

DOT/FAA/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

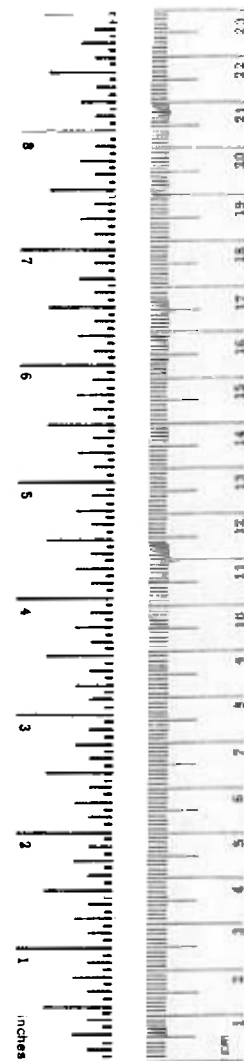
The authors gratefully acknowledge the technical assistance of Douglas Elliott, of the Systems Simulation Branch at the Technical Center, in setting up the analog computer for the experiment described herein. Mr. Elliott also contributed the section found in appendix C of this report entitled "Tracking 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	Symbol
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	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
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

Symbol	When You Know	Multiply by	To Find	Symbol
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				
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	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

EXECUTIVE SUMMARY	
INTRODUCTION	1
Purpose	1
Background	1
Behavioral Measures	2
METHODOLOGY	5
Critical Tracking Task	5
Test Subjects	6
Equipment and Test Procedure	6
DATA ANALYSIS	11
Task Difficulty Versus Effort Rating	11
Secondary Variables	24
Questionnaire Data Analysis	27
RESULTS SUMMARY	30
CONCLUSIONS	31
RECOMMENDATIONS	32
REFERENCES	32
APPENDICES	
A Workload Evaluation Participant Briefing	
B Workload Evaluation Questionnaires	
C Tracking Task	

LIST OF ILLUSTRATIONS

Figure		Page
	Ten-Point Workload Rating Scale	
2	Laboratory Setup	10
3	Mean Effort Rating Versus Difficulty Level for Nonpilot and Pilot Groups	19
4	Histograms of Effort Rating Versus Difficulty Level for Nonpilot and Pilot Groups	20
5	Rating Delay Versus Difficulty Level for Nonpilot and Pilot Groups	22
6	Mean Critical Lambda Versus Time Into Experiment for Nonpilot and Pilot Groups	22
7	Mean Operating Lambda Versus Difficulty Level for Nonpilot and Pilot Groups	23
8	Mean Critical Lambda Versus Difficulty Level for Nonpilot and Pilot Groups	23
9	Mean Natural Logarithm of Tracking Error Versus Difficulty Level for Nonpilot and Pilot Groups	24

LIST OF TABLES

Table		Page
1	Variables Recorded on the Chart	7
2	Research Design	9
3	Identification of Variates	12
4	Definition of Measures Used in Data Analysis	13
5	Mean Values of Measures Used in Data Analysis	14
6	Summary of Overall Repeated Measures Analysis of Variance for Eight Measures: Significant F Ratios	15
7	Simple Effects One-Way Repeated Measures Analysis of Variance: Significant F Ratios	16
8	Simple Effects Repeated Measures Analysis of Variance, Multiple Comparison Tests Among Means	17
9	Relationship of Effort Rating to Stick Deflection and Ln Tracking Error	26
10	Discriminant Function Analysis to Separate Nonpilots from Pilots	26
11	Workload Questionnaire Results	28
12	Questionnaire Intervariable Correlations	29

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:

$$\text{Total Workload} = \text{Physical Workload} + \text{Mental Workload}$$

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 λ 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).

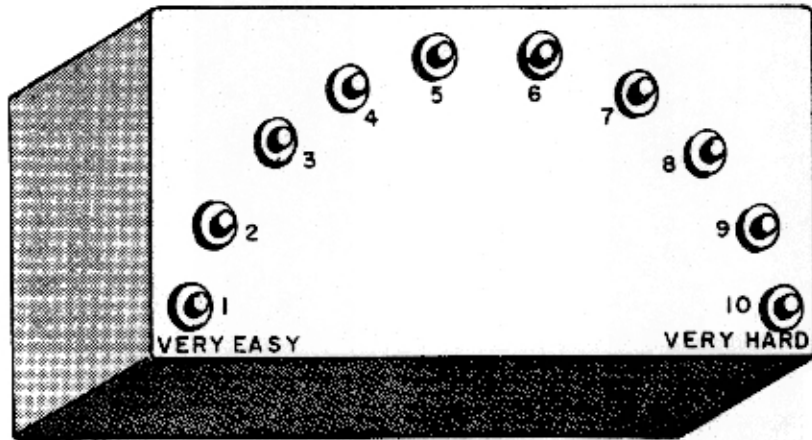
Once it is determined whether or not the workload measurement technique is effective in measuring a subject's assessment of how hard he/she is working at a given time, the scale can be used in simulation studies and actual flight tests.

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.