other measures of workload for greater reliability. Perhaps, the best known subjective measure in aviation is the Cooper-Harper scale. This scale was developed to assess aircraft handling qualities, and it has been modified to focus on pilot workload rather than on the aircraft itself (Sheridan and Simpson, 1979). Katz (1980) applied modified scales to workload measurement in a simulated instrument approach landing and found high reliability to be a major benefit in the subjective rating scale. Test participants were asked to view a video replay of their flights and reassess their workload using his scale. The reassessments were markedly similar to the original rating, and in most cases, the original rating remained unchanged. Additionally, participants in the Katz study were asked whether they felt that it was possible to judge or perceive their own workload, and all responded affirmatively.

Workload research based on subjective measures does have some weaknesses, however. Most subjective workload evaluations have been performed after a flight as part of a debriefing session. This post-hoc approach suffers some deficiencies; i.e., more recent or typical events tend to have greater impact on judgment: the judgments tend to be a time average of the entire run, and information on minimum and maximum workload during a run is often lost (Rosenberg, 1981). Since subjective measures remain the most widely accepted workload measurement to date, what is needed is a minimally intrusive data collection technique which avoids deficiencies inherent in the former approaches. This technique involves recording subjective workload estimates and response delays at equal intervals during task execution. The workload measurement technique described in this paper was developed in response to this need.

SPARE MENTAL CAPACITY. Another workload measurement technique that falls into the general category of behavioral measures of mental workload is the evaluation of the concept of spare mental capacity (Williges and Wierwille, 1979). This concept is based on the assumption of a limited channel capacity sampling model of the human operator. This theory assumes that an upper bound exists on the operator's ability to gather and process information. Spare mental capacity is the difference between the total workload capacity of the operator and the capacity needed to perform the task:

Total Workload Capacity - capacity needed to perform task = spare mental capacity

Williges and Wierwille describe three general methodological approaches for the measurement of workload using the spare mental capacity hypothesis. They are task analytic, secondary task, and occlusion procedures.

Task analytic methods rely heavily on mathematical/theoretical methods from the field of system engineering, and data are usually obtained through laboratory and simulation tests rather than through actual flight evaluation. The underlying assumption is that all task components are performed serially and require specific lengths of time to complete. If the actual time available for completion exceeds the sum of theoretical time for performing task components, the inference is that spare mental capacity exists. Stress and task queuing occur when time is insufficient to perform the tasks.

The secondary task procedures provide an additional task for the human operator to perform when the main (or primary) task has been satisfied. Secondary task performance becomes an indirect measure of operator workload based on the theory that performance of the additional task decreases as the attentional demand of the primary task increases.

Occlusion is similar to the secondary task technique in that it is a time-sharing technique, and it can be used in cases where primary informational inputs are visual. The procedure for using occlusion includes suppressing visual information inputs. For example, the operator may wear a helmet, or hat, fitted with an opaque visor which can be closed by external control, or the electronic displays can be blanked out to accomplish blocking. Results of driving tests where the occlusion method was used revealed that the less frequent the observations, the slower the driver's speed. The faster the speed, the more numerous the driver's observations, as would be expected. Studies that used visual interruption to assess driver's sensitivity to degraded conditions found that this method was sensitive to task difficulty and operator skill (Williges and Wierwille, 1979).

The major underlying hypothesis for the primary task performance assumes that as the mental workload of a human operator increases, the performance of that operator may change, usually in the direction of degradation. Such a change is assumed to be an indication of increased workload. A secondary hypothesis suggests that successful completion of a mission is a measure of workload in itself. If a mission cannot be completed successfully, then one can infer that the operator is overloaded. Workload studies using primary task measures are divided into three major categories: single measures, multiple measures, and mathematical modeling. The greatest applicability of the primary task measures, either single or multiple, is in a high workload situation, as revealed by various research findings. In a low workload situation, primary task measures have not been demonstrated to be useful due to the fact that the operator adapts to maintain output at an acceptable level.

MATHEMATICAL MODELING. Mathematical modeling studies with workload implications are fairly recent, although mathematical modeling using dynamic or mathematical equations of human operator performance in systems have a longer history. Several studies have been examined describing functions and similar models in manual control systems. A describing function refers to the mathematical representation of the behavior of the human operator in a feedback control system. The study cited by Jex, McDonnell, and Phatak (1966), describes the results of a critical tracking task in conjunction with a describing function model to assess workload. This served as the basis for the tracking task used in this study.

Just as there is no one universal definition of workload, there is also a multiplicity of approaches used in the measurement of workload. Assuming that workload is multidimensional, its measurement will have to reflect this complexity. The current research described in this report accepts the multidimensional concept of workload, encompassing the overtly physical elements of input demands and operator behavior. These are directly observable. It also includes the intellectual events which have been classified under such headings as information processing, planning, problem solving, and decisionmaking. These can only be inferred based on what the individual says and does. For the purposes of the current project, it is assumed that if you ask someone how hard they are working, the response will reflect both the physical and nonphysical demands of their task.

METHODOLOGY

CRITICAL TRACKING TASK.

The study employed a nonflying computer driven task, called a critical tracking task, in which difficulty level was clearly definable and controllable. The purpose was to determine if there is a relationship between participant responses on a 10-point workload scale administered during task performance and the task difficulty level which was predetermined. It was hypothesized that the relationship would at least be ordinal and demonstrate some consistency across participants.

The critical tracking task requires that the subject keep a point of light (pip) centered on a screen. The pip diverges from the center if no control is used. The tracking task is analogous to balancing a broomstick on the tip of one's finger with the stick slowly becoming shorter. The shorter the stick, the faster it tends to fall, and the more difficult the task becomes. The length of the stick at the time it falls or, in control theory terms, the divergence rate at the time at which closed loop control is lost (critical lamda in radians per second) is the performance limit of the subject. The performance limit has been shown to reliably change as a function of such factors as blood alcohol level, drug use, fatigue from long term truck driving, etc. This critical tracking task is not unlike the task of flying an aircraft and can be compared with instrument approaches to a localizer. Higher levels of tracking task difficulty correspond to an Instrument Landing System (ILS) task near touchdown.

While tracking the pip, the subject is asked to evaluate his current level of workload once every minute. In a single action, the subject provides both (1) a subjective estimate of his workload during the immediately preceding minute, and (2) objective measures in the form of latency in responding to the workload query stimulus and in the form of missed responses.

There are several advantages in using this subjective workload measurement technique in conjunction with the tracking task. The tracking task has been shown through previous research to be a highly motivating task because the displayed "error" target quickly drifts off center and requires the subjects' continuous attention and effort using the control in order to compensate for the drift and to keep the pip centered (Jex, McDonnell, Phatak, 1966). At the same time, however, it is a task whose difficulty can be precisely controlled, and could provide a structured, single-task environment in which to validate the 10-point subjective measurement scale. Subjects appear to understand fully both the task and the relationship between the workload rating scale and task performance. The scale is simple, easy to understand, and anchored at 1 (very easy) and 10 (very hard).

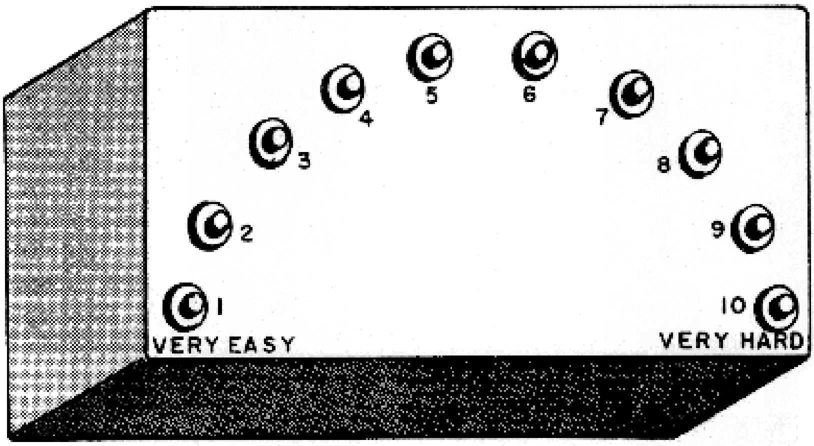
Once it is determined whether or not the workload measurement technique is effective in measuring a subject's assessment of how hard he/she is working at a given time, the scale can be used in simulation studies and actual flight tests.

TEST SUBJECTS.

Two major groups of participants were involved in the study. The first group was comprised of 12 nonpilots. This group included 9 males and 3 females, with a mean age of 40.50 years. The second group included 12 pilots; 9 males and 3 females. The average age for the pilot group was 38.18 years. Their experience in flying hours varied from a low of 70 hours to a high of 7,200 hours, with a median of 625 hours.

EQUIPMENT AND TEST PROCEDURE.

The analog computer used for the critical tracking task was programmed to make a point of light drift from the center of an oscilloscope in a random fashion. The degree of instability called lambda (λ) can be varied, using vernier controls on the computer to provide different objective levels of workload. Lambda can be increased from low values of 0.5 (rad/sec) to higher values (3.0 rad/sec). Using a switch box containing an array of 10 pushbuttons (figure 1), workload responses were made once every minute in response to a "query" tone. The switches were wired through the computer to a strip chart recorder. Through the use of the computer and recorder, the variables listed in table 1 were continuously recorded.



82-66-1

FIGURE 1. TEN-POINT WORKLOAD RATING SCALE

TABLE 1 VARIABLES RECORDED ON THE CHART

VERTICAL DEVIATION
HORIZONTAL DEVIATION
DIFFICULTY LEVEL
RADIAL ERROR
INTEGRATED RADIAL ERROR
RADIAL STICK DEFLECTION
INTEGRATED RADIAL STICK DEFLECTION
WORKLOAD RESPONSE AND RESPONSE DELAY

Prior to participating in the tracking task, each participant was briefed in a conference room about his/her rights and the general tone of the experiment. A copy of this briefing can be found in appendix A. At the completion of the briefing, the researcher administered a short questionnaire (Subjective Units of Discomfort Scale (SUDS)) which focused on the participant's current level of stress and motivation. (See appendix B, Workload Evaluation: Preliminary Questions.) They were then escorted into the experimental room in which the equipment was located. They were seated at one-armed desks facing a CRT display. The participant's dominant hand (as determined by the experimenter asking) was placed on the joystick. The keyboard and joystick were adjusted for participants who were left-handed. Subjects were then briefed on the specific nature of the task, which was to keep the pip centered on the screen by moving the joystick. A practice period followed in which the subject was instructed to "fly" the pip clockwise, counterclockwise, diagonally, and across the horizontal and vertical axes. This phase was completed by attempting to keep the pip centered. During this time, the difficulty level (operating lambda) was set at 0.5 units (Jex, McDonnell and Phatak, 1966). The purpose of this training was to provide the opportunity for the participant to learn at a low level of difficulty. The training was terminated when the oscillation in the radial error was reduced to approximately 3 millimeters (mm) in magnitude. After a brief rest period, another session of centering practice was conducted with difficulty set at 1.0 units. During this period, training with the response box was accomplished. Participants were instructed to keep their nondominant hand physically on the box and to think continuously about how hard they were working. When they heard the query tone, they were instructed to push the button of their choice from 1 (very easy) to 10 (very hard) in response to how hard they felt they were working. Their response was indicated on the chart recorder. At the completion of this training session, actual data collection started. Figure 2 shows the general laboratory setup.

The tracking task generated by the analog computer (see appendix C) could be set to any level of difficulty, from very simple to very difficult. Because people vary in their ability after initial training, the maximum performance or critical tracking difficulty (critical lambda) was measured on each person prior to data collection. Each person was assigned his/her own unique administration of four levels of task difficulty (operating lambda) which were set at 0.25, 0.50, 0.75, and 1.0 of the individual's best performance (critical lambda). The research design is presented in table 2, which shows the balanced presentation order, across participants, that was developed to remove potential order effects from the design.

To measure the individuals maximal performance level or critical lambda, the researcher started with a low level of difficulty (0.5) and increased the difficulty until the individual lost control as defined by the pip hitting the border of the scope. When this occurred, difficulty was decreased until control was regained (defined by oscillations in radial error not exceeding 5 mm). The process was repeated again and the individual's maximal performance was taken as the highest prior to loss of control of the two trials. This was chosen based on preliminary research that indicated that averaging ascending and descending trials or selecting the lower value of the two trials did not adequately stress participants when exposed to values at their λ_c. Once this value was determined, the participant was exposed to two 4-1/2-minute blocks in accordance with the research design in table 2.

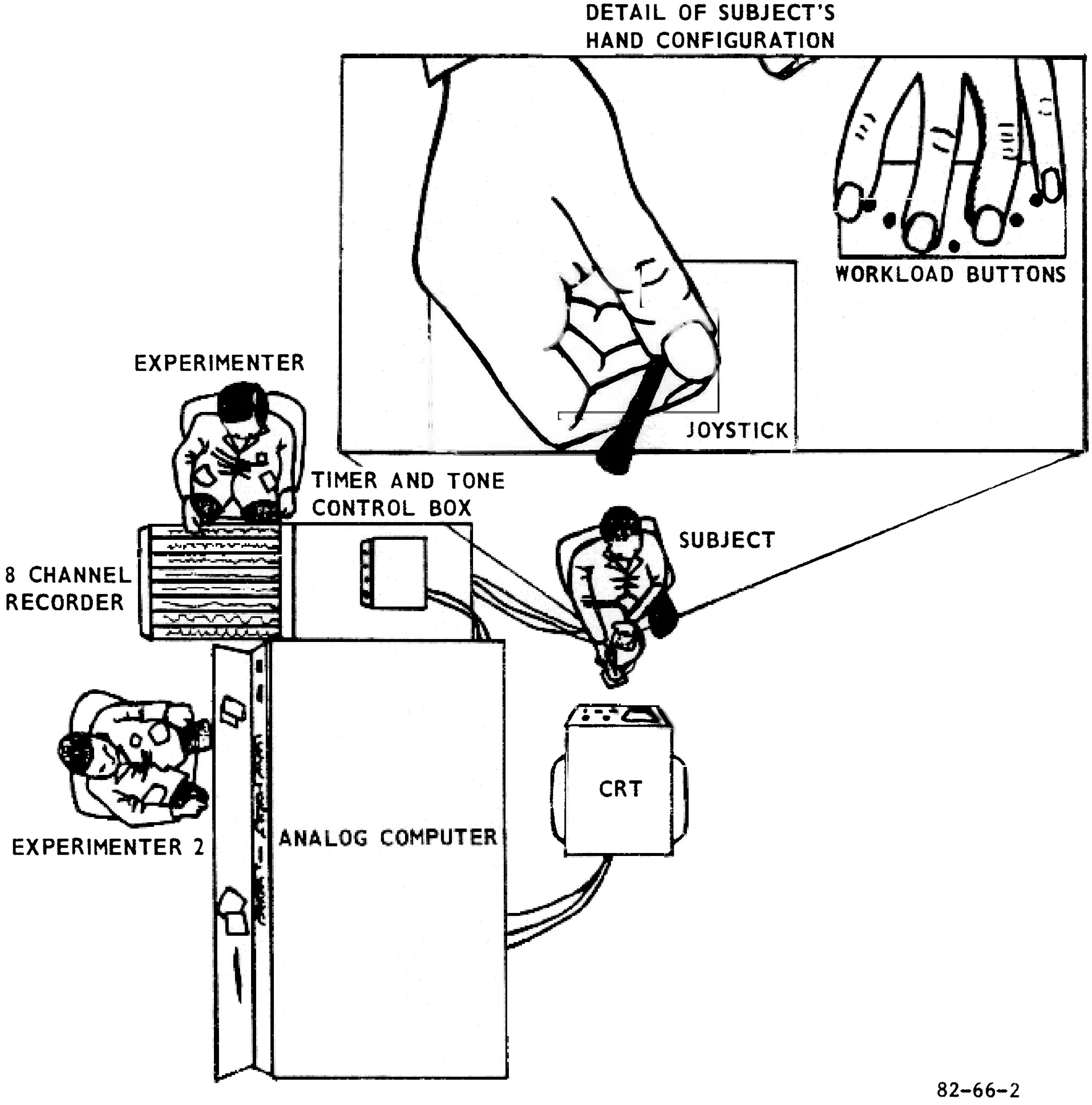


FIGURE 2. LABORATORY SETUP

TABLE 2. RESEARCH DESIGN

Balanced Sequential Block Presentation Order						
	SEQ. Block	SUBJ No.	Order of Presentation of Difficulty Levels			
			D1	D2	D3	D4
NONPILOTS G-1	1	1	3rd	4th	1st	2nd
		2	2	3	1	4
		3	1	2	3	4
		4	1	3	4	2
		5	4	1	2	3
		6	4	2	3	1
	2	7	2	4	1	3
		8	3	4	2	1
		9	2	3	4	1
		10	1	2	4	3
		11	3	1	2	4
		12	4	1	3	2
PILOTS G-2	3	13	1	4	2	3
		14	2	4	3	1
		15	4	3	1	2
		16	3	2	1	4
		17	2	1	3	4
		18	3	1	4	2
	4	19	1	3	2	4
		20	1	4	3	2
		21	4	2	1	3
		22	4	3	2	1
		23	3	2	4	1
		24	2	1	4	3

The operating lambda of each block represented any proportion of the participants λ_c from 0.25 to 1.0. Participants started tracking 30 seconds prior to the first query tone. However, the first 30 seconds of tracking and the response made at the first query tone were viewed as a familiarization phase, and these data were not used in the analysis. Four valid subjective responses and their consequent delays were collected in each trial block. There was a brief rest period between blocks of approximately 4 minutes. After the first two blocks were completed, the individual's critical lambda was again measured. This was done so that compensation could be made for the effects of learning and experience. The difficulty level blocks 3 and 4 were based on this second computation of λ_c. When the last two blocks were completed, a final measurement of λ_c was calculated as a check that the individual's measured ability had not changed drastically in one direction or another. This was quickly followed by the verbal administration of the SUDS and subsequent completion of the remaining questions in writing by the participant. A copy of the questionnaires is included in appendix B. The last step in the experiment was a debriefing of the participant. This was required so that any experimentally induced stress could be identified and reduced through discussion.

DATA ANALYSIS

TASK DIFFICULTY VERSUS EFFORT RATING.

The main purpose of the critical tracking task experiment was to determine if there was a relationship between various levels of objective task difficulty and subjective effort ratings made by participants during tracking. The results of the data analysis clearly show that such a relationship exists and that effort rating correlates with task difficulty. (See Results Summary.)

The major portion of the data analysis used the Analysis of Variance (ANOVA) technique with the variates listed in table 3. These include difficulty level, trial, group, and subjects within groups. The measures used in the data analysis are defined fully in table 4 and include critical lambda, operating lambda, effort rating, rating response delay, mean tracking error, mean stick deflection, operator gain, the transformed variables mean log tracking error, and mean log stick deflection.

Table 5 shows the means averaged across subjects and trials, and table 6 indicates the overall ANOVA results. Table 7 shows the F ratio for simple effects, one-way analysis of variance, and table 8 shows the results of multiple comparison tests among the means from the ANOVA in table 6.

TABLE 3. IDENTIFICATION OF VARIATES

VARIATE	SYMBOL AND DEFINITION OF LEVELS	
DIFFICULTY (Fixed)	Di, 1 = 1,2,3,4	Proportion of critical lambda where

	D1 = 0.25 times critical lambda
	D2 = 0.50 times critical lambda
	D3 = 0.75 times critical lambda
	D4 = 1.0 times critical lambda	

TRIAL (Fixed)	Tj, j = 1,2,3,4	jth minute of 4-minute block of tracking task trials at a constant difficulty

	T1 = 1st minute
	T2 = 2nd minute
	T3 = 3rd minute
	T4 = 4th minute	

GROUP
(Fixed)
	Gk, k = 1 and 2
	G1 = a group of 12 nonpilots
	G2 = a group of 12 pilots	

SUBJECTS WITHIN GROUPS (random)	S = 1,2,....12 S = a random variate described more fully in the text.	

TABLE 4. DEFINITION OF MEASURES USED IN DATA ANALYSIS

Critical Lambda Maximum value of divergence rate (radians per second) measured at the beginning, middle, and end of the experiment. It is a measure of the minimum continuous, dynamic reaction time of the subject, serving as a baseline (denominator) for the determination of operating lambdas representing the four difficulty levels.

Operating Lambda Value of the divergence "rate" (adjusted to each individual's maximum level) held constant within each of the four four-trial blocks.

Effort Rating Subjective rating of effort on a 10-point scale with 1 verbally anchored as "very easy" and 10 as "very hard," obtained every minute during the run.

Rating Response Delay Delay in making the rating in response to a query tone presented every minute.

Mean Tracking Error One-minute integral of radial tracking error

Mean Ln Tracking Error Average of the natural logarithm of each minute-by-minute tracking error value.

Mean Stick Deflection One minute integral of radial joystick deflection in volts. (See appendix A for more detailed information.)

Mean Ln Stick Deflection Average of the natural logarithm of each minute-by-minute stick deflection value.

TABLE 5. MEAN VALUES OF MEASURES USED IN THE DATA ANALYSIS

Difficulty Level	Group	Critical Lambda	Operating Lambda	Rating	Rating Delay	Tracking Error(TE)	Ln* TE	Stick Deflection(SD)		Operator Gain
0.25 Times Critical Lambda	Nonpilot	1.85	0.46	3.63	.20	42.44	3.70	55.48	3.86	1.34
	Pilot	2.41	0.60	2.67	.39	32.00	3.42	33.71	3.30	1.05
0.50 Times Critical Lambda	Nonpilot	1.86	0.93	6.15	1.05	102.31	4.49	105.10	4.51	1.13
	Pilot	2.39	1.20	5.10	.38	110.73	4.63	99.27	4.50	0.91
0.75 Times Critical Lambda	Nonpilot	.89	.42	7.60	.12	291.63	5.52	298.56	5.42	0.96
	Pilot	2.46	1.84	5.65	1.37	171.58	5.06	145.60	4.87	0.85
1.0 Times Critical Lambda	Nonpilot	1.75	1.75	8.63	1.05	458.06	6.04	409.06	5.89	0.88
	ilot	2.33	2.33	7.83	.40	377.23	5.86	321.85	5.66	0.84

*These values do not equal $\log_e$ (tracking error). They are the means of the Ln of the individual CELL values.

TABLE 6. SUMMARY OF OVERALL REPEATED MEASURES ANALYSIS OF VARIANCE FOR EIGHT MEASURES: SIGNIFICANT F RATIOS

Efffect Category	Variate (Source Of Effect)	Critical Lambda	Operating Lambda	Effort Rating	Rating Delay	Tracking Error	Ln Tracking Error	Stick Deflection	Ln Stick Deflection
Main Effects	Difficulty Levels	0	267.91**	70.72**	0	106.36**	206.13**	48.17**	152.59**
	Trials			0	0	5.29**	4.82**	3.61*	4.91*
	Groups	1.72**	13.27**	5.28*	0	4.61*	4.38*	0	0
Two-Way Interaction	D x T			0	0	1.94*	0	2.03*	0
	D x G	0	5.74**	0	0	3.45*	3.12*		3.15*
	T x G			0	0	0	0		0
Three-Way Interaction	D x T x G			0	0	0	0		0

* Sig. $\alpha \leq 0.05$
** Sig. $\alpha \leq 0.01$
F Ratios not sig. $\alpha < 0.05$ shown as zero
– Constant Across Trials

TABLE 7. SIMPLE EFFECTS REPEATED MEASURES ANALYSIS OF VARIANCE F RATIOS

Source Of Effect	Levels Within	Critical Lambda	Operating Lambda	Effort Rating	Rating Delay	Ln Tracking Error	Ln Stick Deflection
Difficulty Levels	G1 (Nonpilots)	0	70.43*	40.35*	0	83.92*	70.88*
	G2 (Pilots)	0	288.00*	32.18*	0	141.07*	85.01*
	D1	5.95*	5.97*	0	0	8.50*	6.59*
Groups	D2	10.64*	10.64*	0	0	0	0
	D3	8.90*	8.88*	6.62*	0	5.92*	5.08*
	D4	12.73*	12.73*	0	4.22	0	0

* F ratios significant at or beyond the $\alpha \leq 0.05$ level
0 Indicates no significance at the α 0.05 level

TABLE 8. SIMPLE EFFECTS REPEATED MEASURES ANALYSIS OF VARIANCE, MULTIPLE COMPARISON TESTS AMONG MEANS

Source of Effect	Levels Within	Critical Lambda	Operating Lambda	Effort Rating	Rating Delay	Ln Tracking Error	Ln Stick Deflection	Operator Gain
Difficulty	G1 (Nonpilots)	0	D1 D2 D3 D4	D1 D2 D3 D4	0	D1 D2 D3 D4	D1 D2 D3 D4	$\overline{\text{D4 D3}}\,\overline{\text{D2}}$ D1
Levels	G2 (Pilots)	0	D1 D2 D3 D4	D1 $\overline{\text{D2 D3}}$ D4	0	D1 D2 D3 D4	D1 D2 D3 D4	D4 D3 D2 D1
Groups	D1	G1 G2	G1 G2	0	0	G2 G1	G2 G1	0
	D2	G1 G2	G1 G2	0	0	0	0	0
	D3	G1 G2	G1 G2	G2 G1	0	G2 G1	G2 G1	0
	D4	G1 G2	G1 G2	0	(G1 G2)	0	0	0

Note: Lines joining means indicate no significant difference using the Neuman-Keuls test at an alpha level ≤ 0.05. Means not covered by the same line differ significantly.

A zero (0) indicates no significant difference exists between any of the means.

Except for the questionnaire data, the results that are discussed are based on the ANOVA, unless otherwise noted. The major question under consideration was to determine if there was a relationship between various levels of objective task difficulty and participants' subjective effort rating. Recall that the purpose of the experiment centered on this question. Since the pilot/nonpilot distinction was also of interest, tests of this variable were included in the analysis of variance. The task difficulty by pilot/nonpilot group interaction was not significant (table 6) which allowed us to examine the main effects directly. The influence of group membership; i.e., pilot/nonpilot and difficulty level, was evaluated separately. Figure 3 is the most informative representation of this data. As difficulty level increases for both groups, the effort rating increases also, in a very reliable manner. From figure 3, a difference between pilots and nonpilots also appears, with the nonpilots assigning generally higher effort ratings. The ANOVA shows that both the group membership and task difficulty variables produced significant main effects across the two groups and across the levels of difficulty. A test of multiple comparisons takes a closer look at these data and determines between which pairs of difficulty levels, for example, differences exist. It was decided to treat the data as if there had been an interaction, in order to remove any overlapping variance generated by difficulty and groups. This was done because of the differing pattern between pilots and nonpilots (note the dip in the line at D-3 for pilots in figure 3).

This procedure proved to be profitable. The simple effects (table 7) are main effects with overlapping variance removed. The results of this and subsequent post-hoc tests are shown at the bottom-right of figure 3. The nonpilots effort ratings at every difficulty level were significantly different from every other difficulty level. Pilots, however, tended not to discriminate difficulty across the two intermediate levels.

Pilots and nonpilots differed significantly only at D-3, an intermediate difficulty level. An alternate way of representing the data is shown in figure 4. Figure 4 shows histograms of effort rating versus difficulty level for both groups of participants. Each histogram contains 48 points representing 12 subjects with 4 trials each. Since the rating scale consists of 10 discrete levels (pushbutton) with a lower limit of 1 and an upper limit of 10, the distribution for the lowest difficulty level is skewed upward and for the highest difficulty is skewed downward. Due to the obvious deviation from a normal distribution (on which parametric techniques such as ANOVA are based) nonparametric analyses were performed. The results for the Friedman Analysis of Variance and multiple comparisons agreed with the results for the parametric analyses reported above.

Both pilots and nonpilots are willing and able to make effort judgments while tracking. Pilot and nonpilot effort ratings were not significantly different except at one (out of four) intermediate level of difficulty. This was of interest since nonpilots could possibly be used in future workload research where nonflying tasks are involved.

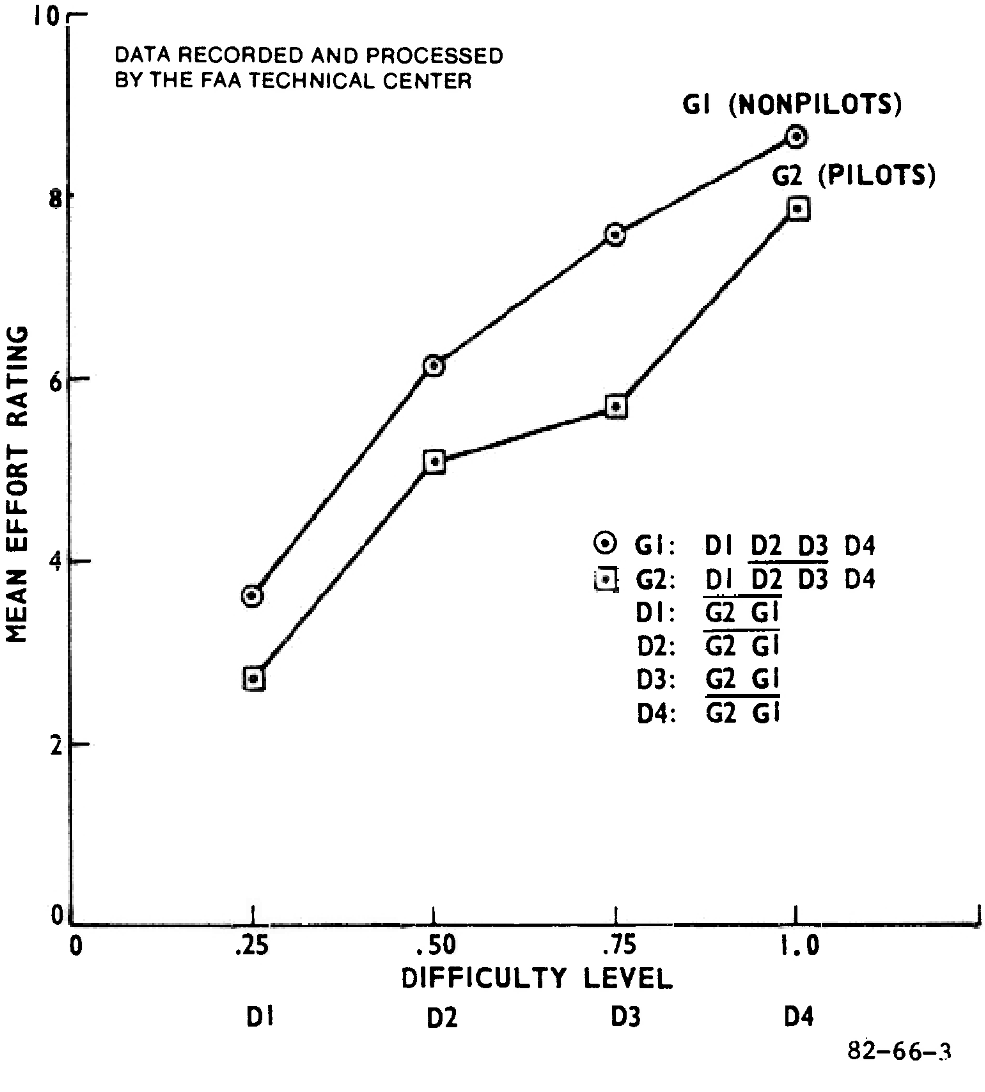
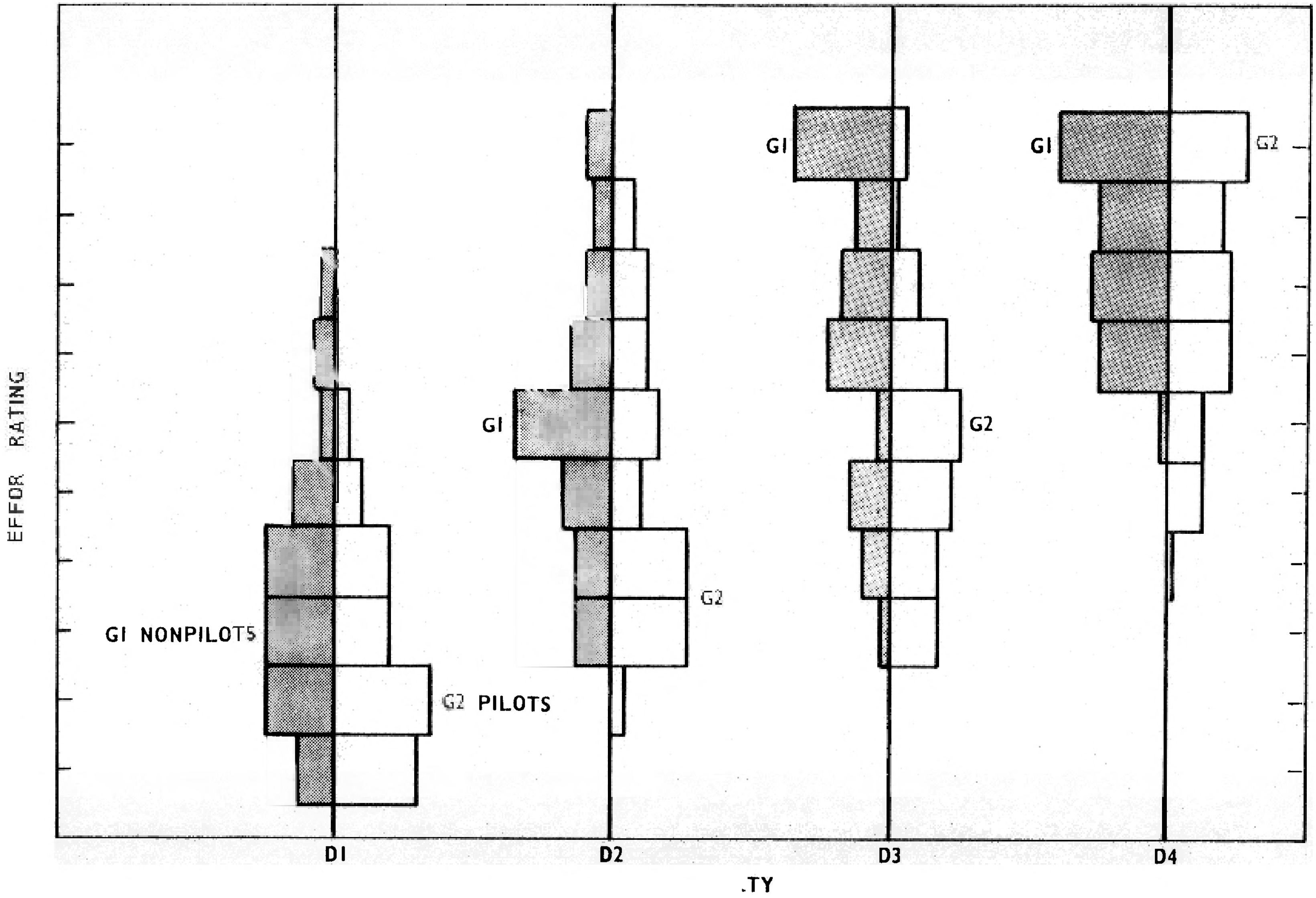


FIGURE 3. MEAN EFFORT RATING VERSUS DIFFICULTY LEVEL FOR NONPILOT AND PILOT GROUPS



GURE OGRAM EFF(RS .TY EV ON AN LOT

It was hypothesized that participants' delay in making a workload response would be related to task difficulty and would be an objective measure of effort. Rating delay is defined as the time between an audio prompt requesting an effort rating and the time the response was entered. Figure 5 presents the results and shows that across difficulty levels there is no systematic change in rating delay for either group. This finding was borne out by the ANOVA reported in table 6. The initial expectation was that the rating delay would increase as difficulty level increased. This did not happen. The reason for the lack of effect of difficulty level on rating delay may be due to the nature of the control task. Since there was only one input modality and one response modality, the switching of response was limited to the use of the POSWAT keyboard. It is recommended that further testing in a more complex task environment be done before reaching a final conclusion on the usefulness of rating delay as a workload measure.

It was of interest to determine if tracking experience gained by the participants during the experiment affected a person's level of performance (critical lambda) and further, to determine if pilots differed from nonpilots. Critical lambda reflects the individual's ability to deal with a maximum difficulty level based on his/her unique abilities and is a measurement of the maximum divergence rate of the pip from the center of the screen.

The results indicate that both pilots and nonpilots were slightly lower at the beginning and middle than at the end of the experiment. These results are shown in figure 6. It is interesting to note that pilots have significantly higher critical lambdas than nonpilots. This is not surprising given that they have more experience in complex perceptual-motor coordination tasks through flying modern aircraft.

The type of pattern for operating lambda that emerged across difficulty levels was evaluated, as well as whether or not a different pattern was seen for pilots and nonpilots. The ANOVA revealed that mean operating lambda across levels was not the same for pilots and nonpilots difficulty by group interaction (table 6). Operating lambda is the absolute divergence rate generated by the computer which serves as a fixed proportion of the individuals maximum or critical lambda. As one increases the difficulty level, operating lambda has to increase with the possible exception of recalibration. This is not a function of how the participant performs during the test trials but only during critical lambda measurement trials. The reason there was an interaction between difficulty and groups appears to be due to a more steady increase in the pilots operating lambda than that shown by the nonpilots. (See figure 7.) This means that pilots were operating at higher lambda levels throughout and confirms the findings already discussed.

Recall that each participant's critical lambda score was recalibrated midway through the experiment, in order to compensate for effects of fatigue or experience on an individual's critical lambda score. The order of presentation of difficulty levels was counterbalanced to further remove order effects. To determine what effect, if any, recalibration had, the average of mean critical lambda for both pilots and nonpilots was computed and plotted against difficulty level. (See figure 8.) What is apparent from this graph and confirmed by the ANOVA is though the critical lambdas of pilots and nonpilots significantly differed, there were no significant differences across difficulty levels, proving that the counterbalanced-experimental design was successful in removing both order and recalibration effects.

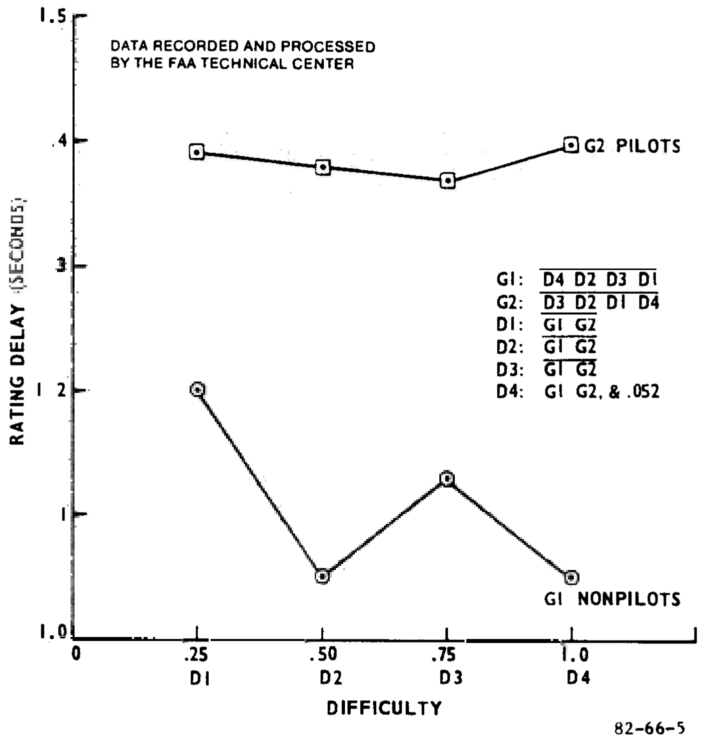


FIGURE 5. RATING DELAY VERSUS DIFFICULTY LEVEL FOR NONPILOT AND PILOT GROUPS

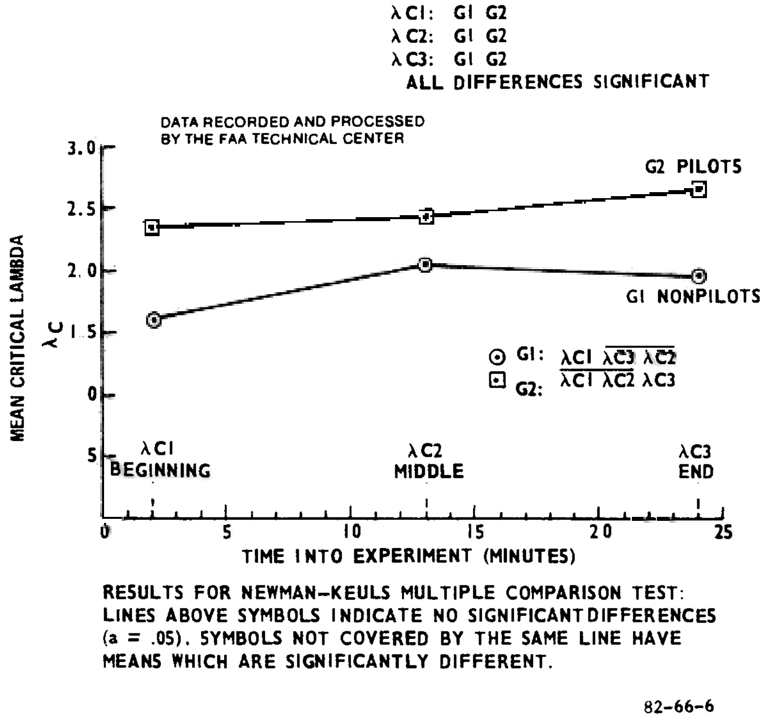


FIGURE 6. MEAN CRITICAL LAMBDA VERSUS TIME INTO EXPERIMENT FOR NONPILOTS AND PILOTS

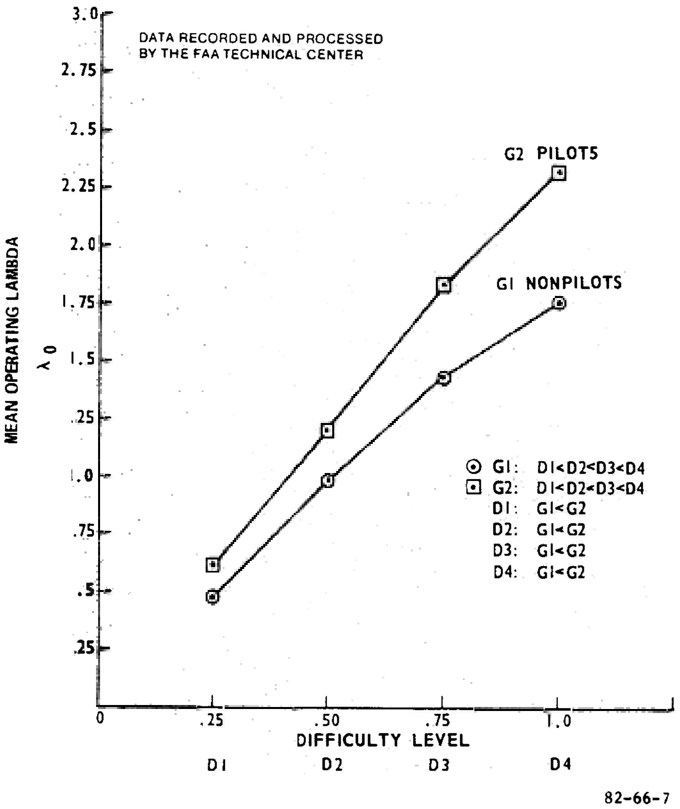


FIGURE 7. MEAN OPERATING LAMBDA VERSUS DIFFICULTY LEVEL FOR NONPILOT AND PILOT GROUPS

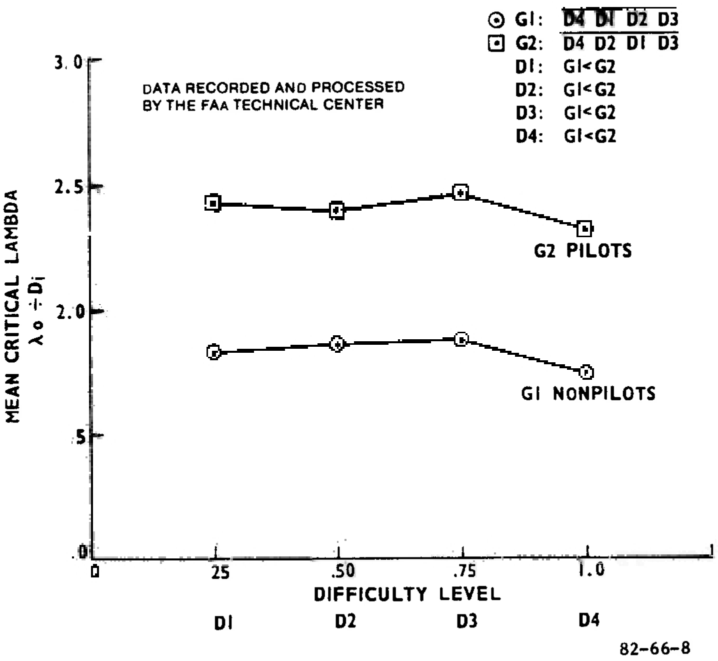


FIGURE 8. MEAN CRITICAL LAMBDA VERSUS DIFFICULTY LEVEL FOR NONPILOT AND PILOT GROUPS

SECONDARY VARIABLES.

Secondary variables which might be useful in the analysis were also considered. Two such variables were stick deflection and tracking error. Log transformations of these data were used to homogenize the variance to satisfy the assumption for the ANOVA. Figure 9 shows the relationship between Ln tracking error and difficulty. As difficulty increases, the magnitude of tracking error increases for both pilots and nonpilots. The crossing of the plots for the two groups indicates that there is probably an interaction between difficulty and groups (i.e., the two groups may be behaving differently). The analysis of variance indicated that this was in fact the case (tables 6 and 7). Tracking error at every difficulty level was significantly different from that at every other (table 8). Pilots and nonpilots were not making significantly different errors on two levels of difficulty, D2 and D4, but were making significantly different errors on D1 and D3. This was an interesting finding since pilots had such consistently higher critical lambdas. However, it should be recalled that the experimental design was adjusted for individual ability by setting difficulty level as a proportion of the operators' maximum performance or critical lambda. If this had not been done, it is likely that pilots would have made smaller errors throughout, since they were not adequately challenged.

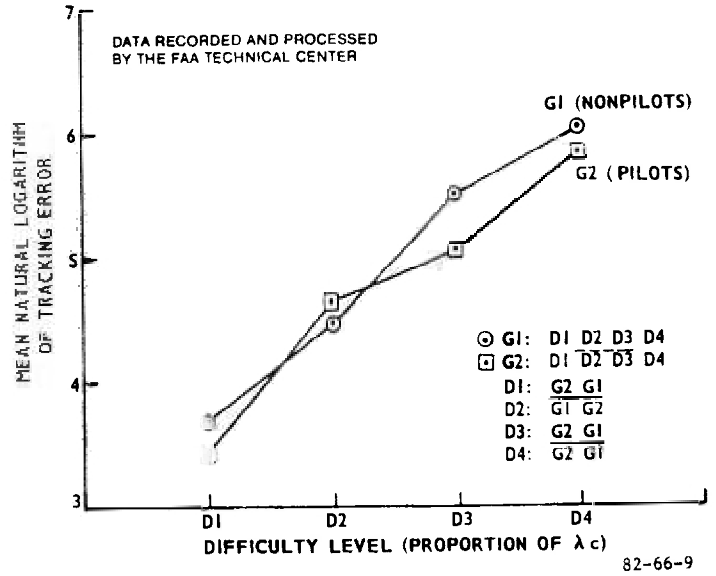

DATA RECORDED AND PROCESSED BY THE FAA TECHNICAL CENTER

FIGURE 9. MEAN NATURAL LOGARITHM OF TRACKING ERROR VERSUS DIFFICULTY LEVEL FOR NONPILOT AND PILOT GROUPS

The analysis for stick deflection was very similar to that for tracking error. The more difficult the task, the more stick movement the operator was required to use to control the system. (See tables 6, 7, and 8.) It also indicates that pilots and nonpilots differed at two levels of difficulty, D1 and D3, with the pilots recording less stick deflection. It seemed that pilots were not putting in as much physical effort at these two levels of difficulty. Why this difference was not consistent for D2 and D4 is unclear. Reported effort between the two groups was only different at D3 where the pilots indicated a lower level of effort. The level of agreement between effort rating and the two variables discussed above was evaluated by means of correlation.

Table 9 indicates what must already be apparent. Difficulty drives effort ratings, and difficulty is directly related to tracking error and stick deflection. Thus, it is easy to see the relationship between these two variables and effort rating. Tracking error and stick deflection may be viewed as indicators of physical effort; and the higher they are, the higher reported effort ratings are. It should be noted that this relationship is far from perfect and confirms that there is more to effort rating than can be seen in observable operator behavior.

The question of what data collected had potential for discriminating between pilots and nonpilots was also investigated. The technique employed for this analysis was discriminant function analysis. This statistical tool attempts to produce a weighted linear combination of variables which would best distinguish between membership in two nonoverlapping groups. The advantage of such analysis is that it may be instrumental in deciding what types of measures might later be used to separate personnel on a performance continuum. This procedure was repeated for all four difficulty levels. The percentages of correct classifications are reported in table 10.

A chi square analysis was applied to determine if the assignment of participants to the two respective groups was accurate beyond chance. In other words, could we have done equally well by randomly labeling participants as pilots and nonpilots without knowing anything about their performance in the experiment? Using a weighted combination of variables, group assignment was more accurate than could be expected by chance alone for all difficulty levels except for the least difficult, D1. It would appear that as difficulty increases, the differences in performance between pilots and nonpilots becomes easier to identify using the pool of measures employed in this experiment.

TABLE 9. RELATIONSHIP OF EFFORT RATING TO STICK DEFLECTION AND Ln TRACKING ERROR

Variable	Group	Correlation
Ln Tracking Error	Nonpilots	0.80
	Pilots	.75
Ln Stick Deflection	Nonpilots	.67
	Pilots	.63

TABLE 10. DISCRIMINANT FUNCTION ANALYSIS TO SEPARATE NONPILOTS FROM PILOTS

	Percent Correctly Classified				
Difficulty	Nonpilots	Pilots	Total	Chi Square	Variables Used
D1	66.7	66.7	66.7	2.67	LnTE
D2	75.0	75.0	75.0	6.00*	LnEffort Rating
D3	75.0	83.3	79.2	8.22*	LnTE, Ln Delay
D4	75.0	91.7	83.3	10.97*	LnTE

*Significant 0.05

QUESTIONNAIRE DATA ANALYSIS.

The post-tracking questionnaire provided useful data for addressing additional areas. For example, after the experiment was completed, was it possible for participants to recall the relative difficulty and subsequent effort during the administration of the four difficulty levels. What was really desired was knowledge of whether traditional post-task questionnaires are reliable indicators of what participants experienced during the experiment. The second question in the workload evaluation task questionnaire (see appendix B) asked the participant to rank-order the four levels of work difficulty. If the participants received the administration of difficulty at 0.25, 0.50, 0.75, and 1.0 of critical lambda, then a correct response would have been to rank-order from least difficult to most difficult as follows: A, B, C, and D. Recall that each participant received the levels of difficulty in a counterbalanced order. Results indicated that personnel were not able to correctly recall the difficulty order after tracking was completed. The median correct percentages are presented at the bottom of table 11. Percentage correct was computed by determining how many of the four difficulty levels were correctly assigned a rank position. The importance of minute-by-minute effort rating data collection during the experiment cannot be overstated, given the poor recall of participants on this critical question.

Finally, what was the attitude of personnel towards this experiment? This question is intentionally broad in order to encompass a number of problems that were addressed with both pre- and post-experiment questionnaires.

A preliminary questionnaire asked participants to rate their anxiety level from 1 (at ease) to 100 (very tense) and also to evaluate their performance motivation on a 10-point scale. These questions were meant as a rough estimate and were not standardized on a sample. Results are presented in table 11. No significant difference between pilots and nonpilots is reported, although the anxiety scale mean appeared lower for the pilots while their performance motivation appeared higher. There was a great deal of individual variation within each group.

After the experiment, the subjects were again asked to subjectively rate their level of anxiety from 1 to 100. Although anxiety appeared to increase for both pilots and nonpilots, the increase was not significant. Also, the performance motivation scale from the preliminary questionnaire correlated negatively with the anxiety scale after the experiment (table 12). In other words, the more motivated the participant said he/she felt before the experiment, the less anxious he/she indicated after completion.

TABLE 11. WORKLOAD QUESTIONNAIRE RESULTS

	NONPILOTS		PILOTS	
QUESTION	MEAN	SD	MEAN	SD
Subjective units of discomfort (1)*	33.75	25.83	23.50	20.59
Subjective units of discomfort (2)	46.42	31.88	35.33	26.85
Performance Motivation*	7.5	1.73	8.25	1.22
Tracking Task ** Difficulty				
Demanding	7.5	1.62	7.66	2.10
Exciting	5.18	1.99	6.16	1.99
Boring	3.66	1.92	2.75	2.22
Undemanding	2.25	1.86	1.72	.90
Workload Buttons				
Comfortable	5.73	2.32	5.0	1.70
Distracting	5.64	3.36	6.17	3.53
Accurate	4.33	2.02	6.33	2.10
Query Tone				
Too Loud	3.83	2.69	2.08	1.16
Too Frequent	3.83	2.52	2.17	1.34
Difficulty Rank Order				
Median % Correct	37.5%		12.5%	

*These values measured on pretest questionnaire

**Note that numbers beyond 5.5 indicate agreement with the descriptor in the left column, while numbers below 5.5 indicate disagreement.

When asked to evaluate the tracking task, participants indicated agreement that the task was difficult and disagreement with statements which indicated that it was boring. Subjects were also asked to evaluate the workload buttons they had to push every minute. Responses to these questions were inconclusive, and the average response fell mid-scale between agreement and disagreement.

Finally, regarding the tone used to signal a workload response, participants generally agreed that it was neither too loud nor too frequent. An additional item of interest was the interrelationship of the questionnaire items to each other, specifically, to determine if there was much redundancy across the questions. Were subjects asked the same thing more than once using different words? Intercorrelations were computed between the responses to all pairs of questions. The majority of these relationships were not significantly different from zero. The remainder, which exceeded the cutoff for significance (0.404 or -0.404), are reported in table 12.

It is apparent that even the significant correlations were low to moderate at best. This demonstrated that, for the most part, participants responded to each question independently. There was little redundancy in the questionnaire data which insured sampling from a variety of areas of individual attitude toward the experiment.

TABLE 12. QUESTIONNAIRE INTERVARIABLE CORRELATIONS

Scale Title	r
Performance Motivation--Post SUDS	-0.507
Performance Motivation--Task Exciting	.442
Performance Motivation--Measure Accurate	.420
Task Demanding--Task Undemanding	-.463
Task Exciting--Task Boring	-.427
Task Exciting--Buttons Comfortable	-.432
Task Boring--Buttons Comfortable	.404
Tone Too Loud--Tone Too Frequent	.612

RESULTS SUMMARY

Both pilots and nonpilots discriminated between the difficulty levels using the pushbutton effort rating system. The pilots, however, did not separate the two intermediate levels of difficulty. While nonpilots reported somewhat higher effort across all four difficulty levels, the difference between their ratings and those of the pilots was only significant at D-3, an intermediate level.

The rating delay made by both groups did not differ significantly across difficulty levels. This measure should not be rejected, however, without further study to determine if its failure was an artifact of experimental design.

Critical lambda, the maximum performance ability of participants, did not change significantly as participant experience with the tracking task increased. This indicated that critical lambda may be a stable measure of individual ability which transcends situational experience. Pilots achieved significantly higher critical lambdas and, as a consequence, their operating lambdas (the result of applying a proportion to critical lambda for each level of difficulty) were also higher.

As difficulty increased, the amount of tracking error also increased for both pilot and nonpilots. The magnitude of pilot error was significantly less than nonpilots on two (D1 + D3) out of the four difficulty levels. If the difficulty had not been adjusted for individual ability, it is probable that pilots would have had lower error scores on all four difficulty levels.

The amount of tracking error correlated relatively well with the amount of effort reported by both groups of participants. This relationship also existed between effort and control input (stick deflection) but was not as strong.

A discriminant function analysis proved that pilots and nonpilots could be separated by their performance in the experiment. This separation was significant at the three higher levels of difficulty, but not at the lowest level D1. Such techniques could be potentially useful to separate personnel into appropriate performance categories in other studies.

After the experiment was completed, participants were unable to accurately recall the order of difficulty presentation they had experienced. This adds to the importance of minute-by-minute effort rating data collection during the experiment itself.

Attitudes toward the experiment did not differ significantly between pilots and nonpilots. The higher the stated motivation was before the experiment, the lower the reported stress was after the experiment. The subjects indicated that the tracking task was difficult and was not boring. They were unclear as to whether the effort rating buttons were distracting, and they indicated that the response query tone was not annoying.

It was apparent that both pilots and nonpilots were willing and able to distinguish between counterbalanced levels of difficulty in a tracking task similar to that imposed by instrument approaches using a localizer. A distinction between difficulty levels was reflected in subjective effort ratings. This finding was in direct contrast to the results of the questionnaire data analysis that indicated participants were not able to accurately recall difficulty levels for each trial. This was an anticipated result, and it serves as justification for measuring effort ratings during the tracking experiment rather than at the conclusion of an experimental session.

CONCLUSIONS

The Pilot Objective/Subjective Workload Assessment Technique (POSWAT) used to measure effort rating on a regular basis during this experiment was found to be practical, minimally intrusive, and informative. The concept merits further evaluation in a cockpit environment.

The critical tracking task is a valuable research tool for investigating workload rating scales providing, as it does, accurate and easily controllable difficulty levels and objective measures of performance.

Subjects were able to discriminate levels of effort involved in controlling a critical tracking task at four different divergence rates (difficulty levels) using the POSWAT rating scale. In the one case in which there was a nonsignificant change in rating with an increase in difficulty level, the rating curve closely matched those for Ln tracking error and Ln stick deflection. This indicates that subjective effort ratings faithfully reflect differences in objective performance and level of difficulty.

Effort rating varied as a function of tracking error or of stick deflection more closely than with difficulty level as defined by proportion of critical lambda.

Rating delay did not vary in any reliable manner as a function of difficulty level. Pilot subjects reported significantly lower effort ratings and obtained significantly higher critical lambda values than the nonpilot subjects.

Participants were unable to identify difficulty level presentation order in the post-test debriefing session. This contrasts with their generally accurate discrimination obtained from the minute-by-minute effort rating using the 10-key POSWAT keyboard.

Discriminant function analysis was found to be a useful technique for determining which measures could best be used to differentiate between participant groups.

RECOMMENDATIONS

A continuation of the development of POSWAT for use as part of a comprehensive pilot effectiveness measurement test battery is recommended. Also, the use of the critical tracking task prior to and following flight experiments is recommended as a measure of the pilot's level of psychomotor functional ability. This test may help to account for day-to-day and fatigue-induced variation in subject performance. It may also be useful for the categorization of the skill level of subjects in future studies and for an investigation of scaling and anchoring questions.

It is recommended that in future studies, the amount of training on the critical tracking task be increased to avoid more than random variability in the results.

REFERENCES

1. Albrecht, A.P., Development of a Program Document to Support the Proposed Enhancement Human Factors Program, April 1981.

2. Chiles, W., and Alluisi, R.A., On the Specification of Operator or Occupational Workload with Performance Measurement Methods. Human Factors, 1979, 21(5) 515-528.

3. Eggemeir, F.T., Some Current Issues in Workload Measurement. Proceedings of The Human Factors Society - 24th Annual Meeting, 1980, 669-673.

4. Goerres, H.P., Subjective Stress Assessment as a Criterion for Measuring the Psychophysical Workload on Pilots. Proceedings of the AGARD Conference on Methods of Assessing Workload, AGARD-CP-216, 1977, b11-1-b11-8.

5. Hess, R.A., Nonadjectival Rating Scales in Human Response Experiments. Human Factors, 1973,, 15(3), 275-280.

6. Hicks, T.A., and Wierwille, W.W., Comparison of Five Mental Workload Assessment Procedures in Moving Base Simulator,Human Factors, 1979, 21, 129-143.

7. Jex, K.R., McDonnell, J.D., and Phatak, A.V., A Critical Tracking Task For Manual Research Control Research. IEEE Transaction on Human Factors in Electronics, Vol. HFE-7, 1966, 138-145

8. Johannsen A., Workload and Workload Measurement. In N. Moray (Ed.: Mental Workload: Its Theory and Measurement. New York: Plenum Press, 1977, 3-11.

9. Katz, J.A., Pilot Workload in the Air Transport Environment: Measurement Theory and the Influence of Air Traffic Control, Flight Transportation Laboratory, MIT, Cambridge, Mass., May 1980 (FTL Report R80-3).

10. Kreifeldt, J.A., Cockpit Displayed Traffic Information and Distributed Management in Air Traffic Control.Human Factors, 1980, 22 (6), 671-691.

11. Philipp, V., Reiche, D., Kirchner, J.H., The Use of Subjective Rating. Ergonomics, 1971, 14, 611-616.

12. Poulton, E.C., Bias in Ergonomic Experiments. Applied Ergonomics, 1973, 4(1). 17-18.

13. Rehmann, J.T., Cockpit Display of Traffic Information and the Measurement of Pilot Workload: An Annotated Bibliography. FAA Technical Center Technical Report, FAA-EM-81-9, 1982.

14. Rolfe, J.M., and Lindsay, S.J., Flight Deck Environment and Pilot Workload: Biological Measures of Workload, Applied Ergonomics, 4, 1973, 199-206.

15. Rosenberg, B., Working Paper 81.2R: Description of a Pilot Subjective/Objective Workload Assessment Technique, January 1981.

16. Sheridan, T.B., and Simpson, R.W., Toward the Definition and Measurement of the Mental Workload of Transport Pilots, Technical Report, Flight Transportation/Man Machine Lab, MIT, Cambridge, Mass., DOT-OS-70055, 1979.

17. Willeges, R., and Wierwille, W., Behavioral Measures of Aircrew Mental Workload. Human Factors, 1979, 21, 549-574.

APPENDIX A

WORKLOAD EVALUATION PARTICIPANT BRIEFING

WORKLOAD EVALUATION
PARTICIPANT BRIEFING

1. Personal Introduction

Hello, my name is ________________. I will be briefing you on what you will be doing for the next hour or so. If you have any questions at any time, feel free to stop me and I will try to answer them.

2. General Project Information

The FAA is working on a joint project with NASA to evaluate the usefulness of a new concept in pilot information displays. This is referred to as the Cockpit Display of Traffic Information (CDTI). It will provide a pilot with an awareness of other aircraft in the vicinity of his own ship. In order to properly investigate this concept we must develop new measurement techniques so that we can determine how the CDTI will affect pilot performance and workload. In our preliminary research we are employing participants to see if we can establish adequate measures of workload or how hard the individual is working. I will explain the tasks shortly. I think you will find them interesting and challenging.

3. Voluntary Participation and Privacy

You are here as a volunteer and we sincerely appreciate your help. You may terminate your participation at any point. However, if you do, the effort you have put in to that point will be wasted for our data collection purposes. Your privacy is being protected because we are not recording your name on any of our forms or in our records. We are not interested in evaluating your performance as an individual but rather in using your efforts to demonstrate the sensitivity of our measurement systems.

a) have subject complete preliminary SUDS scale

4. Specific Task information

(Individual is seated in the experimental room with the scope in front of him/her). What you will be doing today is controlling the movement of a spot of light in front of you by using a joystick which operates very much like its namesake in an airplane. When you wish to move the light upward you pull back on the stick. Likewise downward motion involves pushing the stick forward. Right or left motion is self-explanatory. Try now to move the light up, down, right, and left. Now that you have the feel of the joystick, I will explain the use of the grey box with the buttons on it. You will use this box to indicate how hard you are working at a given point in time. You will make this response each time you hear a tone which sounds like this: (query tone is sounded). You must evaluate how hard you are working from 1 (very easy) to 10 (very hard). You should make this response as quickly after the tone as you can. We suggest that you think about how hard you are working between tones and count the buttons from left to right by touch to approximate your current level. When the tone sounds you should be within one button more or less of your current evaluation.

APPENDIX B

WORKLOAD EVALUATION QUESTIONNAIRES

Once we begin the experiment the point of light will be centered on the screen and your job will be to keep it there by moving the joystick. The light will "wander" from the center unless you continually move it back. It is very important that you try as well as possible to keep the light centered. We are recording the amount of time that it remains off center.

Now we will being a practice period so that you can learn to operate the equipment. This will last about 5 minutes.

WHAT CERTIFICATES/RATINGS DO YOU HOLD?

_____STUDENT
_____PRIVATE
_____COMMERCIAL
_____ATP

_____SINGLE ENGINE
_____MULTI ENGINE
_____LAND
_____SEA
_____INSTRUMENT
_____CFI

TOTAL FLYING TIME________________________HOURS

HOURS IN LAST TWELVE MONTHS____________

Date Participant Number

____

WORKLOAD EVALUATION
PRELIMINARY QUESTIONS

You have just been briefed on what you will be doing for the next hour. If you have any questions at any time feel free to

Before we begin the experiment,we have a few questions. Please be as honest as you can. Remember that your name is not being recorded. This data will be used for research purposes only.

1. First, we would like to know how you feel at this moment. Imagine the range of your feelings from being very calm, relaxed and at ease (1 to being very tense, excited and upset (100). Assign a number from 1 to 100 which best describes how you feel at this very moment

____
Write your number here

2. Next, we need a measure of your performance-motivation By this is meant your evaluation of how hard you tend to work at tasks. Recognizing that this will vary from one task to next, try to evaluate based on averaging across the different things you do at home and at work. Choose a number from 1 (average - try to get by) to 10 (high-work very hard at everything)

(circle one)
Average 1 2 3 4 5 6 7 8 9 10 High

Thank you for your help. The next step is to participate in our experiment which you should find interesting.

Date Participant Number

____

WORKLOAD EVALUATION
TASK QUESTIONS

You have just completed your participation in our tracking task exercise. We appreciate your help and very much need your honest answers to the following questions inorder to perfect our measurement system. Again, we remind you that your name is not being recorded and no attempt will be made to identify you in our records. Data will be used for research purposes only.

1. Now that you have completed our exercise, we would like to know how you feel at this moment. Imagine the range of your feelings from being very calm, relaxed and at ease (1 to being very tense, excited and upset (100). Assign a number from 1 to 100, which best describes how you feel at this very moment.

____
Write your number here.

2. During this experiment you were exposed to four levels of work difficulty, which were presented in a scrambled order. Assume the order you received was: A, B, C, D. Please rank order these levels from most to least difficult to accomplish. Fill in the letters as indicated below.

Most Difficult 1
2
3
Least Difficult4

Place 1 letter next to each number

The next series of questions each involve a statement followed by a scale of agreement or disagreement. Circle a number from 1 (strongly disagree) to 10 (strongly agree) which best describes your level of agreement with the statement.

4. The tracking task I participated in was:

Strongly Disagree	1	2	3	4	5	6	7	8	9	10	Strongly Agree

circle one

Demanding	1	2	3	4	5	6	7	8	9	10
Exciting	1	2	3	4	5	6	7	8	9	10
Boring	1	2	3	4	5	6	7	8	9	10
Undemanding	1	2	3	4	5	6	7	8	9	10

5. The workload buttons which I had to push every minute were:

Comfortable	1	2	3	4	5	6	7	8	9	10
Always Distracting	1	2	3	4	5	6	7	8	9	10
An Accurate measure of work load	1	2	3	4	5	6	7	8	9	10

6. The tone used to signal my workload reponse was:

Too Loud	1	2	3	4	5	6	7	8	9	10
Too Frequent	1	2	3	4	5	6	7	8	9	10

7. Feel free to comment on anything you feel is important in our development of this experiment.

APPENDIX C

TRACKING TASK

TRACKING

The tracking task is a divergent positive feedback loop shown in block diagram form in figure 1.

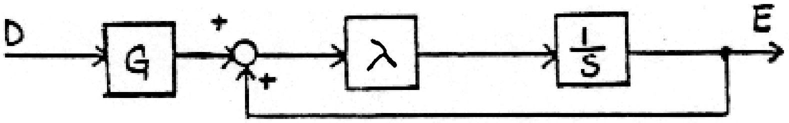


FIGURE C-1. BLOCK DIAGRAM OF TRACKING TASK

This loop has the transfer function:

$$\frac{E}{D} = \frac{G}{(s/\lambda)-1}$$

Where: D = Stick deflection input
E = Target Error Output
G = Control gain
λ = Error Rate gain, lambda, in radius per second
1/s = Time integration

In this loop, the rate of divergence of the error output is proportional to the error magnitude plus stick deflection. In analog computer form, the tracking task is represented by the diagram of figure 2, with potentiometers set for fixed values of G and λ :

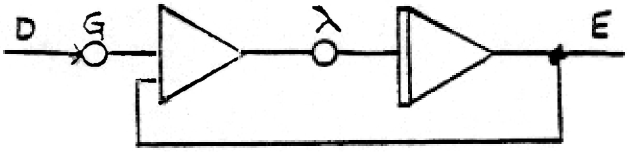


FIGURE C-2. ANALOG DIAGRAM OF TRACKING TASK

For the case of the varying λ , a multiplier is substituted for the λ potentiometer, to accept a variable λ input signal, as shown in figure 3. Multiplier input connections are arranged to preserve positive feedback in the loop.

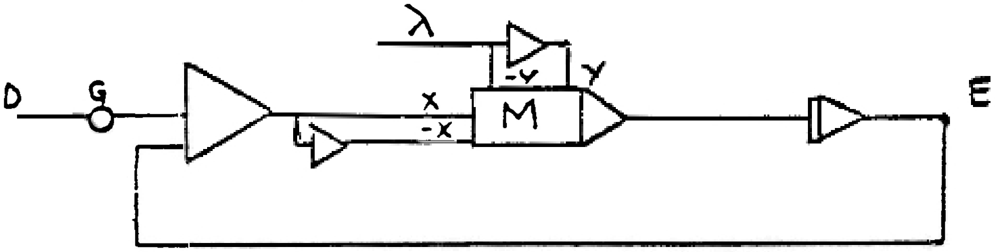


FIGURE C-3. ANALOG CIRCUIT FOR TRACKING TASK WITH VARIABLE LAMBDA

A tracking task system was implemented on a Donner Model SD-80 analog computer to develop vertical and horizontal output signals in response to manually operated joystick deflections. The block diagram of figure 4 illustrates this system. Nomenclature for the diagram is presented in table 1.

Two divergent positive feedback loops are used, one for the vertical tracking task and one for the horizontal. Longitudinal stick motions provide inputs to the vertical task, and lateral stick motions to the horizontal task. The output responses from the task represent target error signals, which are applied to the appropriate vertical and horizontal deflections of an oscilloscope. The error signals deflect a dot which moves about the face of the oscilloscope in response to joystick inputs. At zero error, the dot is centered on the oscilloscope.

Centering bias adjustments are provided on the computer to trim the stick input signals to zero when the stick is centered. A low-amplitude sine wave function is also added to each stick input to keep the signal active when stick signals are small. The sine wave frequencies and amplitudes are individually adjustable

The error rate gain, λ, is controllable in a number of ways: (a) the value of λ can be held at a constant value by closing the reset switch on the integrator and selecting the desired λ value with the initial λ setting; (b) the λ value can be caused to increase from the initial value at a constant rate by placing the rate input switch in the positive position, selecting the desired λ rate setting, and opening the reset switch on the λ integrator; (c) variation of λ can be stopped at any existing magnitude by placing the rate input switch in the center position, (d) a decreasing λ value can be produced by placing the rate input switch in the negative position, when the reset switch is open. The decrease will occur at the rate set on the λ rate control; and (e) the value of λ can be returned to the initial setting by closing the reset switch on the λ integrator.

Either of the tasks can be immobilized, to leave only a single axis active, by closing the reset switch on the desired error integrator. Both error signals can be returned to the center to restart the problem by closing the reset switches on both error integrators. This can be accomplished by placing the analog computer to reset.

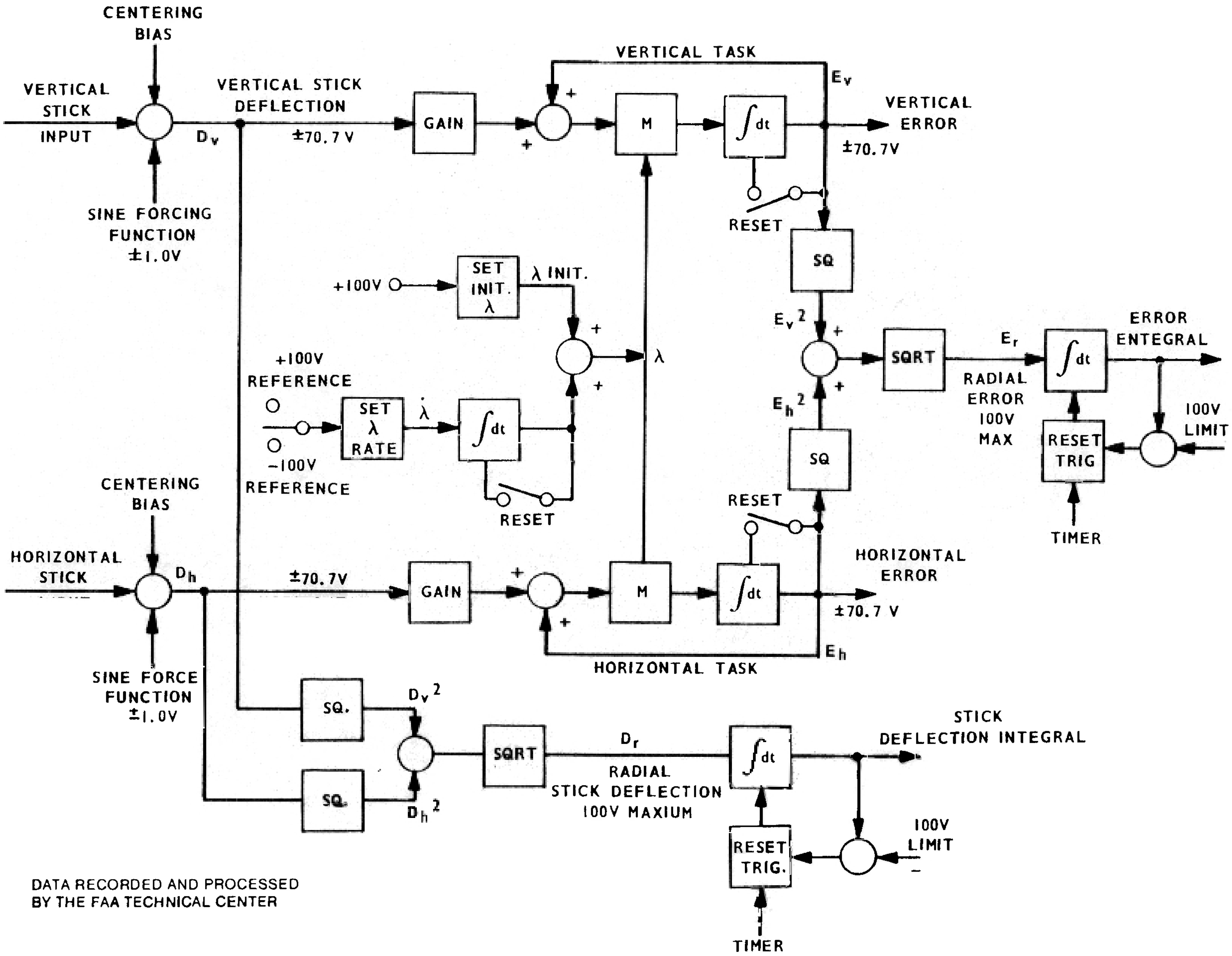
A circuit is provided to convert the vertical and horizontal stick deflection magnitudes into a single radial deflection value. This is accomplished by a hypotenuse computation which calculates the square root of the sum of the squares of the vertical and horizontal magnitudes.

An integrating circuit provides a summation of the radial stick deflections over a period of time. When the sum of the deflections reaches 100 volts, a reset trigger circuit returns the summation to zero. An external relay input is also provided to return the summation to zero when a reset clock pulse is received from a test period timer, each minute. At the end of a test period, the total summation of stick deflections is determined by the number of resets plus the final integrator magnitude. Identical circuits are provided for developing the radial magnitudes of the error output signals, and integrating these magnitudes to provide an error summation. Records of the variables, including stick deflections, error magnitudes, and their summations are made as time histories on a eight-channel Brush strip-chart recorder.

Analog computer mechanization of the tracking task system is illustrated by two figures. Figure 5 presents an analog diagram of the vertical and horizontal tracking tasks, with provisions for controlling the λ value. Figure 6 presents the hypotenuse computation which develops the radial values of stick deflection or error magnitude. This diagram also includes the integrator circuit for summation of these magnitudes.

TABLE C-1. NOMENCLATURE

D_V = Vertical Stick Deflection
D_H = Horizontal Stick Deflection
D_R = Radial Stick Deflection
E_V = Vertical Error Magnitude
E_H = Horizontal Error Magnitude
E_R = Radial Error Magnitude
M = Multiplication
SQ = Squaring Computation
SQRT = Square Root Computation
S = d/dt, differential operator, 1/sec
λ = error rate gain, reciprocal of first order divergence time constant, 1/sec



82-66-C-4

FIGURE C-4. TRACKING TASK BLOCK DIAGRAM

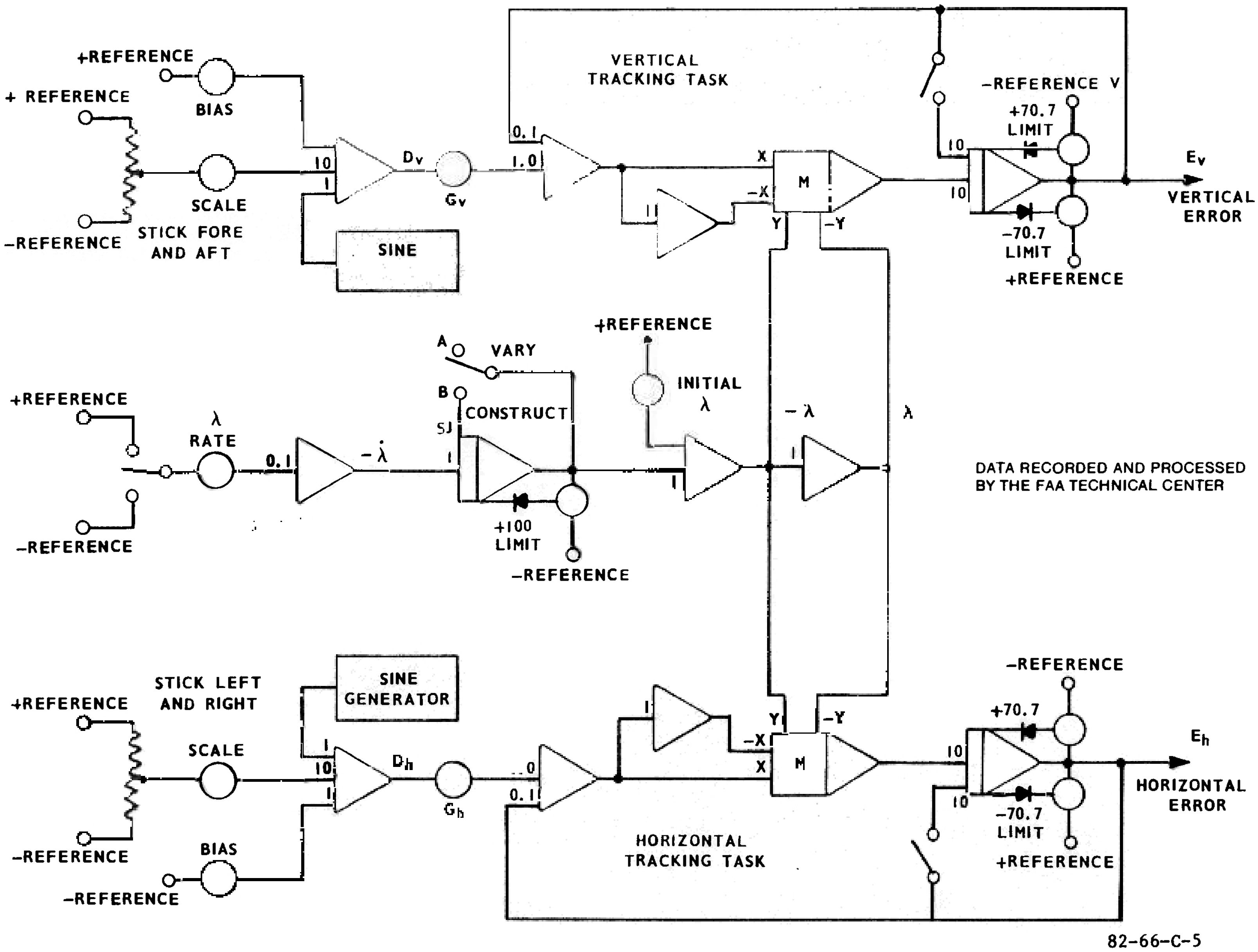
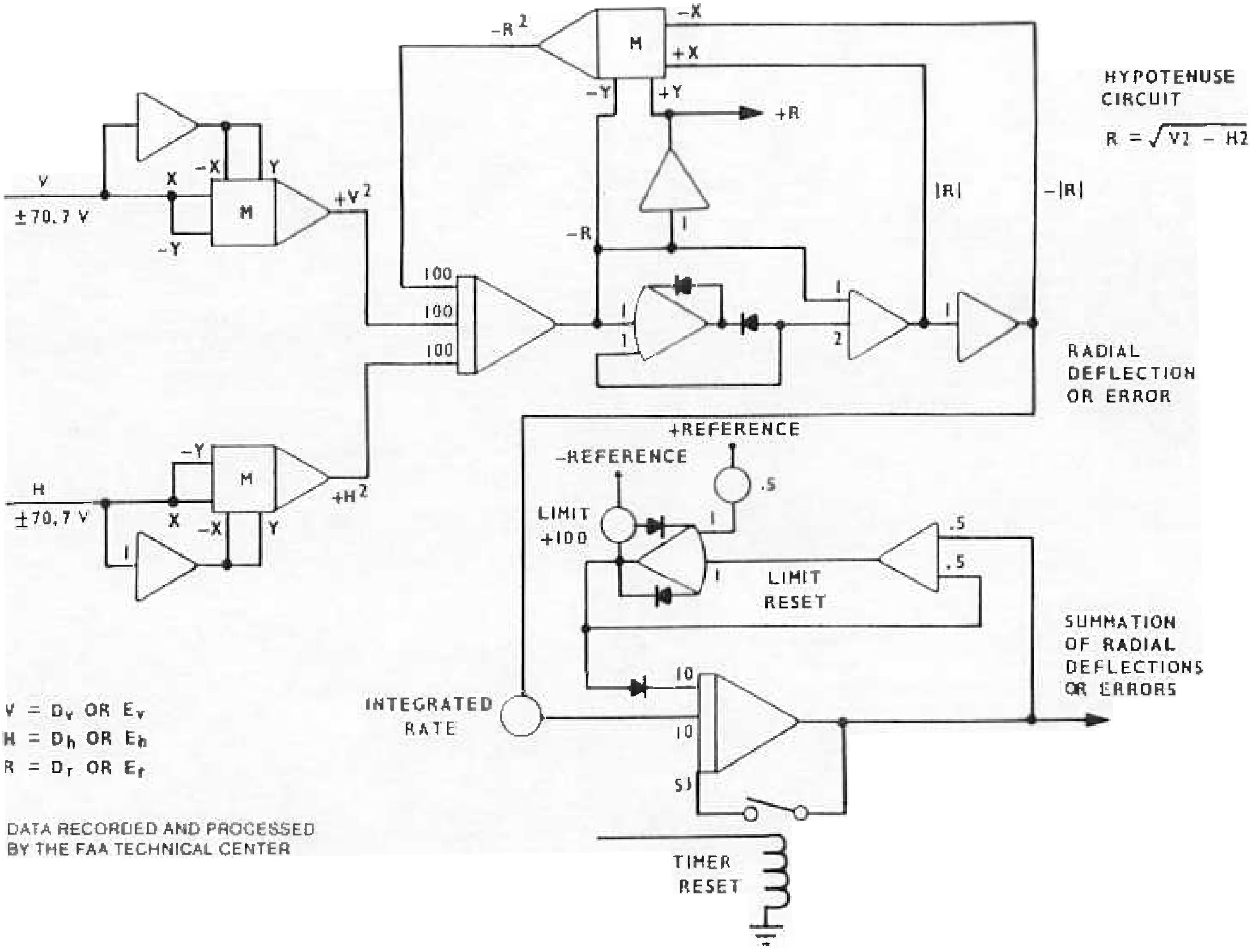


FIGURE C-5. TRACKING TASK ANALOG DIAGRAM



DATA RECORDED AND PROCESSED BY THE FAA TECHNICAL CENTER

C-6. TRACKING TASK ANALOG DIAGRAM (TWO IDENTICAL