A SENSITIVITY/INTRUSION COMPARISON OF MENTAL WORKLOAD ESTIMATION TECHNIQUES USING A SIMULATED FLIGHT TASK EMPHASIZING PERCEPTUAL PILOTING BEHAVIORS

by

John Gordon Casali

Dissertation submitted to the
Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in

Industrial Engineering and Operations Research

APPROVED:

Dr. W. W. Wierwille, Chairman

Dr. R. D. Dryden

Dr. E. S. Geller

Dr. D. L. Price

Dr. D. R. Sebolt

July, 1982
Blacksburg, Virginia

ACKNOWLEDGEMENTS

The author expresses special thanks to his major advisor, Dr. Walter W. Wierwille, whose untiring guidance, assistance, and enthusiasm made this dissertation possible. Committee members Dr. Robert D. Dryden, Dr. E. Scott Geller, Dr. Dennis L. Price, and Dr. Don R. Sebolt also provided much needed advice and support throughout the course of the research described herein, and are sincerely thanked for their contributions.

This research was sponsored by NASA, Ames Research Center located at Moffett Field, California. Thanks are due to Mrs. Sandra Hart who served as grant technical monitor. Virginia Polytechnic Institute and State University also contributed to this research though cost sharing and through the award of a Cunningham Dissertation Fellowship to the author.

Several individuals assisted with certain aspects of the research. Dr. Klaus H. Hinkleman, Professor of Statistics, is thanked for his advice on the parametric analyses. Mr. Wolfgang Durheimer is thanked for the role he played in serving as an associate experimenter for collection of secondary task data and subsequent data reduction. Appreciation is also due to Mr. Richard E. Cordes for his aid in fabricating the respiration measurement equipment.

Very special thanks are given to the author's parents, Mr. and Mrs. J. Tony Casali, for their steadfast encouragement and support throughout the educational career of their son.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

INTRODUCTION

Background

In modern military and civilian aviation, a prime consideration is that aircrew personnel should not be overburdened. Flight-deck instrumentation, communications procedures, and aircraft dynamics are progressively increasing in complexity, demanding more attention, information processing, and subsequent action on the part of today's pilot. Such demands, coupled with higher airspeeds and more dense traffic conditions, increase the "mental workload" level of the pilot.

The criticality of the mental workload requirements placed on the pilot and aircrew is quite variable. If mental requirements are excessive there may be a measurable degradation in the performance of simple tasks associated with flying, such as instrument monitoring. Often, more critical tasks will receive the pilot's immediate attention and other tasks will be time-shared or even ignored. Mental <u>overload</u> may further result in significant pilot errors in aircraft control, possibly culminating in an accident. Overload may occur instantaneously or it may be sustained. In any case, overwhelming the pilot/aircrew with time-constrained information-handling responsibilities is certainly undesirable, both from operational efficiency and safety standpoints.

In the thrust of improving the pilot-aircraft interface, significant technological advances in avionics ultimately have found their way into cockpits. Generally, these efforts have been directed toward enhancing pilot information gathering and processing. Additions such as automated aircraft control devices, stability augmentation systems, computerized navigational aids, predictive displays, integrated flight instrument cathode ray tube displays, and even additional aircrew members have and will continue to be included in aircraft cockpits. The necessity of these additions is in part a result of increased aircraft capabilities, which in turn often increases the demands placed on the pilot. Despite these recent advances and impressive capabilities, the ultimate success of any aircraft or flight mission is predicated on a single common denominator: the human pilot. For this reason, the need for empirical assessment of pilot mental workload is particularly cogent.

During the design, development, and evaluation of a new aircraft, it is important for the cockpit design engineer to work for an optimum pilot-aircraft interface while also striving to maintain mental workload requirements at a level within the pilot's capabilities. It is preferable to maintain workload at a level well-within these capabilities to insure that reserve capacity is available for unanticipated workload "peaks" or emergency situations. In addition to new aircraft, it is also important to assess the workload implications of retrofitting new equipment or introducing new procedures in older aircraft. With the task of bounding pilot mental workload in research and development efforts, the accurate measurement of workload then becomes essential.

Mental Workload Measurement

Numerous methods, test instruments, and analytical techniques have been purported as useful in the assessment of operator mental workload in flight-related and other tasks (Wierwille and Williges, 1978). In some procedures the workload level is measured on an absolute scale, while in other cases, several situations of interest are compared against each other in terms of relative workload requirements. Often, it is desirable to concentrate mental workload assessment procedures on particular tasks of interest to a researcher or designer. These tasks are usually referred to as "primary tasks" in an experimental paradigm.

The utility and viability of any mental workload estimation technique is largely dependent on two factors: sensitivity and intrusion.

Sensitivity. As of this writing, the ability of state-of-the-art mental workload estimation techniques to measure and quantify workload on an absolute ratio scale of say, zero to 100 percent loading, has not been demonstrated. Prior to such an endeavor, it appears prudent that a variety of techniques should be compared as to their sensitivity to changes in mental workload level, as in the present study. With this objective in mind, a sensitive workload estimation technique can be defined as one which reliably discriminates between significant differences in operator mental loading requirements of a given task. Significant changes are those which lead to degradations in aircraft control, those which reduce pilot ability in accomplishing mission objectives, and those which increase the pilot's attentional demands to a degree that time-sharing or omission of tasks occurs.

Intrusion. Intrusion refers to an undesirable, artificial variance in primary task performance, solely due to the concurrent use of a workload estimation procedure or associated equipment. Intrusion is objectionable for two reasons. First, its presence contaminates workload assessment because primary task performance is altered and the indicated or measured workload level may not be representative of task requirements alone. Highly intrusive techniques may also degrade aircraft controllability and create safety hazards. The occurrence of intrusion in workload measurement is well-known. Specific instances are discussed in the following section.

OVERVIEW OF MENTAL WORKLOAD ESTIMATION LITERATURE

Existing Literature Reviews on Mental Workload Estimation

A plethora of extremely diverse research literature concerning mental workload and associated estimation techniques has surfaced in the past two decades. Numerous documents, in the form of literature reviews, annotated bibliographies, and categorization schemes are presently available to the workload researcher. Several review articles which are particularly noteworthy are discussed below in chronological order. These review articles are those which survey the general area of mental workload estimation. Reviews which sample a specific workload subcategory are covered later.

Reising (1972) provides a primer of workload estimation techniques intended for use by engineers involved in pilot workload problems. This early review concentrates on the definition of mental workload and stress and on the dichotomization of workload measures as either physiological or psychological. Another early, but more complete categorization of workload research literature appears in Jahns (1973a;1973b). Jahns organized pre-1973 mental workload research into four categories: information processing studies, operator activation level studies, time and motion studies, and equipment design studies. Based on the review, Jahns concluded that there exist three critical, functional aspects of mental workload, including input load, operator effort, and operator work performance. Gerathewohl (1976) provides a brief summary of the measurement techniques for perceptual and mental workload used specifically for pilots. and aircrews. In his report, techniques are classified as physiological, psychological, or operational. Also limited to workload methodologies applicable to aircrew systems is the selected annotated bibliography of Schiflett (1976).

Again in 1976, one of the first critical surveys of workload assessment appeared. Gartner and Murphy (1976), performed a conceptual analysis of workload and fatigue in pilots and further classified methodologies as either task demand analyses, pilot effort assessments based on task performance, psychophysiological indicants of effort, or subjective opinion assessments. Considerations and limitations concerning the utilization of various methodologies are also specified in the report. Rolfe (1976) reviewed assessment techniques for human response in vehicular control tasks. Categories of response assessment identified by Rolfe include performance requirement, response level, response effectiveness, environmental impairment, and between-operator comparison. Other measures mentioned include observational techniques, subjective assessment, loading tasks, and physiological measurement.

More recently, Butterbaugh (1978) reviewed pilot/aircrew mental workload measurement and prediction methods. The adequacy of various workload estimation techniques for measuring mission-derived workload and for individual task application within the mission was examined. Butterbaugh concluded that the workload measurement/prediction technology (circa 1978) was not sufficient for whole-mission assessment but was adequate for part-task requirements.

Mental workload measurement has received recent attention as the subject of several AGARD (Advisory Group for Aerospace Research and Development) Symposia reports. In an AGARD report edited by Roscoe (1978a), individual attention is devoted to various categories of mental workload methods. Within this report, Roscoe (1978b) specifically discusses various physiological methods, Ellis (1978) addresses subjective opinion methods, and Chiles (1978) reviews various objective techniques including laboratory, analytic, synthetic, vehicular simulation, and in-flight methods (see also Chiles and Alluisi, 1979). A more recent AGARD report, edited by Hartman and McKenzie (1979), includes 19 papers by various researchers covering a wide variety of workload assessment techniques. A useful collection of position papers on mental workload also appears in Moray (1979). It should be noted that individual papers from each of these reports are cited later in this overview, under the appropriate workload estimation technique subheading.

Perhaps the most definitive textual review and classification system concerning mental workload estimation techniques appears in Wierwille and Williges (1978). This document reviews over 400 workload-related references within the framework of a two-dimensional categorization scheme. Two mutually-exclusive subsets of the magnanimous results of this technical report have been published in the open literature as Williges and Wierwille (1979) and Wierwille (1979). More recently, the literature review portion of the technical report has been updated by Wierwille and Williges (1980) in an annotated bibliography addressing mental workload measurement. This bibliography contains more than 600 citations categorized by the two-dimensional scheme.

Because the Wierwille and Williges (1978;1980) manuscripts provide a useful basis for bounding the scope of the literature review in this report, a summary of the two-dimensional categorization scheme appears below. One dimension of the Wierwille-Williges classification scheme provides a means of classifying human behaviors that might be elicited during flight-related tasks. An inordinate amount of workload research has been performed in non-aviation settings; therefore, to draw accurate conclusions from this body of research, some means of relating general human operator behaviors to aircrew performance is essential. The universal operator behavior classification system outlined in Berliner, Angell, and Shearer (1964) was adopted by Wierwille and Williges (1978) for this purpose (Table 1).

This widely-used tri-level hierarchical system initially separates operator behaviors into four major processes: perceptual, mediational, communicative, and psychomotor. Next, the four processes are broken into six activities and finally into 47 mutually-exclusive specific behaviors. The orthogonality of the component behaviors insures a minimum of disagreement among investigators as to the particular behaviors inherent in a particular pilot/aircrew problem. It should be noted, however, that truly pure behaviors are rare in applied aviation settings. Therefore, it is common to speak of a particular task or mission as "emphasizing a particular behavior process," such as perception.

The Berliner et al. (1964) classification scheme may also be used to provide a framework for a boundary definition of mental workload. Previous efforts to define mental workload have resulted in considerable disagreement, as evidenced by the definitions and position papers reviewed in Reising (1972) and Moray (1979). Using the Berliner et

TABLE 1

Universal Operator Behavior Classification Scheme*

Processes	Activities	Specific Behaviors
1. Perceptual processes	1.1 Searching for and receiving information	1.1.1 Detects
		1.1.2 Inspects
		1.1.3 Observes
		1.1.4 Reads
		1.1.5 Receives
		1.1.6 Scans
		1.1.7 Surveys
	1.2 Identifying objects, actions, events	1.2.1 Discriminates
		1.2.2 Identifies
		1.2.3 Locates
2. Mediational processes	2.1 Information processing	2.1.1 Categorizes
		2.1.2 Calculates
		2.1.3 Codes
		2.1.4 Computes
		2.1.5 Interpolates
		2.1.6 Itemizes
		2.1.7 Tabulates
		2.1.8 Translates
	2.2 Problem solving and decision-making	2.2.1 Analyzes
		2.2.2 Calculates
		2.2.3 Chooses
		2.2.4 Compares
		2.2.5 Computes
		2.2.6 Estimates
		2.2.7 Plans
3. Communication processes		3.1 Advises
		3.2 Answers
		3.3 Communicates
		3.4 Directs
		3.5 Indicates
		3.6 Informs
		3.7 Instructs
		3.8 Requests
		3.9 Transmits
4. Motor processes	4.1 Simple/Discrete	4.1.1 Activates
		4.1.2 Closes
		4.1.3 Connects
		4.1.4 Disconnects
		4.1.5 Joins
		4.1.6 Moves
		4.1.7 Presses
		4.1.8 Sets
	4.2 Complex/Continuous	4.2.1 Adjusts
		4.2.2 Aligns
		4.2.3 Regulates
		4.2.4 Synchronizes
		4.2.5 Tracks

*Berliner, Angell, and Shearer, 1964

al. classification, mental workload is said to occur when an operator exhibits any of the specific universal operator behaviors shown in Table 1. Within this conceptual structure, experimental tasks of interest (or primary tasks) may be designed to elicit certain behaviors, therefore, enabling various workload estimation techniques to be investigated as to their relative sensitivity to changes in either perceptual, mediational, communicative, or psychomotor load.

Returning to the Wierwille and Williges (1978) workload classification system, the second dimension is oriented toward mental workload assessment techniques. The various methods for mental workload estimation are first logically separated into the four major categories of subjective opinion, spare mental capacity, primary task, and physiological (Table 2). These four major categories, which are discussed later in detail, are further subdivided into 28 individual techniques. Intricacies of the two-dimensional classification system are discussed in Wierwille, Williges, and Schiflett (1979). Furthermore, the two dimensions of operator behaviors and workload techniques are conjointly applied to a stepwise algorithm, enabling a user to select a particular workload method for a specific aircrew problem (Wierwille et al., 1979). Of course, this selection procedure is based solely on the results of the

TABLE 2

Mental Workload Estimation Techniques Classification Scheme*

- 1. Subjective Opinion
  - 1.1 Rating Scales
  - 1.2 Interviews and Questionnaires
- 2. Spare Mental Capacity
  - 2.1 Task Analytic
    - 2.1.1 Task Component, Time Summation
    - 2.1.2 Information-Theoretic
  - 2.2 Secondary Task
    - 2.2.1 Nonadaptive, Arithmetic/Logic
    - 2.2.2 Nonadaptive, Tracking
    - 2.2.3 Time Estimation
    - 2.2.4 Adaptive, Arithmetic/Logic
    - 2.2.5 Adaptive, Tracking
  - 2.3 Occlusion
- 3. Primary Task
  - 3.1 Single Measures
  - 3.2 Multiple Measures
  - 3.3 Math Modeling
- 4. Physiological Measures
  - 4.1 Single Measures
    - 4.1.1 FFF
    - 4.1.2 GSR
    - 4.1.3 EKG
    - 4.1.4 EMG
    - 4.1.5 EEG
    - 4.1.6 ECP
    - 4.1.7 Eye and Eyelid Movement
    - 4.1.8 Pupillary Dilation
    - 4.1.9 Muscle Tension, Tremor
    - 4.1.10 Heart Rate, Heart Rate Variability, Blood Pressure
    - 4.1.11 Breathing Analysis
    - 4.1.12 Body Fluid Analysis
    - 4.1.13 Handwriting Analysis
  - 4.2 Combined Physiological Measures
  - 4.3 Speech Pattern Analysis

*Wierwille and Williges, 1978

literature review and does not imply that the selection of a technique is empirically derived for each aircrew situation.

The next section of this literature review addresses individual workload estimation techniques selected for investigation in the present study. In the proposal stage, a number of techniques were considered for investigation as candidate measures. In the final selection of techniques for this report, a minimum of two techniques from each of the four major classes (subjective opinion, spare mental capacity, primary task, and physiological) were chosen. Selected techniques are shown in Table 3. Selection of these particular estimation techniques was in large part based on their apparent promise for use in pilot/aircrew problems. Techniques identical to or closely akin to those in Table 3 have previously demonstrated some sensitivity to changes in mental loading requirements in various tasks relevant to the perceptual requirements of flight. Also, all techniques selected are feasible for implementation in either simulated or full-scale flight in terms of hardware requirements. The number of selected techniques, eight, was constrained only by the pilot resources in the experiment locality. Nearly all local licensed pilots were contacted and/or participated in the study described herein.

Literature addressing each major category of workload estimation technique (according to the Wierwille and

TABLE 3

Mental Workload Estimation Techniques Investigated in the Present Study

Subjective Opinion	Modified Cooper-Harper Scale
	Multi-Descriptor Scale
Spare Mental Capacity (Secondary Task)	Time Estimation
	Tapping Regularity
Primary Task	Danger Condition Reaction Time
	Control Movements
Physiological	Heart Pulse Rate Variability
	Respiration Rate

Williges (1978) classification) and each specific candidate technique is reviewed below. Where possible, the literature concerning a specific technique is presented in chronological order.

Subjective Opinion

In flight test and evaluation problems, many mental workload assessment procedures often entail some form of subjective opinion. Pilot/aircrew opinions concerning a particular aircraft design, operational procedure, or instruction are frequently obtained via interviews, structured and unstructured questionnaires, and psychometrically-derived rating scales. Due to its quantitative and systematic nature, the rating scale approach is most common in controlled research efforts. Questionnaires and interviews are generally considered as those procedures not based on psychometric scaling considerations, and as such, are often employed for qualitative ratings of system aspects.

Opinion ratings offer several advantages for the workload researcher (Wierwille and Williges, 1978). First, pilot acceptance of opinion ratings is generally quite high. It is sometimes evident that the rater's ego is bolstered by the fact that s/he has a personalized input into the design of a system. If so, the subject may perceive a strong sense of internal locus of control. Also, if instructions are carefully designed, rating techniques are often easily understandable. Ease of implementation, simplicity of test instruments and scoring, and lack of intrusion are other advantages. However, as noted by Wierwille and Williges (1978), awareness on the part of an aircrew member that an opinion rating will be obtained may somewhat influence or intrude upon his/her behavior on the primary task. Even so, this modification of behavior patterns may actually improve the accuracy of the ratings because of the increase in attention to the task at hand.

Despite these cogent advantages, opinion ratings do have some shortcomings (Wierwille and Williges, 1978). The most prominent of these is the fact that the majority of rating scales previously used in workload research have not been developed according to rigorous psychometric principles. Without such development, the reliability and validity of a rating scale instrument are suspect. Another problem centers around the appropriate application of statistical procedures to scores derived from rating scales. Many rating scales do not provide an interval level of measurement and as such, the application of parametric statistical procedures to scale data must be carefully qualified. Several authors have outlined procedures for scale construction and psychometric scaling to avoid most pitfalls (e.g., Edwards, 1957; Nunnally, 1967). Others (Dyer, Matthews, Wright, and Yardawitch, 1976; Geer, 1977) have presented recommendations concerning the application of subjective opinion techniques in applied settings.

Another problem associated with opinion ratings is that the rater may adapt to a particular system, exhibiting improved task peformance over time, and supply ratings which are overly high or low. In some cases, pilots may not be able to accurately differentiate between mental and physical workload. A particularly difficult problem exists in situations where aircrew members may simply not be aware of the incident workload level. In other words, actual task loading may differ substantially from perceived task loading. Other problems include variance in ratings due to experience, fatigue, emotion, feelings of inadequacy, and apathy.

Application of subjective opinion in workload research. Initially, two workload related, but not flight-related, investigations which have demonstrated the successful use of opinion techniques as indicants of task load will be discussed. Philipp, Reiche, and Kirchner (1971) employed a zero-to-nine point Likert-type scale to obtain air traffic controllers' subjective ratings on two dimensions of workload: stress of time and difficulty of the control task. Nonparametric rank correlation was applied to the data, and in several instances objective measures of task load correlated significantly with the subjective ratings.

Hicks and Wierwille (1979) devised an ordinal rating scale directed especially toward operator mental workload in a driving task. The rating scale demonstrated significant sensitivity to changes in vehicle handling (influencing driving task difficulty), and little intrusion on driving task performance.

For the less-structured opinion techniques of interview and questionnaire, several studies on flight-related issues warrant mention. Soliday (1965), after presenting a moving-base, G-seat simulation of low altitude high-speed flight, employed a question and answer debriefing session to obtain pilots' reactions to the flight task. Cantrell and Hartman (1967) utilized self-report, 30-minute interval, logs of stress-inducing activities to obtain data on typical military transport pilot workload levels. A similar technique was applied by Soutendam (1977) to air traffic control tower operators. Steininger (1977) used a questionnaire consisting of 82 seven-step rating scale items, coupled with a semi-structured interview to assess pilot workload in short haul jet transport aircraft. Cockpit and instrument layout, handling qualities, and overall system operation were each investigated as to pilot workload involved. Recommendations were made in regard to questionnaire and interview strategy. Rohmert (1977) provides additional quidelines for workload-related questionnaire development.

Structured rating scales have also been successfully employed in both simulator and in-flight research. In 1971, Spyker, Stackhouse, Khalafalla, and McLane used a non-continuous, forced-response rating scale to measure subjects' perceived task workload in conjunction with several physiological measures. The primary task load was varied by changing aircraft pitch dynamics and wind gust disturbances in a simulated aircraft tracking task. While no rigorous statistical tests were performed, the authors reported that subjective evaluations of task difficulty increased as the task indeed became more difficult. Baker and Intano (1974) proposed the use of a post-flight structured questionnaire, specified in their report, for investigation of the effects on helicopter pilot workload of yaw axis augmentation dynamics. No follow-on study appears to have been performed. In 1975, Crabtree used a bipolar-adjective "opinionnaire" to assess the relative workload involved during six different modes of aircraft control in simulated instrument flight rule (IFR) approaches. Objective workload data was also recorded and subjected to an analysis of variance to determine sensitivity. Unfortunately, no statistical analysis was performed on the subjective data.

Using an actual aircraft, Geiselhart, Schiffler, and Ivey (1976) conducted flight tests to investigate the feasibility of reducing aircrew size in a KC-135 aircraft from four to three members. A post-flight multi-item rating scale/questionnaire produced qualitative responses indicating that on several types of missions, the reduced aircrew size resulted in excessively high workload levels. In an investigation of simulated landing approaches with varying degrees of failure of the autopilot system, Johannsen, Pfendler, and Stein (1976) used a five-point graphic rating scale as an indicant of perceived workload level. Qualitative comparisons revealed that the workload ratings were sensitive to increases in workload accompanying autopilot failures.

In a full-scale helicopter (Alouette Model III) study, Smit and Wewerinke (1978) assessed pilot workload during instrument hover and instrument navigation (tracking) tasks. Performance, physiological, and subjective opinion measures of workload were obtained. From a multivariate analysis of the data, the authors concluded that the three rating scales for 1) degree of effort expended, 2) controllability and precision, and 3) pilot demand, discriminated between loading levels on the flight tasks. All scales were ten-step ordinal scales, with objective descriptors on the controllability and precision and pilot demands scales. Another multi-measure workload investigation was designed to determine the minimum aircrew size and minimum crew systems required for accomplishment of a tactical transport mission in an Advanced Medium Short Takeoff or Landing (STOL) Transport (Madero, Sexton, Gunning, and Moss, 1979). A structured questionnaire/rating scale was used in conjunction with flight performance measures in a simulator to determine required crew and system size and complexity. Other investigators who have used rating scales as a complement to other workload measures in flight simulator studies include Murphy and Gurman (1972), and Kreifeldt, Parkin, and Rothschild, (1976). Additional full-scale flight studies utilizing rating scales in this regard include Stackhouse (1973), Lebacqz and Aiken (1975), and Helm (1975,1976).

As noted by Williges and Wierwille (1979), in most of the rating scale applications above, the scale was situation-specific, nontransferable, and had not been empirically demonstrated as a reliable nor valid measure of pilot mental workload.

In contrast to the conglomeration of rating scales reviewed above, two systematically-derived rating scales have seen considerable application in aircraft development and multi-situation pilot/aircrew workload assessment. These ratings scales are the Cooper-Harper (1969) scale and the recent Workload-Compensation-Interference/Technical Effectiveness (WCI/TE) scale, as described in Donnell

(1979). Both scales have undergone rigorous testing and some psychometric development. Both scales also provide a quantitative assessment of pilot workload for a specific task. However, due to the specificity of terms and definitions used in the scale axes, the WCI/TE rating scale (Figure 1) is not readily applicable to primary tasks which are perceptual in nature. Therefore, the WCI/TE scale was not selected for use in the present investigation. Nevertheless, because the WCI/TE scale represents a prominent, current effort in mental workload measurement, it will be briefly reviewed here.

Workload-Compensation-Interference/Technical Effectiveness (WCI/TE) scale. The WCI/TE scale consists of two labeled scale axes or dimensions: workload, compensation, and interference required and inherent in the system and perceived technical effectiveness of the system. Ratings are independently performed on each axis, providing a separate ordinal, categorical rating for workload-compensation-inteference and technical effectiveness. This pair of ratings then specifies the appropriate cell in the two-dimensional square matrix. Ratings on the non-interval WCI scales are combined through conjoint measurement procedures to provide an approximate interval metric of flight task/system operability (Donnell and O'Connor, 1978; Donnell, 1979). Each cell of the rating matrix has

CIRCLE THE APPROPRIATE RESPONSE FOR EACH DIMENSION.

DEFINITIONS:

Workload. The integrated physical and mental effort required to perform a specified piloting task.

Compensation. The measure of additional pilot effort and attention required to maintain a given level of performance in the face of deficient vehicle characteristics.

Interference. The extent to which attention devoted to task decreases the attention available for other task(s).

Technical Effectiveness. The utility of a system as determined by its design, hardware and software in accomplishing a task and accomplishing desired performance.

TECHNICAL EFFECTIVENESS	1	2	3	4 Column	Row
Multiple Tasks Integrated	(48)	(26)	(10)	(0)	4
Design Enhances Specific Task Accomplishment	(57)	(35)	(19)	(9)	3
Adequate Performance Achievable; Design Sufficient to Specific Task	(68)	(46)	(30)	(20)	2
Inadequate performance Due to Technical Design	(100)	(88)	(62)	(52)	1
	Workload Extreme; Compensation Extreme; Interference Extreme	Workload High; Compensation High; Interference High	Workload Moderate; Compensation Moderate; Interference Moderate	Workload Low; Compensation Low; Interference Low	

WORKLOAD/COMPENSATION/INTERFERENCE
(Mental and Physical)

Figure 1. Workload-Compensation-Interference/Technical Effectiveness rating matrix, with scale values and term definitions shown (Donnell, 1979).

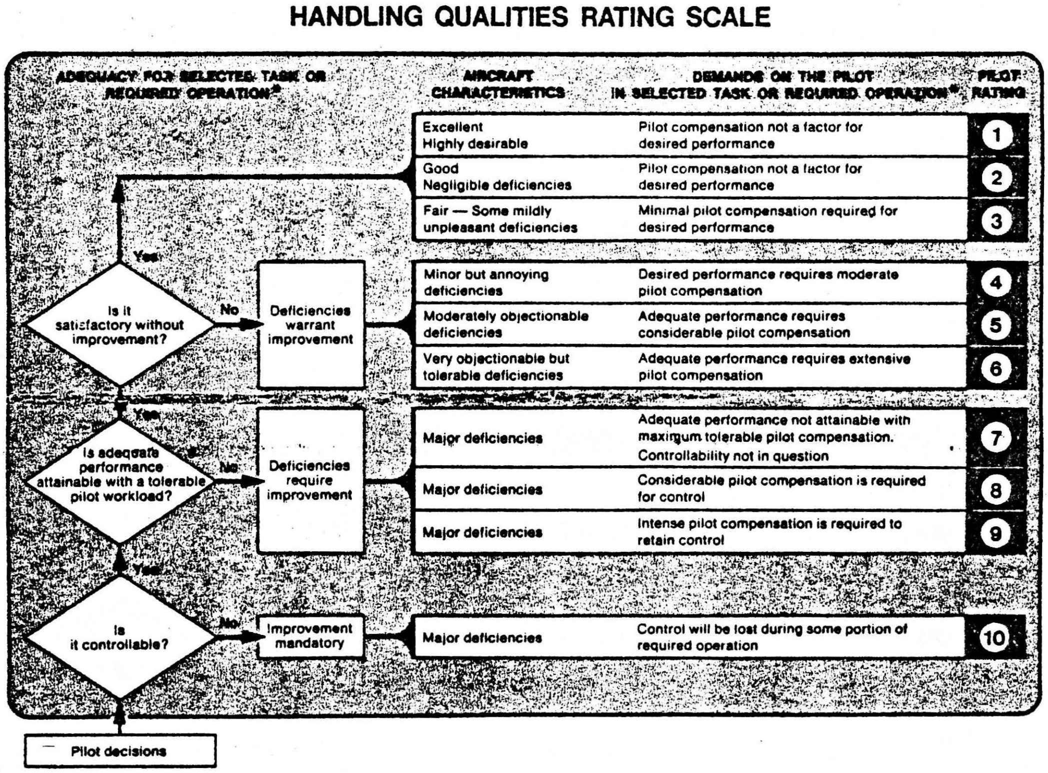
associated with it a predetermined score, shown in parentheses in Figure 1, for quantization. These scores were obtained by the Delta scaling method, outlined in Appendix B of Donnell (1979).

Most WCI/TE related studies have focused on the development and validation of the scale itself. In 1977, O'Connor and Buede presented a one-dimensional adjectival rating scale for obtaining pilots ratings of workload required in various F18 aircraft flight tests. This preliminary scale was ordinal at best and O'Connor and Buede recognized the desirability of obtaining two separate ratings of the F18 system, one dealing with pilot workload required and one addressing the technical effectiveness of the system. Donnell and O'Connor (1978) devised a two-dimensional rating matrix for combining pilot workload and technical effectiveness scores into a single interval metric. Initial scaling was performed on data obtained from a group of pilots who ranked the cells of the matrix from best to worst on a dimension of flight task/system operability. Donnell and O'Connor (1978) also had three pilots rate the F18 aircraft system on 258 task conditions using an earlier version of the WCI/TE rating matrix referred to as the pilot workload/technical effectiveness (PW/TE) rating matrix. Applying discrimination analysis procedures to the data, the authors found that PW/TE scores were valuable for discriminating between tactical and nontactical tasks/systems, for identifying critical task/system deficiencies, and for identifying system strengths.

The next major step in the scale development was reported by Donnell (1979). The technical effectiveness dimension of the PW/TE scale remained essentially unchanged; however, Donnell modified the pilot workload dimension by deleting all terms which could imply that the rater should make a value judgement as to aircraft control or workload level, such as "excessive," "major," "minor," and "as anticipated." Instead, the pilot workload factor was renamed as workload-compensation-interference (WCI) and four levels of WCI were used: "low," "moderate," "high," and "extreme." These alterations reflected an intent to make the rating matrix more applicable to human-machine systems other than aircraft. Finally, the Delta method of additive conjoint scaling was applied to the modified scale (see Donnell, 1979). Donnell applied the new WCI/TE rating matrix in an assessment of the mission operability of an A7-Echo single-seat aircraft on 202 flight tasks. Eight experienced pilots participated in the study. The overall mission operability score was 68 out of a perfect score of 100, as indicated by aggregated WCI/TE ratings. Via further discrimination analyses, the operability deficit (32) could then be reduced by correcting prominent deficiencies where pilot workload was high and mission effectiveness was low.

The utility of the WCI/TE rating matrix in both the A7-Echo and F18 aircraft calculations is therefore quite evident.

Cooper-Harper scale. While the WCI/TE scale is gaining popularity for use in specific flight-task situations, the Cooper-Harper (1969) scale has a considerably larger following in the workload literature. The standard form of the scale is shown in Figure 2. Following a series of iterative approximations, this final form of the scale was produced. A detailed description of the scale's development is presented in Cooper and Harper (1969). The original intention of the scale was to provide a vehicle for pilot input to the design and evaluation of aircraft handling dynamics. In use, a set of instructions is presented to the pilot, describing the proper procedure for making a rating and defining the terms used within the scale. Ratings on the scale are performed by following decision tree logic to an ultimate numerical rating of aircraft control performance, based on the amount of pilot effort and compensation required. Therefore, pilot workload and aircraft performance interact in the rating process. In its original form, the Cooper-Harper scale is somewhat specific to the assessment of handling qualities. Because this study was directed toward pilot mental workload, the Cooper-Harper scale was altered slightly for use in the experiment described herein. These modifications are described in the forthcoming apparatus section.


HANDLING QUALITIES RATING SCALE
ADEQUACY FOR SELECTED TASK OR REQUIRED OPERATION*
Yes
Is it satisfactory without improvement?
No
Deficiencies warrant improvement
Yes
Is adequate performance attainable with a tolerable pilot workload?
No
Deficiencies require improvement
Yes
Is it controllable?
No
Improvement mandatory
Pilot decisions

AIRCRAFT CHARACTERISTICS	DEMANDS ON THE PILOT IN SELECTED TASK OR REQUIRED OPERATION*	PILOT RATING
Excellent Highly desirable	Pilot compensation not a factor for desired performance	1
Good Negligible deficiencies	Pilot compensation not a factor for desired performance	2
Fair — Some mildly unpleasant deficiencies	Minimal pilot compensation required for desired performance	3
Minor but annoying deficiencies	Desired performance requires moderate pilot compensation	4
Moderately objectionable deficiencies	Adequate performance requires considerable pilot compensation	5
Very objectionable but tolerable deficiencies	Adequate performance requires extensive pilot compensation	6
Major deficiencies	Adequate performance not attainable with maximum tolerable pilot compensation. Controllability not in question	7
Major deficiencies	Considerable pilot compensation is required for control	8
Major deficiencies	Intense pilot compensation is required to retain control	9
Major deficiencies	Control will be lost during some portion of required operation	10

Figure 2. Cooper-Harper rating scale (Cooper and Harper, 1969).

Despite the substantial effort vested in its development and initial testing, the Cooper-Harper scale has not been subjected to reliability and validity testing. Nevertheless, since its introduction, the scale has seen notable popularity for use in both simulator and full-scale workload-related studies.

For instance, Schultz, Newell, and Whitbeck (1970) employed the Cooper-Harper scale for pilot rating of his/her ability to maintain acceptable localizer and glideslope accuracy in performing a simulated ILS (instrument landing system) task. The objective of the study was to assess the relationship between actual pilot-aircraft performance and subjective rating of performance. The loading task emphasized the use of psychomotor flying skills; load was varied by altering the simulated airplane dynamics. An analysis of variance revealed that the Cooper-Harper ratings obtained were sensitive to changes in both short-period frequency and damping ratio parameters of the simulated aircraft dynamics. These results should be qualified somewhat in that application of parametric analysis of variance to ordinal data, such as that obtained with the Cooper-Harper scale, violates the parametric statistics assumption of equal interval data.

In another study, Dick, Brown, and Bailey (1976) used Cooper-Harper ratings and other objective measures to compare workload levels across several landing conditions in a Boeing 737 flight simulator. Workload was varied by changes in turbulence and instrumentation. A multiple linear regression analysis demonstrated that the number of control movements (aileron, elevator, and thrust) accounted for 73% of the variance in Cooper-Harper ratings. Similar results were obtained by Waller (1976), who varied pilot workload level in simulated ILS approaches by turbulence changes and by the presence or absence of flight director glide slope and localizer command bars. Waller's results revealed that oculometer-obtained instrument scanning measures could predict pilot rating on the Cooper-Harper scale with a standard error of 0.8 unit. These results show that pilot ratings, properly acquired, may demonstrate strong positive relationships with other, more objective measures of flight task load.

Using a full-scale JUH-1H helicopter, Sanders, Burden, Simmons, Lees, and Kimball (1978) investigated perceptual-motor workload during a hover maneuver. Workload level was manipulated by the amount of cyclic and pedal control augmentation used in the helicopter. Cooper-Harper ratings monotonically decreased in value as the amount of control augmentation was decreased. Unfortunately, no statistical analyses were performed on the opinion data. In another full-scale helicopter study, O'Connell (1978) used a CH-47B helicopter to assess control/display parameters. Subject pilots flew several approaches, each terminating in hover and vertical landing, while using a head-up display with separate vertical and horizontal situations. The pilots' Cooper-Harper ratings varied with the particular segment of the approach being rated (e.g., final approach to hover point) and also with the ambient wind conditions.

Multi-Descriptor scale. The second opinion measure employed in the present study was the mental workload multi-descriptor scale. This rating scale was developed by Dr. W. W. Wierwille in the Vehicle Simulation Laboratory at Virginia Polytechnic Institute and State University (VPI&SU). As this study represents the first application of this rating scale, no prior literature could be cited. However, rating scales with similar properties to the multi-descriptor scale have been successfully used in several studies mentioned previously (e.g., Philipp, Reiche, and Kirchner, 1971, Crabtree, 1975; Smit and Wewerinke, 1978). A very recent paper by Hart, Childress, and Hauser (1982) also addresses the description of mental workload using multiple descriptor ratings. The multi-descriptor scale and associated instructions appear in the forthcoming apparatus section of this report.

Spare Mental Capacity

The second major category of workload assessment techniques is that of spare mental capacity. This large body of techniques is directed toward evaluation of an operator's spare, residual, or reserve mental capacity while performing a task. As introduced by Knowles (1963), the fundamental assumption of the spare capacity concept is that the human operator functions as a finite-capacity, single-channel, sampling device for assimilating and processing information relevant to the task at hand. Mental workload is measured by determining the operator's spare mental capacity at a particular point in time or over a period of time by a host of direct or indirect techniques. As operator loading on the primary task increases, the remaining capacity of the finite-bound input channel will decrease, and the measurement of this capacity then supposedly reflects primary task loading. A substantial amount of research has been directed toward various ramifications of spare mental capacity assumptions, including multi-channel versus single-channel processing, factors influencing upper channel capacity (e.g., emotion, stress, and fatigue), and bottlenecks in the input channel. This research is summarized in Kahneman (1973).

Spare mental capacity workload estimation techniques have been divided into three major types, according to the

Wierwille and Williges (1978) classification scheme. These types include task analytic, occlusion, and secondary task methods as shown in Table 2. Task analytic methods usually entail the use of mathematical modeling and theoretical methods for evaluating residual mental capacity. Senders (e.g., 1964, 1970) and Siegel and Wolfe (e.g., 1969) have authored much of the fundamental framework for task analytic research as thoroughly reviewed in Wierwille and Williges (1978). Occlusion is a time-sharing technique and usually involves blocking or occluding the operator's visual input during primary task performance. The occlusion is usually performed with a movable opaque visor over the operator's eyes or by blanking the visual display itself. Depending upon the procedure used, the blank or off time is varied either by the experimenter or by the operator. Workload on the primary task is assumed to increase as the number of "on" viewing samples required for adequate performance of the primary task increases per unit time. Wierwille and Williges (1978) also provide an overview of occlusion research.

All spare mental capacity techniques in this investigation were of the secondary task category; neither occlusion nor task analytic approaches were addressed. Task analytic methods, due to their modeling and theoretical predicates, are not readily adaptable to a multivariable comparison of techniques in a flight simulator. Occlusion is probably not feasible for use in actual flight because of the hazards inherent in blocking pilot's visual input. Also, in previous studies, occlusion has shown little promise as an indicant of mental workload. For example, Hicks and Wierwille (1979) found that the occlusion technique was not significantly sensitive to changes in loading in a simulated driving task. Furthermore, occlusion undesirably intruded on driving performance.

Secondary task review. The majority of spare mental capacity research applications have involved the use of secondary tasks. In any secondary task procedure, the operator first performs a primary or main task and in his/her spare time (away from the primary task), attends to an additional or secondary task. Performance on the secondary task is then an indirect measure of the attentional demand of the primary task. Secondary task performance theoretically should decrease as primary task requirements increase.

In certain types of secondary task procedures, scores on the secondary task are first obtained during a baseline period (secondary task presented alone) and later in conjunction with the primary task (primary and secondary task presented concurrently). In these particular procedures, percentage workload on the primary task of interest is usually computed using an equation of the following general form (Hicks and Wierwille, 1979):

$$W = 100(1 - Ssp/Ss)$$

where

W = primary task load

Ssp = score on secondary task concurrent with primary task

Ss = score on secondary task alone (baseline).

In other secondary task procedures, no baseline score is necessary. These procedures usually attempt to tap the relative workload of two or more primary task conditions, rather than rate absolute percentage workload. Often, variance or standard deviation scores are obtained for the secondary task and compared across primary task conditions.

An extensive amount of literature concerning the use of a multitude of diverse secondary tasks has been reviewed by Rolfe (1973), Levine, Ogden, and Eisner (1978), Chiles (1978), and Ogden, Levine, and Eisner (1979). Of these reviews, the Levine et al. (1978) overview and annotated bibliography is probably the most comprehensive. Throughout these reviews, one prominent assumption concerning the use of secondary tasks is evident. All nonadaptive secondary task procedures assume that the human operator performs the secondary task only when s/he has spare or free time available from the primary task. Implicit in this assumption is that the primary task remains fully attended throughout its duration, while secondary task performance varies according to available reserve mental capacity. A related postulation is that primary and secondary task performance are additive in nature and mutually noninterfering (Wierwille and Williges, 1978).

This noninterference assumption underscores the major problem of secondary task workload estimation--that of intrusion on the primary task. It is difficult to assume that a subject can consistently attend to the primary task, only performing the secondary task during free time. Subject instructions are critical in the effective application of secondary tasks. That is, subjects must be instructed not to timeshare between the primary and secondary task. If the primary task is not under full control, the secondary task should be ignored. The existence of secondary task intrusion is well-documented in the workload literature for a wide variety of tasks (e.g., Hall, Passey, and Meighan, 1965; Trumbo and Noble, 1972; Wickens, 1974; Hawkins, Church, and de Lemos, 1978; Wierwille and Gutmann, 1978; Hicks and Wierwille, 1979; Wickens and Tsang, 1979; Whitaker, 1979). The relative intrusion of two secondary task workload estimation techniques were assessed in the present study.

Recognizing the intertask interference problem, several researchers have devised cross-adaptive tasks to minimize the intrusion associated with secondary task evaluations (e.g., Kelley and Wargo, 1967). In cross-adaptive procedures, the difficulty of the secondary task is usually adjusted as a function of operator performance on the primary task. While cross-adaptive techniques have demonstrated limited success, the additional instrumentation required for the adaptive logic presently precludes their use in a flight environment. As such, both secondary task procedures investigated in this study were nonadaptive.

Next, each secondary task procedure is discussed individually.

Time estimation. Requiring subjects to estimate the passage of a specified duration of time has been a popular secondary task, as evidenced in the review of time estimation research by Hart, McPherson, and Loomis (1978). The use of time estimation tasks is particularly well-suited for pilot/aircrew use because of the many aspects of flying which require timing, such as holding patterns, ground speed calculations, and nonprecision approaches. As noted by Hart and Bird (1980), pilots tend to perform both conscious and unconscious mental estimates of time, regardless of the fact that most aircraft instrument panels have time clocks. Time estimation has other advantages which include ease of implementation, instrumentation, and scoring, ease of understanding, and unobtrusiveness.

Various methods which are available for time estimation tasks include verbal estimation, production, reproduction, and comparison, as described by Hart (1976). For flight tasks, the method of production is readily adaptable and most commonly used (Wierwille and Williges, 1978). It was used in the research described herein.

In the production method, the operator is presented with an onset signal to start the desired interval of time, the length of which is predesignated by the experimenter (usually 10 seconds). The operator then presses a key or switch to indicate the end of the estimated interval after the onset signal. Accuracy of the time estimates which are performed concurrently with the primary task, is assumed to be influenced by the mental workload required for performance of the primary task. The magnitude of this influence is believed to be a function of the strategy that the operator uses in making the estimates (Hart and McPherson, 1976). In low to moderate primary task load conditions, the operator may make a conscious, sustained effort to monitor the passage of time (an active estimate) while performing the primary task. As primary task load increases, causing an increase in concurrent activity, attention is diverted from the active mode of time estimation. This usually results in estimates which are generally too long and more variable. Further increases in primary task load eventually prevent the operator from making ongoing, active estimates of time due to the strong interference of concurrent activity. In this case, estimates are generally made by memory on the basis of the events that occurred during the period since the last onset signal. In this "retrospective" mode, overestimation of elapsed time usually occurs, resulting in estimates which are too short. More importantly, time estimates generally are expected to become more variable as concurrent task loading increases. Therefore, the standard deviation of time estimates is often used as a measure.

Although time estimation is a relatively new measure, it has been applied as a workload measure in several recent flight-related studies. For example, using a transport aircraft simulator, NASA-Ames Research Center (1975) manipulated pilot workload by varying cockpit instrumentation (presence or absence of flight path predictor and wind vector) and incident wind gusts (either zero or 15 miles per hour). Ten-second time estimates by the production method demonstrated sensitivity to changes in loading due to instrumentation and wind. Also using a flight simulator, Hart and McPherson (1976) found an increase in positive skewness of the distribution of production time estimates, a decrease in medium estimate length, and an increase in estimate variability as wind velocity was increased, and as a flight path predictor and graphic wind vector were added to the cockpit moving-map display.

In another application, time estimation was used as a dependent measure in demonstrating that synthesized cockpit warning messages require more attentional demand as the redundancy of the message is decreased (Hart and Simpson, 1976; Simpson and Hart, 1977). Subsequently, Hart, McPherson, Kreifeldt, and Wempe (1977) found that during various simulator approaches, pilots used an active time estimation mode less than a retrospective mode. This was indicative of high attentional demand during final approach. Actively reported estimates were consistently shorter than retrospective estimates. Similar results were found by Hart (1979), who reported that time estimation was a useful measure of pilot workload at various junctures of final approaches in crowded terminal areas.

In a simulation study directed at determination of avionics requirements for pilot-copilot transport aircraft, test aircrews flew four segments of an airdrop mission including takeoff, cruise, preparation for airdrop, and airdrop (Gunning, 1978). Productions of ten-second intervals were obtained during the cruise, preparation for airdrop, and airdrop segments of the flights. Mean time estimates produced by the pilots during flight reliably increased with respect to the mean baseline estimates. A large increase in variance, skew, and kurtosis of the estimates also occurred across the flight segments for pilots. Copilots' time estimates, however, did not exhibit significant differences from baseline to in-flight measurement. These results indicate that pilot's workload was higher than copilots' workload solely on the basis of the time estimation measure. However, these results were not consistent with subjective ratings obtained from the aircrew members themselves. Copilots rated their perceived workload higher than pilots. Gunning stressed the need for extended research to determine the efficacy of the time estimation procedure. Finally, McCauley, Kennedy, and Bittner (1979) found that time estimation day-to-day test-retest reliabilities were quite high (r = 0.80), attesting to the stability of the measure.

Tapping regularity. Michon (1964) proposed a nonverbal secondary task which he purported to be a particularly useful indicator of perceptual task load. In his task, commonly referred to as interval production, Michon required subjects to tap a foot pedal, press a key, or nod their heads in as regular a sequence of intervals as possible. Usually, subjects were instructed to tap or nod at a constant rate while concurrently performing a primary task of interest. Tapping regularity was expected to decrease as the difficulty of concurrent activity increased. A complex scoring procedure, taking into account the variance of tapping sequences, is explained in detail in Michon (1966). This procedure yields a measure of variability referred to as "pels," Michon's units of perceptual load. The necessary computational formulae are presented in the appendices section of this report.

Tapping regularity measurement offers several advantages. First, the required apparatus is minimal. A portable tapping apparatus was developed and discussed by Michon and van Doorne (1967). Also, Michon found no significant intrusion effect of tapping on primary task performance (e.g., Michon, 1964). Finally, the tapping task is easily learned by subjects and stabilization of tapping regularity quickly occurs during practice trials in baseline (secondary task alone) format (Michon, 1964).

Research evidence exists for the sensitivity of the tapping task to changes in mental load requirements of the primary task. Michon (1964) found that tapping regularity indices (pels) significantly discriminated between two levels of choice reaction task difficulty. He also demonstrated the sensitivity of the measure to changes in difficulty <u>across</u> different types of primary tasks, including maze performance, screw sorting, multiplication, letter detection, and Bourdon-test performance. Similar results were obtained by Michon in 1966. In the latter study, Michon also reported that tapping regularity results were in general agreement with task difficulty subjective ratings performed by the subjects.

In the realm of flight-related studies, it appears that only a single study by Johannsen, Pfendler, and Stein (1976) employed tapping regularity as a mental workload measure. In this experiment, pilots performed an instrument landing approach in a fixed-base STOL aircraft simulator. Workload was manipulated by varying the degree and type of autopilot failures. The results showed that tapping irregularity tended to increase as pilots' workload ratings of the flight task increased.

The specifics of the tapping regularity task used in the present study are discussed in the method section of this report.

Primary Task Measures

The third category of workload estimation techniques concerns the measurement of certain task-related variables that directly reflect operator performance on the primary task. In primary task performance measurement it is generally assumed that increased mental loading requirements of the task at hand will consequently degrade operator performance of the task. Fundamentally, primary task measures differ from the other three workload technique categories (opinion, spare mental capacity, and physiological) in that primary task variables supposedly are direct indicants of task performance.

Some disagreement as to the viability of primary task measures exists. Several authors (e.g., Cooper and Harper, 1969) have reported that primary task measures are in many cases not sensitive to changes in task-related mental workload because the human operator possesses a wide range of adaptability (or reserve capacity) through which s/he holds performance relatively constant across changing loads. In other words, as task difficulty increases, the operator merely summons more effort to hold the overall system performance constant. If this is indeed the case, primary task measures that do not rely solely on overall system performance or ultimate system output, but instead on some other task performance parameter, should be examined as potential mental workload measures. If the adaptation assumption is true, the operator must still alter his/her strategy to compensate for changes in loading. The selected measure must then also tap the strategy itself. In a general sense, Wierwille and Williges (1978) in their review on pilot workload concluded that primary task measures were often sensitive to changes in task difficulty in high mental workload conditions but were typically insensitive in low workload conditions.

Primary task measures are often difficult to obtain in full-scale aircraft, largely because of the specialized instrumentation that must be carried on-board and interfaced with the aircraft control systems. In some cases, air-to-ground telemetering of data is necessary. Conversely, these measures are usually easily adaptable to flight simulators because much of the required instrumentation is often present in the simulator's computational dynamics. An additional advantage is that primary task methods of workload estimation do not usually intrude on task performance (Wierwille and Williges, 1978).

Primary task measurement review. Numerous techniques have appeared as primary task indices for a multitude of diverse cognitive and motor tasks. Most of the literature is not relevant to the present investigation because primary task measures, by nature, are quite _task specific_. New workload measurement situations usually require development of new primary task measures. For the interested reader, a comprehensive annotated bibliography of primary task-related literature appears in Clement (1978). Wierwille and Williges (1978) also present a thorough review of primary task measurement procedures in their operator workload assessment report. Several flight-related primary task measures, directly relevant to the present study, will be briefly reviewed below.

Control movements. Probably the most popular primary task measure in vehicular studies is that of control movements. In automobile driving, steering wheel, accelerator, and brake inputs are often recorded. In flight-related studies, aileron, elevator, rudder, and throttle movements are commonly obtained as measures of control difficulty that implicitly reflect pilot psychomotor load. While control movements generally are not considered as indicants of perceptual motor load, they were used as a primary task measure in the present study. The analysis of control movements was performed in an effort to assess the effects of load changes (which were predominately perceptual) on pilot's manual control responses. Several related studies warrant mention.

In a full-scale driving experiment, McLean and Hoffman (1975) found that the number of steering movements per unit time did not correlate with other measures of steering performance. However, steering movements were indeed found to be a sensitive measure of steering task difficulty while driving. Similarly, after conducting moving-base driving simulator studies, Hicks and Wierwille (1979) and Casali and Wierwille (1980) reported results which indicated that steering movements were sensitive to changes in driving task difficulty produced by variations in vehicle dynamics. While in both of these studies the results were statistically significant, the directionality (increasing or decreasing) of steering reversal rates was not consistent between the two studies. Although steering reversal rates are generally expected to increase as driving task demands increase, several studies have revealed a reverse trend (see Macdonald and Hoffman (1980) for a review).

Control movements have also been used as indicants of task difficulty in flight. For example, Dick, Brown, and Bailey (1976) found that aileron, elevator, and thrust control inputs were greatest in frequency under heavy turbulence conditions. The flight task, highly psychomotor in nature, was implemented in a Boeing 737 simulator. The number of pilot control inputs accounted for a major part of the variance in pilots' Cooper-Harper ratings of workload.

Finally, aileron and elevator inputs have been also tested as relative indicants of perceptual workload changes in simulated flight (Rolfe, Chappelow, Evans, Lindsay, and Browning, 1974). In an instrument landing approach task referred to as the perceptual flight plan, Rolfe, et al. obtained several primary task flight performance measures. Pilots' attention to engine instruments was increased by the presence of low oil pressure at two intervals in the approach. Each pilot flew two approaches: one high workload approach and one low workload approach. Aileron and elevator inputs were tended to be higher in the high workload condition, although the differences were not significant by nonparametric analyses of variance. The other primary task measures (localizer deviation, glideslope deviation, and airspeed variability) also each demonstrated similar data trends but no statistical significance.

Detection primary task measures. The second primary task measure used in the present study reflected performance on a danger condition detection-type task. As mentioned in Table 3, a reaction time measure was used. The measure itself is discussed in detail in the method section. Many different detection tasks have been applied in various workload studies, each task utilizing measures specific to that particular task. Several representative studies will be mentioned here.

Laurell and Lisper (1978), used brake reaction distance to a roadside visual detection task as a primary task measure in a highway driving situation. The objective of the study was to validate a secondary task measure, reaction time to auditory signals, as an indicant of driving performance. High correlations between primary and secondary task performance were obtained for various durations of time-on-task. It was concluded that subsidiary reaction time was indeed a valid measure of changes in driving performance.

Using a similar primary-secondary task paradigm, but directed more specifically at workload, Price (1975) compared the effects of three air-to-ground image sensor mounting systems on workload and target acquisition. The three systems differed in the gimbal order of the sensor mount. Gimbal orders included roll-pitch, pitch-yaw, and yaw-pitch. Primary task score, on a target detection-recognition-identification task, was the slant range to the target at the instant of the subject's verbal response. The secondary task required the subject to read aloud visually-presented random digits. Both primary and secondary task measures generally yielded corresponding results. Pitch-yaw gimbal order demonstrated the lowest workload level, while roll-pitch demonstrated the highest level, as indicated by secondary task scores. Similarly, for target acquisition, pitch-yaw was best, while roll-pitch was worst, as evidenced by the primary task scores (slant range).

In 1978, Edwards, Pilette, Biggs, and Martinek used primary task measures in an investigation of the effects of workload changes on operator performance while monitoring ground sensors of enemy activity at a remote location. The results indicated that a percentage measure of correct target detection performance decreased in size as a function of increasing workload. Workload was increased by increasing the number of sensors to be monitored and/or by increasing the level of target activity (more targets per unit time). Errors of omission (missed targets) were much more common than errors of commission (detection of a non-existing target).

Physiological Methods

The physiological category of mental workload estimation techniques includes numerous physiological measures that are purported to reflect changes in operator mental loading (Table 2).

The rationale for using bodily responses as indicants of mental loading is as follows. The fundamental concept is that human physiological processes, such as nervous system activity, circulatory system activity, respiratory activity, and body fluid chemistry, undergo involuntary changes in response to changes in operator workload. Increases in mental workload, as well as physical workload, incur physiological costs on the operator. The amount and type of activity in the autonomic nervous system reflects these physiological costs, in the process of regulation or maintenance of homeostatic internal conditions. Physiological changes in the heart, secreting glands, and involuntary muscles induced by either the parasympathetic or sympathetic branches of the autonomic nervous system are usually independent of conscious thought, with few exceptions (Roscoe, 1978b). Similarly, the autonomic nervous system also functions as a control mechanism for the activation or arousal levels of the operator. The level of activation or arousal refers to the degree of preparedness of the body for the potential release of energy. For workload measurement, the assumption is that changes in arousal level accompany changes in mental workload level. The fluctuations in arousal level may then be measured by appropriate physiological measures. In some cases, workload changes may also induce stress level changes, which in turn affect measurable physiological responses. Stress serves as an intermediary variable and measures of stress may then be related back to workload (Wierwille and Williges, 1978).

Several disadvantages are apparent in the use of physiological responses as mental workload indicants. First, the physiological measurement variable may be contaminated by fluctuations in operator emotional state, mental and physical fatigue, physical exertion, circadian rhythms, metabolic state and a host of other extraneous variables. For this reason, stringent experimental control is of the utmost importance. For example, in some instances stress may be used as an intermediary variable as previously discussed. In other cases, stress and mental load may not covary concurrently, and assumed measures of stress will not accurately reflect mental loading. Another disadvantage is that large inter-individual differences exist as to the directionality and magnitude of physiological changes.

These sources of variability may limit the development of useful metrics of mental workload based on physiological responses. The monitoring apparatus required for certain physiological variables is sometimes highly obtrusive and unwieldy in a cockpit environment. Transducers may be uncomfortable and invasive, such as the catheters often used in blood pressure measurement. Care must be taken to obtain certain physiological measures surrepetitiously, to avoid "conscious" influences in the physiological responses.

Despite these drawbacks, physiological measures do exhibit several attractive features. Foremost, the use of these measures does not assume that the human operator must be a single-channel sampling device. As noted by Wierwille and Williges (1978), a global workload definition may instead be assumed. Also, due to advances in microelectronic and optical technology, physiological measurement transducers are becoming less cumbersome, less invasive, and more reliable and sensitive, increasing their promise for pilot/aircrew application.

Physiological measurement review. Due to the plethora of physiological variables that have received previous research attention, only those variables identical to or closely related to those used in the present study are reviewed in detail below. The interested reader is referred to the following overview articles which include discussion of a broad range of physiological variables: Spyker, Stackhouse, Khalafalla, McLane (1971), Wierwille and Williges (1978), Roscoe (1978b), Wierwille (1979), and Perelli (1979). General discussions concerning the use of physiological methods for mental workload estimation appear in Zwaga (1973), Rolfe and Lindsay (1973), and Ursin and Ursin (1979). Also worthy of mention are two recent texts which provide guidelines for the appropriate application and subsequent implementation and instrumentation of a number of physiological measures (Andreassi, 1980; Stern, Ray and Davis, 1980).

Heart rate. Heart rate, usually measured as the number of ventricular beats per unit time, has received considerable research attention in the workload literature, but with somewhat conflicting results. In general, heart rate or pulse rate (a strong positive correlate of heart rate sensed at a peripheral artery) is expected to increase with increased mental load requirements. Heart rate and other cardiac-related measures are usually obtained from an electrocardiogram, cardiotachogram, phonocardiogram, vectorcardiogram, or plethysmogram. Regardless of the monitoring apparatus, the obtained frequency measure of heart rate is assumed to directly or indirectly covary with mental loading.

First, the findings of several studies using non-flight-related mental loading tasks will be discussed. Mobbs, David, and Thomas (1971) found that heart rate did not signficantly vary with changes in task complexity in paced and unpaced arithmetic tasks, although heart rates tended to rise with increases in error rates. Ettema and Zielhuis (1971) discovered that subjects' heart rates exhibited significant increases as a function of the amount of information handled per minute in an auditory binary choice reaction task. Similarly, Boyce (1974) found a reliable increase in heart rate as difficulty of a mental subtraction task was heightened. However, in laboratory binary choice tasks, Kalsbeek (1973) reported no consistent pattern in heart rate to loading relationships. In a driving experiment, Helander (1975) found that heart rate tended to increase with traffic event difficulty and with the presence of stress-inducing road characteristics such as merging lanes. Conversely, Gaume and White (1975) found no consistent relationship between pulse rate and mental workload on a digital counter monitoring task. Mean heart rate values also showed no sensitivity to driving task difficulty changes produced by degraded vehicle dynamics in a simulated highway driving task (Casali and Wierwille, 1980).

Several studies have used heart rate as a mental load indicant in flight tasks. Bateman, Goldsmith, Jackson, Ruffell-Smith, and Mottodes (1970) found similar heart rate values for commercial pilots over a wide variety of actual flights and simulator flights. In each case, heart rates during flights were higher than resting rates. High-skill inflight maneuvers and stressful inflight segments were accompanied by increased pilot heart rates. Opmeer and Krol (1973) monitored heart rate across four phases of a simulated flight task: (in order of least to most difficult) baseline, level flight, take-off, and approach. Heart rate values consistently increased with the difficulty of the flight phase.

Using a twin-engine civilian transport aircraft, Roscoe (1975) obtained heart rate and subjective measures of mental workload in an evaluation of steep-gradient and two segment noise abatement landing approaches. Heart rate mean values showed a strong positive correlation with the subjective ratings of workload. Later, Roscoe (1976) extended the investigation to include different flight phases, such as take-off, approach, and landing, and also different aircraft. He concluded that heart rate is a worthwhile measure for use in full-scale flight evaluation.

In a simulated precision hover task, Stackhouse (1976) manipulated workload by varying flight control system feedback and by changing displayed information. The results yielded a significant correlation beween heart rate mean and pilot workload.

Finally, Markiewicz, Koradecka, and Konarska (1977) found that in both fixed and rotary wing aircraft in agricultural crop-dusting and sowing flights, pilot's heart rates increased as a function of the particular flight segment. From lowest to highest heart rate, the flight segments were: at rest, before start, start, direct flight, turns, landing, after landing, and flight over obstacle. Only descriptive statistics were presented.

Due to the inconsistency and nondirectionality of results obtained by various researchers using a variety of primary tasks, the heart rate measure of mental workload is somewhat suspect. Therefore, for the present study the frequency measure of heart rate was not investigated. In its place, the more promising measure of heart rate variability was used.

Heart rate variability. The second candidate cardiovascular measure is that of beat to beat irregularity, commonly referred to as heart rate variability or sinus arrythmia. A synonymous term is pulse rate variability, which is sensed at a peripheral bodily location rather than with electrodes on the chest. Measures of variability have typically been extracted from electrocardiograms (EKG), cardiotachograms, and plethysmograms by simple calculation procedures, such as sampling of instantaneous heart rate values or the standard deviation of inter-heartbeat intervals. Complex variability measures, using spectral waveform analysis, have also surfaced in the literature, such as the arrythmia quotient (AQ) used by Laurig and Phillip (1970) or the combined arrythmia measure (CAM) of Luczak and Laurig (1973). Regardless of the measurement used, heart rate variability is generally thought to decrease with increases in mental loading (e.g., Perelli, 1979).

Most of the mental workload research addressing heart rate variability as an indicator has not been flight-deck related. This body of research will be briefly reviewed in chronological order.

In 1967, Kalsbeek and Sykes used a dual-task situation where subjects first performed a visual binary choice reaction task and secondarily performed an auditory binary choice task. The hypothesis that an increase in mental load would coincide with a decrease in heart rate variability was supported. Strong negative correlations were obtained between the channel capacity measure and sinus arrythmia. Using a similar auditory binary choice procedure as their sole loading task, Ettema and Zielhuis (1971) found that heart rate variability obtained from R-R interval measures significantly decreased as the amount of input information per minute increased. These authors noted further that the suppression in variability could also be a function of a change in breathing pattern induced by the task load variation.

In the Mobbs et al. (1971) study, heart rate irregularity did not significantly decrease as arithmetic task complexity increased. Heart rate variability was computed from recordings of instantaneous heart rate values. Using more complex, spectrally-derived, measures of heart rate variability, such as CAM, significant changes in heart rate variability were found to accompany changes in mental loading on a binary choice task (Luczak and Laurig, 1973). Other less complex variability measures did not demonstrate sensitivity to load. Similar results were discussed by Rohmert, Laurig, Philipp, and Luczak (1973). Also in 1973, Kalsbeek (1973) reported that heart beat irregularity was significantly lower for high presentation rates than for low presentation rates on an auditory binary choice task. Finally, Boyce (1974) found a decrease in sinus arrythmia with increases in mental loading on a subtraction task.

In a more applied experimental setting, Hicks and Wierwille (1979) varied mental workload (primarily psychomotor in nature) by changing the center of windgust pressure point of application in a moving-base driving simulator. As workload was increased by moving the point of application rearward on the simulated vehicle, both primary task and rating scale measures reliably reflected the accompanying increase in loading. However, heart rate variability, as computed from instantaneous heart rate samples, did not exhibit a significant decrease with increased mental loading.

Several flight-related studies have incorporated heart rate variability as a measure of mental load. In 1967, Auffret, Seris, Berthoz, and Fatras reported that heart rate variability measures based on instantaneous cardiac frequency values were useful measures of pilot loading. The authors cautioned that establishing heart rate means over long periods could mask true changes in heart rate variability. Stackhouse (1973) found that heart rate variability as measured by R-R interval standard deviation correlated significantly with pilot performance in both simulated and actual flight helicopter hover maneuvers.

Finally, Opmeer and Krol (1973) reported that heart rate variance systematically decreased over four increasingly difficulty flight phases: baseline, level flight, take-off, and approach. These results were obtained in both McDonald-Douglas DC-7 and Beechcraft AT-100 flight simulators. Heart rate variability exhibited greater sensitivity to flight-related mental workload than heart rate alone. Furthermore, the authors concluded that heart rate variance was highly reflective of cognitive load while heart rate was more sensitive to anxiety-inducing tasks.

Respiration rate. Another physiological variable which appears to reflect changes in mental workload is that of breathing or respiration behavior. There is some research evidence that respiration patterns become somewhat more shallow and regular, and often increase in frequency (in terms of inspirations per unit time) in response to increases in load (Wierwille and Williges, 1978). The rate or frequency measure has shown considerable promise and is the easiest of the respiration parameters to obtain. For this reason, respiration rate was investigated in the present study.

Briefly, respiratory parameters have been measured by several different methods which differ substantially as to their degree of obtrusiveness (Wierwille and Williges, 1978). Rough indications of rate and tidal volume can be obtained with thermistors inserted into the operator's nostril. Volume measures and gas component analyses are often obtained by closed-circuit breathing into a Douglas bag or similar apparatus. These two methods are probably too cumbersome for implementation in either actual or simulated flight. Rate and volume may also be obtained indirectly through instantaneous measurement of chest cavity circumference via strain gauges or potentiometer actuation using flexible chest belts. Finally, accurate rate measures can be obtained from movement-sensitive transducers mounted in front of the operator's diaphragm, as discussed in the forthcoming apparatus section of this report.

Several non-flight-related studies have addressed respiration parameters as indicants of mental workload. Jex and Allen (1970) found that breathing was faster and more shallow during performance of a tracking task than during rest. However, no reliable increases in breathing rate occurred with increase in the tracking control dynamics from first to third order. Respiration parameters were obtained using a nasal thermistor. Ettema and Zielhuis (1971), in their binary choice experiment, reported that breathing rate significantly increased with the amount of information handled per minute. Correlations of breathing rate with signal presentation rate were higher than those of heart rate or heart rate variability with presentation rate. Mulder and Mulder-Hajonides van der Meulen (1973) found that in a paced choice reaction task, the most sensitive measure of task load was the number of reversal points in the cardiotachogram. Respiration frequency strongly correlated with this measure. The authors noted that correlations between heart rate variability measures and workload may actually result from the influence of breathing rate fluctuations on heart rate regularity, where breathing rate changes with workload. Finally, Gaume and White (1975) found no reliable increases in respiration rate as workload increased on a digital counter monitoring task. Breathing frequency was obtained using a piezoelectric strain gauge belt.

Respiration rate has also been incorporated as a measure in driving studies. In a full-scale driving task, Lisper, Laurell, and Stenig (1973) found a slight decrease in respiration rate as driving time increased. Casali and Wierwille (1980) noted significant increases in respiration rate from baseline to simulated driving conditions and more importantly, as a function of degraded vehicle dynamics. However, the thrust of neither of these studies was directed specifically at mental workload.

In flight-related workload experimentation, some attention has been devoted to measures of respiratory parameters. Spyker et al. (1971) manipulated workload in a simulated two-axis aircraft control task by changing pitch axis stability and wind gust levels. A positive correlation between respiration frequency and pilot workload was discovered. This correlation was higher than the correlations of heart rate, heart rate variability, and galvanic skin resistance with workload. These results were consistent with the findings of Ettema and Zielhuis (1971).

Both Stackhouse (1973) and Sun, Keane, and Stackhouse (1976) found that several respiration parameters were correlated with operator performance and mental loading levels in simulated helicopter precision hover experiments. Sun et al. (1976) used a strain gauge measure of respiration rate as an indicant of operator loading.

Finally, Smit and Wewerinke (1978) found that pilots' respiration rates reflected their required level of attention in helicopter instrument hover and navigation tasks. In general, respiration rate increased as the level of attention and effort required concurrently increased.

Utility Deficiencies in the Mental Workload Estimation Literature

It is evident from the workload review literature outlined above that little research effort has been directly applied to the problem of specifying a viable workload estimation technique for a given pilot/aircrew problem. Presently, it appears that the best procedure for a particular situation is to apply the literature review algorithm discussed in Wierwille and Williges (1978). The imminent need for assessing individual workload estimation techniques as to their sensitivity to changes in mental loading on specific tasks (which elicit specific behavior processes) is evident. Furthermore, the relative intrusion of individual techniques on primary task performance needs empirical determination. Finally, it would increase the range of utility of an estimation technique if the transferability of the technique from one behavior process (e.g., psychomotor, perceptual, mediational, communicative) to another were known.

The consequences of specifying a nonoptimal technique are considerable. First, an estimation technique which is not reliably sensitive to shifts in mental loading on a particular process may mask true differences in workload. In an overload or near-overload situation, an insensitive technique could ultimately lead to acceptance of a hazardous pilot/aircrew procedure or aircraft design. An intrusive technique may alter performance of a fundamental flight-related task, contaminating the results of a workload sensitivity analysis because the pilot behaves in a non-normal manner. Furthermore, if intrusion is too high, aircraft control may be severely degraded, rendering full-scale flight testing infeasible.

Research Objective

In the research described herein, several workload estimation techniques were comparatively evaluated under identical experimental conditions in a flight simulator. The objective of this comparison process was to determine the relative sensitivity and intrusion of each estimation technique in applications to a piloting situation which emphasized the use of perceptual behaviors. All comparisons were quantified by statistical analyses and subsequent recommendations concerning the selection of a particular technique are made in the conclusion section of this report.

Due to the breadth of pilot behaviors required during the numerous aspects of flight, it would be difficult to investigate all four major categories of universal operator behaviors (Berliner et al., 1964) in a single controlled experiment. Therefore, the investigation concentrated on a single behavior category: perceptual behavior processes. The piloting task that was developed for the study emphasized the use of perceptual behaviors and is discussed in detail in the forthcoming method section. Parallel studies in the Vehicle Simulation Laboratory at VPI&SU will focus on mediational and communicative processes. A very recent study by Connor (1981) addressed similar techniques to the present study, but utilized a psychomotor task. Because several parallels could be drawn between the Connor study and the present study, his research is discussed in the conclusions section of this report (along with the results of the present study). The results of these three studies will be integrated with those of the present study to provide insight into the transferability of various techniques across behavioral processes.

EXPERIMENTAL METHOD

Subjects

A total of 48 male pilots were used as volunteer subjects in this experiment. All available licensed pilots in the vicinity of VPI&SU were contacted and interviewed by telephone to determine their interests and qualifications for participation in the workload research. No interested female pilots were found in the vicinity. Minimum requirements for participation included: (1) a VFR (visual flight rules) private pilot's license; (2) 50 hours of piloting time in full-scale general aviation aircraft; and, (3) no flight time in a Singer-Link flight trainer-simulator over the past five years. The subjects ranged in piloting experience from 60 hours to 2500 hours with a mean of 289 hours. Assignment of subjects to conditions was based solely on piloting experience. This was done to obtain a cross-sectional representation of the range of experience levels for each workload estimation technique. Details of subject assignment are discussed in the experimental design section.

All pilots were paid a gratuity of five dollars per hour for their participation in the two to three hour experiment. Out-of-town subjects were reimbursed for vehicle travel expense at the rate of 18 cents per mile.

Flight Simulator Apparatus

Flight simulator. The fundamental apparatus used in this experiment was the Singer-Link model GAT-1B flight simulator located in the VPI&SU Vehicle Simulation Laboratory. Figure 3 shows the simulator and associated peripheral equipment. The GAT-1B is a high fidelity simulation of a single-engine light airplane, such as a Cessna 120. Incorporated in the simulator is a complete illuminated instrument panel including all basic flight instruments: attitude indicator, altimeter, directional gyro, airspeed indicator, vertical speed indicator, and turn-and-slip indicator (Figure 4). Additional instrumentation for very high frequency omnidirectional range (VOR) navigation and instrument landing purposes, consisting of localizer and glide slope indicators and dual navigation/communication (NAV/COM) transceivers are also included for simulation of instrument flight rules (IFR) conditions (Sanderson, 1978). Automatic direction finding equipment (ADF) is also located on the instrument panel. (The ADF and IFR navigation equipment were not required for the present study).

Engine and fuel conditions are displayed in a bank of six instruments located on the lower left quadrant of the instrument panel (Figure 4). Right and left fuel tank gauges, alternator ammeter, oil pressure gauge, oil

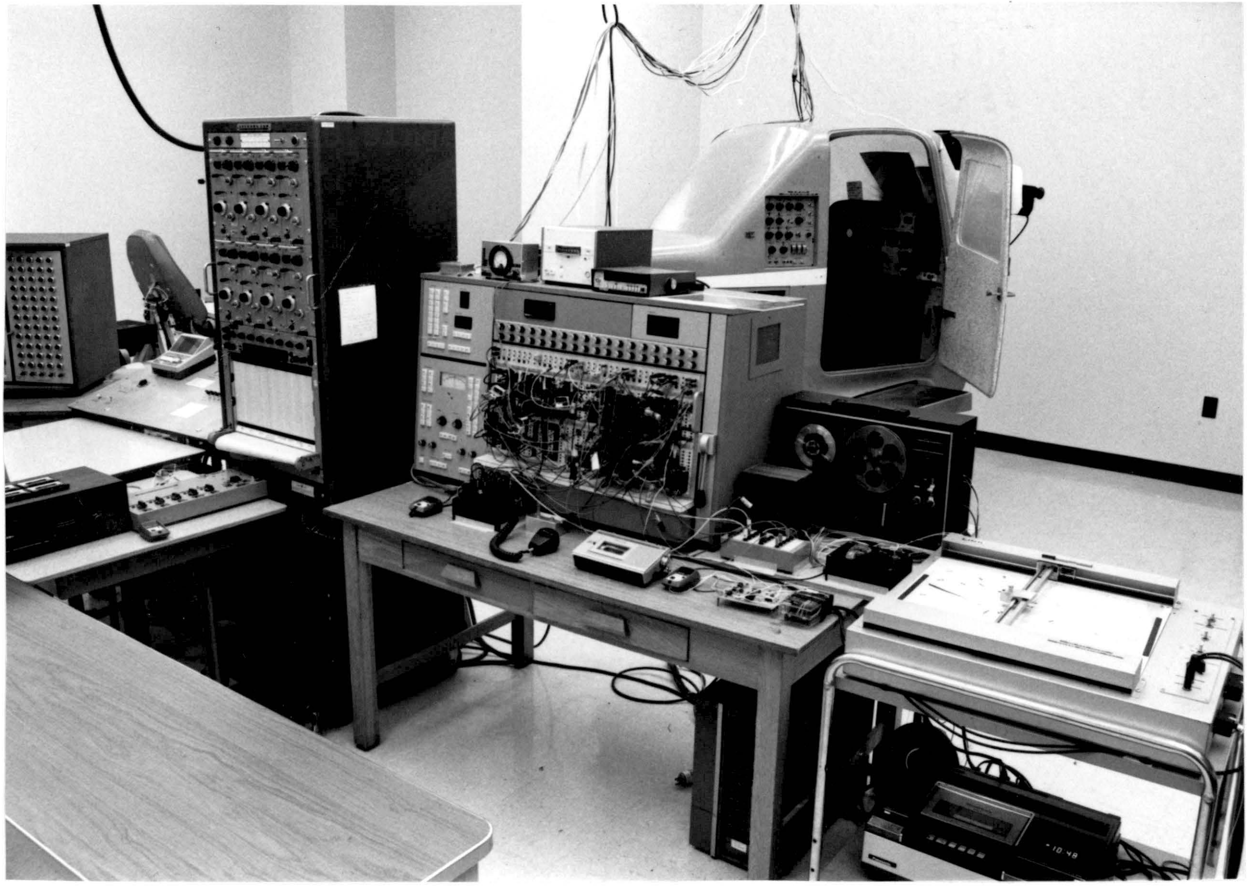


Figure 3. VPI&SU Singer-Link flight simulation laboratory.

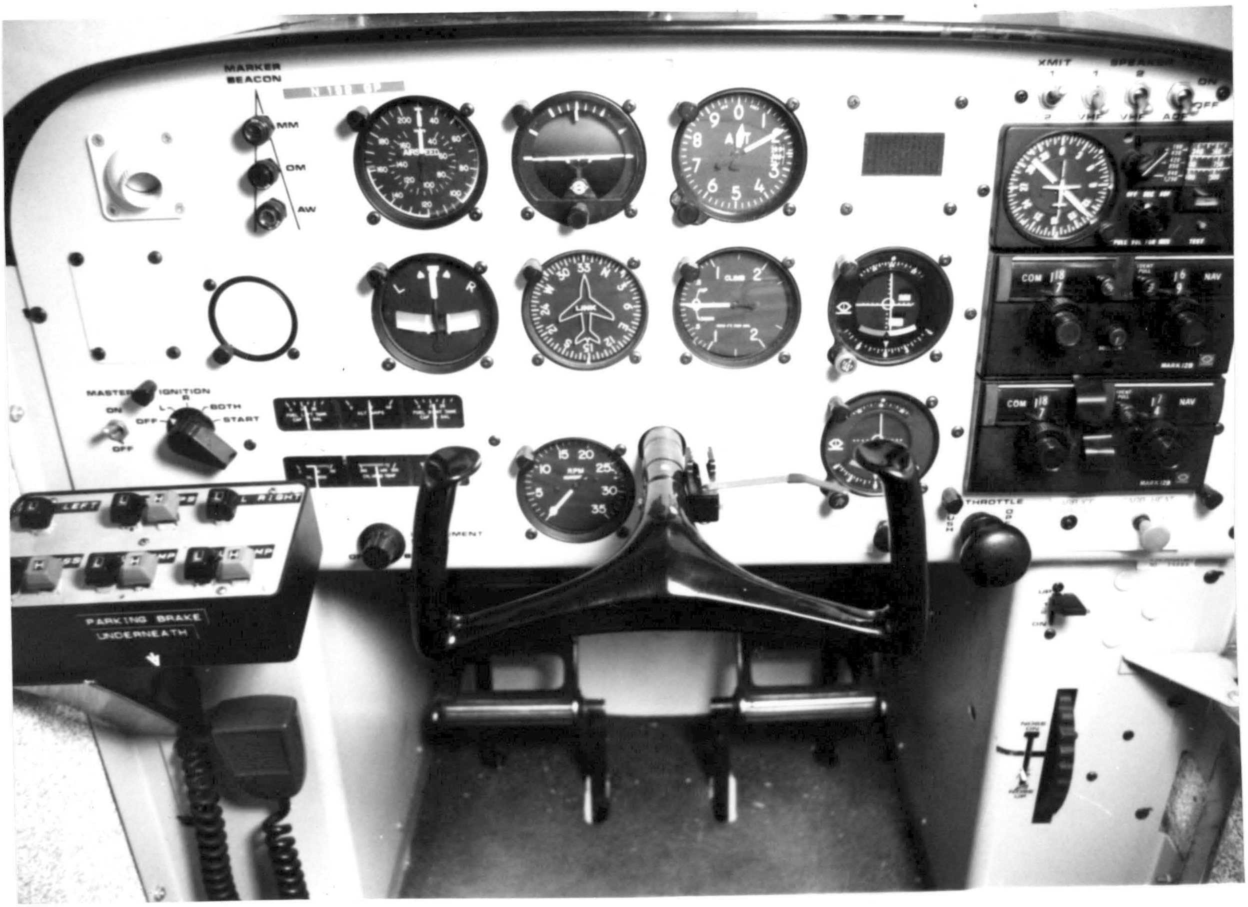


Figure 4. Flight simulator instrument panel.

temperature gauge, and cylinder head temperature gauge are provided. The cil pressure, oil temperature, and cylinder head temperature gauges all have definite redline (danger) positions on the instrument faces and their needle pointer positions are externally adjustable by an experimenter or flight instructor to indicate engine problems. A tachometer displaying engine rpm is also present, as is a time clock.

Simulator controls. The GAT-1B has several controls typical of a single-engine, propeller-driven aircraft. Primary flight controls are yoke-and-column for aileron and elevator position, foot pedals for rudder position, and throttle push-pull knob for engine speed. Other controls include a thumbwheel for minor elevator trim adjustment, toggle switch for wing flaps adjustment, carburetor mixture adjustment knob, carburetor heat adjustment knob, engine ignition switch, parking brake knob, and instrument lighting adjustment rheostat.

Pilots in the GAT-1B communicate with an experimenter via a lapel microphone and cockpit speaker system. (Push-to-talk microphones and headphones were not used in this experiment because of their obtrusiveness and interference with other equipment.) Other features of the GAT-1B include an audio system for engine sound and tire screech upon landing, and an adjustable fan-forced cockpit ventilation system. An aural warning signal sounds when the stall speed of 89 kilometers-per-hour (55 mph) is reached (Singer, 1973).

One of the most important features of the GAT-1B is its three-axis motion system. Pilot's control deflections and experimenter-introduced environmental (e.g., turbulence and barometric pressure), engine, and load distribution parameters serve as inputs to the hybrid electronic airplane dynamics. The dynamics computer is a custom unit manufactured by Singer-Link. This dynamics computer then provides output signals for the motion system electric servo motors, moving the simulator in the roll, pitch, and yaw axes. Motion system signals for the servo drives pass through slip-rings located in the base of the simulator.

The motion system is integrated with the other interactive systems of the simulator including the instrument panel, manual controls, and audio system. Proper coordination among these systems provides for a dynamically-realistic closed-loop light aircraft simulation.

A full fiberglass cab encloses the cockpit of the simulator (Figure 3). The simulator is normally operated under room lights with translucent blinders over each window area. This prevents subject distraction from irrelevant room cues and also maintains a constant level of cockpit illumination. Fortunately, the GAT-1B simulator has no reported history of eliciting "simulator sickness" in subject pilots. Furthermore, the simulator equipment was approved by the VPI&SU Institutional Review Board for use in research involving human subjects, prior to this experiment.

Simulator modifications and primary task measurement apparatus. In preparation for the present experiment, numerous modifications to the standard GAT-1B simulator were made.

First, to enable the difficulty or loading inherent in the primary flight task to be varied on a perceptual (detection) dimension, external control over engine, fuel, and carburetor icing instruments was necessary. Therefore, a switch-operated remote control panel was fabricated and interfaced with the GAT-1B instrument drive circuitry. This control panel enabled the experimenter to position the instrument needle pointers on the oil pressure, oil temperature, cylinder head temperature and alternator ampere instruments at either low or high redline (danger) settings (Figure 4). Pointers on the left and right fuel tank gauges could be individually positioned at an extreme low fuel level (past the last graduation mark on the left of the fuel gauge face). Experimenter control over the needle pointer positions on the engine/fuel instruments was done on an "all-or-none" basis. That is, the needle pointer remained at a near-centered (normal) position unless overridden by the experimenter, in which case the pointer moved completely past the redline (danger) level. The change in pointer position from normal to danger was quite noticeable because of the large deflection angle on the instrument face. The pilot responded to each danger condition by pressing a momentary-contact pushbutton corresponding to that particular condition. One pushbutton was provided for each possible danger condition: low oil pressure, high oil pressure, low oil temperature, high oil temperature, low alternator amperes, high alternator amperes, low cylinder head temperature, high cylinder head temperature, low left fuel tank, and low right fuel tank. Actuation of the correct pushbutton caused the danger condition to return to normal. The pushbuttons were positioned on a module in a one-to-one correspondence with their associated instruments for proper control-display compatibility. This module was located on the lower left of the instrument panel, just below the engine/fuel instrument quadrant.

In addition to the engine/fuel instruments, the remote control panel also provided experimenter control over a carburetor ice warning light. This light, a Monsanto MV5075 red light-emitting diode (LED), was added to the GAT-1B instrument panel on the opposite side from the engine/fuel instruments, so that pilot would need to perform a full scan of the panel to detect all possible danger indications. The LED was clearly labeled as to its function: "CARB ICE

WARNING." When the experimenter switched on the carburetor ice warning, a five-volt signal caused the LED to light. The pilot responded to (extinguished) the LED by pulling an adjacent "CARB HEAT" knob.

During a flight, one, some, or all of the engine, fuel, and carburetor icing instruments could display danger conditions, as is described in the forthcoming experimental design section. Also, the experimenter could simultaneously provide more than one danger condition. It should be noted that the introduction of danger conditions via the remote control panel--GAT-1B instrument interface in no way affected flight performance of the aircraft. Danger conditions were simply needle pointer movements; no disturbances were applied to the simulator's dynamics as a consequence of the presented danger condition. To sequence the presentation of danger conditions during each flight, the experimenter was cued by an audio cassette tape which specified when a particular danger condition should be presented. Different cassette tapes were used depending upon the particular loading level of the primary task. The cassettes were played on a BIC Model T-1 cassette deck through Realistic Nova-50 headphones.

Danger condition reaction times were obtained for responses to the carburetor ice warning LED. When the experimenter switched the LED on, a Cronus Olympian digital timer was started automatically. When the subject responded to the LED by pulling the carb heat knob, the timer was stopped by an electrical pulse from the carb heat switch. Reaction times accurate to 0.01 second were obtained from the timer readout.

Other modifications to the simulator provided a means of obtaining the control movements (per second) primary task measure and two additional primary task measures, pitch and roll high pass deviation, for use in the intrusion analysis. (The intrusion analysis and the specific measures used are fully discussed in the forthcoming experimental design section; only the necessary measurement equipment is discussed here.) The electronic signals for control movements were obtained from standard GAT-1B potentiometers mounted on the yoke-and-column and rudder pedals. A cumulative recording of control movements during a flight was obtained using a Hewlett-Packard Model 1600A digital counter, after being processed by an EAI-380 analog/hybrid computer. For the pitch and roll deviation measures, raw pitch and roll position signals were obtained directly from the GAT-1B dynamics computer and inputted to high-pass filtering and mean square computational programs on the hybrid computer. The final mean square values were displayed as scaled voltages at the end of a flight. Low-frequency deviations were filtered out to insure that differences in aircraft trim level between pilots would not mask true differences in heading- and attitude-following control, as reflected in the pitch and roll angular excursions of the simulated aircraft. The cutoff frequency of the high-pass filter for both pitch and roll was 0.05 radians-per-second. Pitch and roll mean square signals were also displayed as traces on a Sanborn Model 350 chart recorder for visual monitoring purposes during a simulated flight.

All other workload estimation measures, with the exception of the opinion measures, were obtained from sensors located elsewhere on the GAT-1B or directly on the subjects. Signals from these sensors traveled from the simulator to the experimenter's station via a multi-conductor umbilical cable tethered above the simulator cab. Sensing apparatus for each measurement system is discussed below.

Opinion Measurement Apparatus

Modified Cooper-Harper scale. The standard form of the Cooper-Harper scale, shown previously in Figure 2, is primarily intended for pilot assessment of aircraft handling qualities (Cooper and Harper, (1969). Because the scale has been proven effective in previous flight research it was desirable to include a Cooper-Harper-like scale as a workload estimation technique in the present study.

However, the standard form of the scale does not lend itself to ratings of perceptual-type tasks, due to the nature of the descriptors used in the decision tree. The modified Cooper-Harper scale used in this experiment is shown in Appendix B. This scale uses the same decision tree flowchart and numerical scale values as the Cooper-Harper scale; only the verbal descriptors have been altered so that the scale is not limited to ratings of aircraft controllability. An effort was made by the investigators involved in the present study to make the scale more generalizable for a wider variety of task workload applications. As such, new instructions were also needed; these appear in Appendix C.

Multi-descriptor scale. The workload multi-descriptor rating scale, developed by Dr. W. W. Wierwille at VPI&SU, was the second opinion measure used in this experiment. The scale itself appears in Appendix D and the associated instructions, as given to the subjects, appear in Appendix E.

Secondary Task Apparatus

Time estimation equipment. In the time estimation secondary task, the subject was prompted to begin mental production of a 10-second time interval by a tape-recorded "now" message over the cockpit speaker. A Realistic Minisette-10 cassette recorder was used and each prerecorded prompt was separated by approximately 20 seconds on tape. After hearing the "now" signal, the subject pressed a yoke-mounted microswitch to start the interval and again to signal a 10-second lapse. These microswitch depressions provided start and stop signals for a custom-designed logic circuit which subsequently provided switch pulses for a Cronus Olympian digital timer. The timer was accurate to 0.01 second. The clock on the simulator instrument panel was covered so as not to interfere with the time estimation measure.

Tapping regularity equipment. For the Michon tapping secondary task, subjects were required to tap the yoke mounted microswitch as regularly as possible at a rate of one tap every two seconds. The downstroke of the microswitch ended one interval and began the next. Subjects were instructed to this effect. The signals (taps) from the microswitch were inputted to an integration-track store hardwire program on the EAI-380 computer. The output from this program was displayed as a trace on the Sanborn 350 stripchart recorder. The resultant recorded trace allowed accurate subsequent analysis of the length of time between successive taps. These "delta" values were needed for computation of the tapping regularity measure using Michon's (1966) formula (Appendix F).

Physiological Measurement Apparatus

The physiological workload measures required substantial instrumentation of the subject pilot. Of course, the overriding concern when "wiring" human subjects for the purpose of physiological monitoring is the elimination of electrical shock potential. No electrodes were used in the present study. Also, for both physiological transducers, the subject was "floating" in the circuitry, completely isolating him from any electrical signals. An instrumented subject is shown in Figure 5.

Pulse rate variability monitoring equipment. This cardiovascular measure was obtained using a small thermally-insulated, light-activated sensor fitted over the antihelix of the right ear of the subject. This sensor, a Hewlett-Packard plethysmograph, consists of an earpiece module with two separate, but facing sides. One side of the module contains an infrared light source. The other side contains a phototransistor (photosensitive cell) located directly colinear with the infrared source and separated from it by the antihelix wall of the ear. The infrared source remains on continuously, illuminating the phototransistor through the skin of the ear. The plethysmograph, a transmissivity device, is sensitive to quick changes in opacity of the skin of the ear. The heart produces a pressure pulse which propagates from the aorta to the peripheral arteries, such

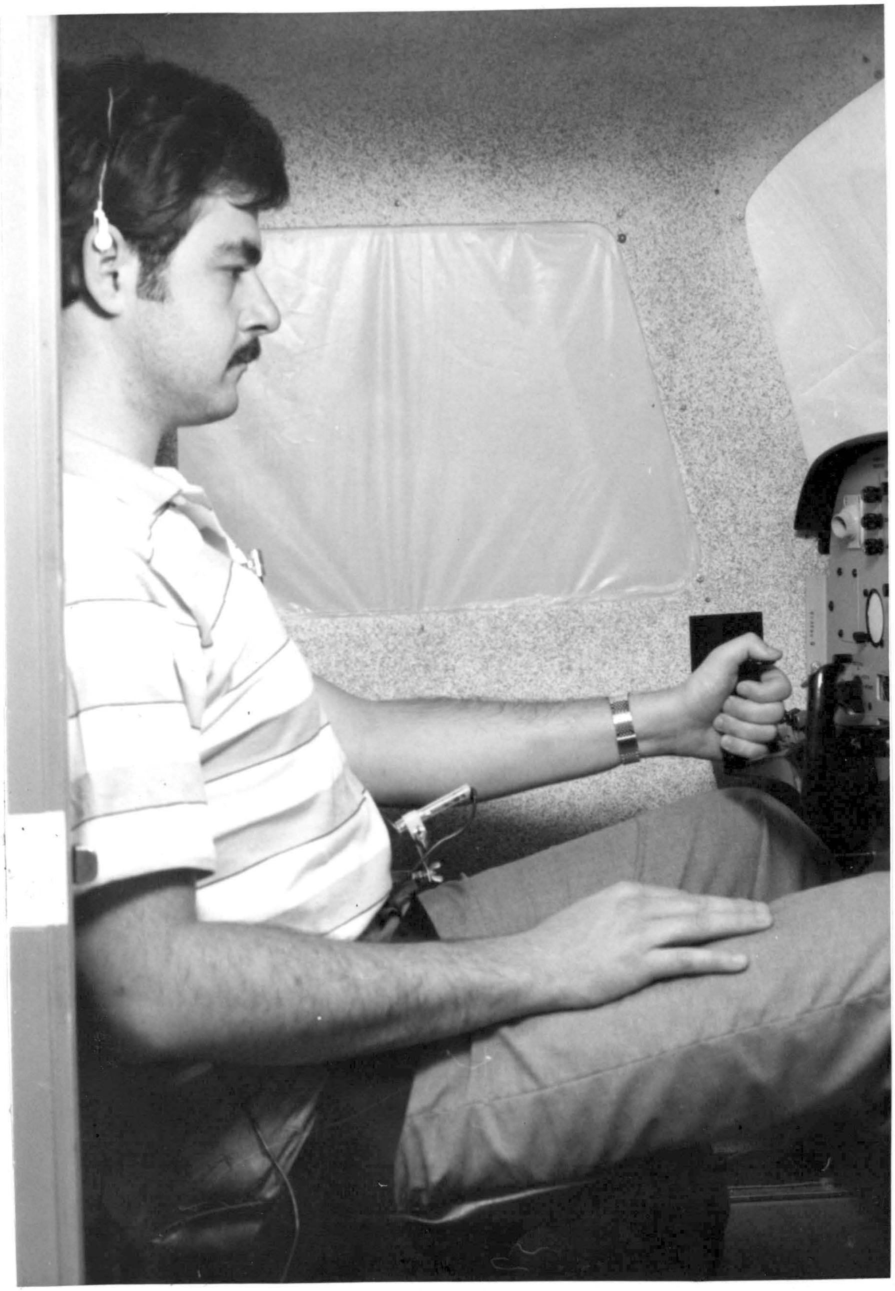


Figure 5. Subject wearing physiological sensors.

as those in the ear. When the volume of blood present in the tissues of the ear increases due to the pulse wave and vasodilation, the opacity of the skin of the ear increases. An increase in opacity reduces the amount of light incident upon the phototransistor, resulting in a change in voltage output level.

The output voltage signal from the plethysmograph inputs via cable to a remote Hewlett-Packard Patient Monitor, Model 78203C, where the signal is conditioned and the heart pulse rate is determined. An analog signal of heart pulse rate and an analog pressure pulse waveform for each ventricular beat are available as output signals from the Patient Monitor. These signals were processed by the EAI-380 hybrid computer for on-line computation of pulse rate and pulse rate mean square values. Both pulse rate and pulse rate mean square values were needed for computation of pulse rate standard deviation. At the end of a data-recording period, voltage values corresponding to these cardiovascular measures were displayed on the EAI-380 digital voltmeter. These voltage values were later converted to actual pulse rate variability scores. Also, a trace of the heart pulse waveform was recorded on the Sanborn stripchart recorder for visual monitoring by the experimenter during data runs.

Respiration rate monitoring equipment. Respiration rate in breaths-per-minute was obtained using a sensor fabricated in the Vehicle Simulation Laboratory (Casali and Wierwille, 1980). The circuit diagrams for this unique sensor appear in Appendix A. This apparatus consists of a flexible metal belt, positioned around a seated subject's upper abdomen. This belt supports the respiration transducer and positions it about 0.5 inch (1.27 cm) in front of the subject and slightly below the bottom of the ribcage. Inhalation and exhalation during breathing produces expansional and contractional movements of the abdomen to which the transducer is sensitive. Basically, the transducer and the subject's body are capacitively coupled. The subject's body acts as an antenna for stray 60 Hz noise. The amount of noise that the transducer receives is a function of the distance between the body and the transducer. The noise signal from the transducer is then actively filtered, amplified, and detected as a slowly-varying voltage signal output. This procedure yields a relatively clean waveform signal from which breathing frequency can be accurately resolved. A permanent recording of this signal was obtained using the Sanborn stripchart recorder.

Experimental Design

In this experiment, two sets of data were collected simultaneously. The major focus of this experiment was a sensitivity analysis applied to the "prime" data set. A subsequent analysis was applied to the intrusion data to determine whether the introduction of certain workload estimation techniques influenced primary task performance. The sensitivity and intrusion data were collected together for efficient use of pilot resources. The two experimental designs are discussed separately below.

Sensitivity analysis design. A mixed three-by-eight complete factorial design was used for sensitivity data collection. The experimental design matrix appears in Figure 6.

Load level was the fixed-effects, within-subject variable. As mentioned previously, load was manipulated by changing the rate and number of engine/fuel/carburetor icing danger conditions occurring during a flight. Each subject flew three flights, where each flight corresponded to a single load level of either low, medium, or high. Using six subjects per technique, it was possible to completely counterbalance the presentation order of load levels across subjects. Counterbalancing served as protection against habituation or practice effects in the simulator. In the low load conditions, only carburetor icing conditions were

		LOAD		
		Low	Medium	High
TECHNIQUE	Modified Cooper-Harper Scale	$S_1 - S_6$	$S_1 - S_6$	$S_1 - S_6$
	Multi-Descriptor Scale	$S_7 - S_{12}$	$S_7 - S_{12}$	$S_7 - S_{12}$
	Time Estimation	S_{13}-S_{18}	S_{13}-S_{18}	S_{13}-S_{18}
	Tapping Regularity	S_{19}-S_{24}	S_{19}-S_{24}	S_{19}-S_{24}
	Pulse Rate Variability	S_{25}-S_{30}	S_{25}-S_{30}	S_{25}-S_{30}
	Respiration Rate	S_{31}-S_{36}	S_{31}-S_{36}	S_{31}-S_{36}
	Danger Condition Response Time	S_{37}-S_{42}	S_{37}-S_{42}	S_{37}-S_{42}
	Control Movements	S_{43}-S_{48}	S_{43}-S_{48}	S_{43}-S_{48}

Figure 6. Experimental design matrix for sensitivity analysis: 3 x 8 mixed-factors (6 subjects per cell).

presented. The average presentation rate on the low loading portion of the primary flight task was one failure every 50 seconds. The medium load condition was limited to left and right fuel tank problems and carburetor icing, at an average rate of one failure per 10 seconds. In the high workload condition, danger conditions appeared on all engine and fuel instruments, in addition to carburetor icing. The average presentation rate of the high level loading portion of the primary task was one failure every five seconds. It should be noted that the loading dimension was ordinal in nature.

Technique, or mental workload estimation technique, constituted the fixed-effects, between-subjects variable. As shown in Figure 6, eight techniques were investigated in the sensitivity analysis. The actual numerical measures and associated computational formulae for each technique are presented in Appendix F. Of course, each subject pilot experienced only one estimation technique.

Subjects, shown within the matrix in Figure 6, were considered as a random-effects variable. As previously mentioned, subject assignment was performed according to pilots' experience level. After the number of piloting hours for each potential subject was determined in the telephone interview, a rank ordering of all experience levels (in hours) was performed. This ranking was then separated into sextiles, with eight subjects in a sextile.

One subject was then randomly selected from each sextile and assigned to the first workload estimation technique condition. This procedure continued for all eight techniques, resulting in a cross-section of six experience levels for each technique. The only other stipulation was that five of the subjects in each technique level were VFR-certified and the sixth subject was IFR-certified.

The experimental design for the sensitivity analysis was univariate, utilizing a single dependent measure called "score." "Score" represents the score or value obtained on each workload estimation technique. Between techniques, there were differences in scaling values for the score, such as beats-per-minute or control movements-per-second; therefore, all scores within a particular technique were converted to standard units prior to application of analysis of variance. This procedure is further discussed in the results section of this report.

Intrusion analysis design. The experimental design matrix for the intrusion analysis was identical to that of the sensitivity analysis shown in Figure 6. The only change was in the type of dependent measure used. For intrusion analysis, four primary task dependent measures were collected concurrently with the sensitivity dependent measure of "score" for each technique. The primary task measures for intrusion analysis included danger condition reaction time, control movements-per-second, pitch high-pass mean square, and roll high-pass mean square. Again, these measures are described in Appendix F. It should be noted that in the two conditions where the workload estimation technique itself was reaction time or control movements, the sensitivity score obtained also served as one of the intrusion analysis measures. Due to the inclusion of four dependent measures, the intrusion analysis was multivariate in nature. The presence of differential intrusion among techniques, including the primary task techniques themselves, would be realized as a main effect of technique in the multivariate analysis of variance.

Experimental Task Procedures

As background to a detailed discussion of the sequence of events which occurred in the experimental procedures, individual experimental tasks (both primary and secondary) are first described below. Subject instruction for all experimental tasks appear in Appendices G-P of this report. The instructions are referred to, in order of presentation, during the discussion of the sequence of events. Of course, all pilots were exposed to the primary task; whereas, only certain pilots were exposed to the secondary tasks, as dictated by the between-subjects technique variable.

Primary task procedure. The "primary task" in this experiment refers to a particular segment of the flight task during which workload level was manipulated and data were obtained. The primary task segment commenced after the subject had taken off, climbed to a predetermined altitude, and leveled off. It is important to note that the primary task was multidimensional. It consisted of flying the aircraft in straight and level attitude (the "navigational control" dimension) while detecting and identifying danger conditions occurring on the engine, fuel, and carburetor icing instruments. The navigational control portion was invariant in difficulty; load was varied solely on the danger condition task. Pilots were instructed to strive to maintain adequate performance on all aspects of the primary task. Adequate performance was defined on the following parameters:

1) maintaining a directional gyro heading of 0 (zero) degrees (0 radians), or due North within ± 10 degrees (0.175 radians),

2) maintaining an altitude of 2000 feet (609.6m) within ± 100 feet (30.5 m),

3) maintaining an airspeed of 100 miles-per-hour (160.9 km/h) within ± 10 miles-per-hour (16.1 km/h), and

4) detecting and identifying <u>all</u> danger conditions on the engine, fuel, and carburetor icing instruments as quickly and as accurately as possible, using the pushbutton module located on the instrument panel.

Each subject flew four flights in which the primary task was presented. The first flight was always a practice flight while the three subsequent flights were experimental flights in which data were obtained. Each experimental flight had a load level (low, medium, or high) inherent in the danger condition detection task. Recall that the actual load levels were specified previously in the experimental design section.

The intent of the experimenters involved in this study was to provide a realistic, "inside-the-cockpit," primary flight task. An essential requirement of this primary task was that load be manipulable on an aspect of instrument flight which emphasized the pilots' use of Berliner, Angell, and Shearer's (1964) "perceptual" behaviors. A "pure" perceptual task, in the strict sense of the word, was not necessary nor feasible for this experiment. A perceptual task limited to the retinal level of sensation was not desired because the realism and pertinence of such a task would have been questionable. Furthermore, such a sensory task would not have been compatible with (and potentially could have undesirably competed with) the more realistic aspects of the primary task, such as the maintenance of specified flight parameters. For these reasons the danger condition detection/identification task was devised. This task required the use of such behaviors as scanning, inspection, observation, location, detection, discrimination, and identification. Also, the danger condition task was particularly germane to simulated instrument flying because pilots are normally apprised of engine, fuel, and carburetor icing conditions during actual flight, due to the criticality of malfunction. The danger condition task was quite compatible with the navigational control portion (maintaining altitude, heading, and airspeed) of the primary task because it simply represented another pertinent aspect of instrument flying required of VFR pilots. Furthermore, as a result of the design of the danger condition task, the pilots were forced to perform complete scans of the instrument panel, from the engine/fuel quadrant on the extreme lower left to the carburetor ice warning light on the extreme lower right.

As discussed previously in the experimental design section, perceptual load was controlled by varying the rate and number of danger conditions occurring during the straight and level portion of experimental flight. Prior to each flight, the subject pilot was informed as to which subset of engine/fuel/carburetor icing instruments would show danger conditions during that flight. In all flights, the number of carburetor icing conditions was invariant, although the subject was not aware of this fact. From subject responses to the carburetor icing LED, performed with the carburetor heat knob, the danger condition response time measure was obtained. The carburetor ice/heat apparatus was instrumented for obtaining reaction time because it was situated by itself on the lower right portion of the instrument panel. As perceptual load heightened with an increase in the rate and number of danger conditions appearing, response time to carburetor ice was also expected to show an increase. Longer response times to carburetor ice was expected to result from a forced decrease in the number of saccadic eye movements from the engine/fuel quadrant or the six primary flight instruments to the carburetor ice LED. Shorter response times in low loading conditions were expected to result from the decreased diversion of attention away from the carburetor ice occurrences. All danger conditions remained "on" (redlined or bright LED) for 15 seconds or until the subject pilot correctly identified the condition by pressing the corresponding pushbutton, whichever occurred first. If the subject had not responded to a carburetor icing occurrence after 15 seconds had elapsed, the experimenter turned off the LED and scored the response time on that trial as 15 seconds. This method of scoring helped insure a conservative statistical posture.

The second dimension of the primary task, that of aircraft navigational control, remained invariant in difficulty across the three experimental flights. If the difficulty of controlling the aircraft had been changed by altering stability or turbulence, psychomotor rather than perceptual behaviors would have been emphasized. However, some external disturbance <u>was</u> necessary to make the flight task realistic from a pilot's viewpoint. As such, crosswinds were included in an effort to force the pilot to scan the attitude gyro, heading indicator, airspeed indicator, and altimeter (the four basic flight instruments) for maintaining altitude, airspeed, and heading. If no external disturbance had been included, the pilot would only have needed to adjust the trim of the simulated aircraft after reaching altitude and occasionally glance at the attitude gyro to maintain instructed parameters. If so, this type of flight task would have been quite unrealistic and one-dimensional, perhaps to the point that desk-top laboratory apparatus could have replaced the simulator as the experimental environment. In each of the three experimental flights, equal amplitude, continuous duration simulated crosswinds were applied from random directions upon the aircraft. The standard GAT-1B random gust generator was used to apply a mild crosswind, having amplitude peaks of approximately 10 miles-per-hour (16.1 km/h).

Secondary tasks procedures. As previously mentioned, two levels of the technique independent variable consisted of secondary task estimation techniques: time estimation and tapping regularity. Both of these secondary tasks were readily adaptable to the present study because they do not require visual input. Additional visual input would have uncontrollably compromised primary task performance on the danger condition detection/identification task. Furthermore, these measures are relatively unobtrusive in terms of essential hardware and are easily adaptable to a flight simulator cockpit. Both tapping regularity and time estimation are easily learned by subjects, and quantitative scoring is quite straightforward (see Appendix F).

For the time estimation task, the method of production was used (Hart, 1976). In both practice and experimental trials, the subject indicated the beginning and end of a 10-second interval by depressing the yoke-mounted microswitch with his right thumb. A single switch depression signaled the beginning of an interval and the next depression signaled the end. Subjects were carefully instructed not to engage in counting, such as verbalizing elapsed seconds, when producing 10-second intervals. As discussed in the apparatus section, the subject was prompted for each trial by the tape-recorded "ready, now" command, with 20 seconds separation between trials. Subjects' time estimates were manually scored using the digital timer circuit during a flight. Subjects were given two opportunities to practice the time estimation task, both alone and concurrent with the primary task, prior to the experimental flights.

In the tapping regularity secondary task, the subject was requested to tap (depress) the yoke-mounted switch at one tap per two seconds with as constant, rhythmical, or repetitive a pace as possible. Again, subjects practiced the tapping task both alone and with the primary task, prior to the experimental flights. Because Michon's (1966) formula for "pels" was subsequently applied to the data, it was necessary to obtain an interval of baseline data (tapping task) alone equal in length to the flight task experimental data intervals (see Appendix F for Michon's formula).

Sequential Experimental Procedure

The perceptual workload experiment consisted of two separate parts or "sessions" within a single two to three hour period. The first session primarily involved the presentation of preliminary forms and instructions and culminated with practice trials on the particular tasks that the subject would encounter during the experimental flights and data collection.

Preliminary/practice session. All simulator and recording systems, with the exception of the GAT-1B sound system, were turned on 20 minutes prior to subject arrival and allowed to stabilize electrically. Also at this time, the levels of the independent variables were prepared and the counterbalancing order for load presentation was determined.

Upon entering the laboratory, the subject was escorted to a desk and presented with a general description of the workload experiment to read (Appendix G). Next, the subject read and signed a participant's consent form, which described policies concerning anonymity of data, knowledge of results, and the privilege of withdrawing from the experiment at any time (Appendix H).

After signing the consent form and agreeing to participate, the subject was rechecked as to his piloting experience level. This log book figure was cross-checked with the experience level obtained in the telephone interview. All pilots' experience levels coincided with their earlier answers. Next, depending upon his particular workload estimation technique condition, the subject read an appropriate set of instructions. Rating scale subjects were given a sample rating scale, either Modified Cooper-Harper or Multi-Descriptor, along with detailed instruction on how to make a rating on that scale (Appendices B-E). Secondary task subjects read either time estimation or tapping regularity instructions, shown in Appendices I and J, respectively. Subjects in the physiological measurement conditions (pulse rate standard deviation and respiration rate) read a general description of the physiological sensors that they would wear during flight (Appendix K). Subjects were not informed as to the function of each sensor, in an effort to avoid such biases as a "conscious" breathing pattern. No preliminary instructions were necessary for subjects in the two primary task techniques of danger condition response time and control movements.

After a period in which the subject could ask questions concerning his instructions, he was escorted to the flight simulator. The subject was seated in the cockpit, the seat was adjusted, and the lap safety belt was buckled. If physiological sensors were required, they were fitted at this time. The subject was then familiarized with the GAT-1B controls, flight instrument panel, communication system, engine/fuel/carburetor icing instrumentation, pushbutton module for identifying danger conditions, carburetor heat knob, and secondary task microswitch. Simulator egress procedures were also discussed.

The next events occurred while the subject was seated in the stationary, quiet simulator. Physiological measurement subjects were told to relax and sit quietly for

12 minutes. This waiting period in the simulator was to allow the subject's physiological responses to stabilize to a near resting level prior to experimental measurement. Secondary task subjects were allowed to practice the secondary task (either time estimation or tapping regularity) by itself for two minutes. After practice, tapping regularity subjects performed the tapping task by itself for an additional five minutes for baseline data collection for application to Michon's (1966) formula. Finally, all subjects read and completed a danger condition rating form (Appendix L). On this form, the subject was instructed to rank order (ties allowed) in order of most severe to least severe, the potential danger conditions which could occur on the engine/fuel/carburetor icing instruments. After ranking, the subjects read part two of Appendix L, which basically instructed him to disregard any pre-bias he may have as to the relative severity of the danger conditions. He was told to respond to all danger conditions as if they were equally critical or severe to flight performance and safety.

The next step in the practice period for all subjects, regardless of their technique condition assignment, was detailed instruction concerning the danger condition detection/identification portion of the primary task. These instructions, which appear in Appendix M, explicitly described the nature of the danger conditions as well as the appropriate responses to them. It was stressed that the subject need only detect and correctly identify the presence of danger conditions as quickly and as accurately as possible. It was further emphasized that the indicated danger conditions would in no way affect performance of the simulated aircraft. The pilot was instructed not to diagnose or compensate for the danger conditions in any manner.

The final event during the preliminary/practice session was the practice flight itself. The purpose of the practice flight was to familiarize the subject with all aspects of the primary task, and where appropriate, the secondary task coupled with the primary task. After reading the instructions shown in Appendix N, the subject was cleared for takeoff. The practice flight proceeded as follows. After takeoff, the subject pilot climbed to an altitude of 2000 feet, leveled off and trimmed the aircraft at 100 miles-per-hour with a due North heading. Upon reaching a "trimmed-out" attitude, the pilot was informed that turbulence would begin. For the remainder of the practice flight, the level of turbulence used in the experimental flights was added. The navigational control portion (maintaining altitude, heading, and airspeed within specified tolerances) of the primary task was practiced for four minutes. At the end of four minutes, the danger condition detection/identification task was presented in conjunction with the navigational control task. Danger conditions were presented over the next three minutes: minute one was representative of low loading, minute two was medium loading, and minute three was high loading. All subjects except those in secondary task conditions then landed the simulated aircraft, ending the practice flight. Secondary task subjects continued flying and practiced the appropriate secondary task in conjunction with the primary task for three more minutes. Secondary task subjects then landed the simulated aircraft. In this manner, all aspects of the primary and secondary tasks were first practiced individually and then concurrently.

Following the practice flight, the subject was allowed to exit the simulator for a short intermission. Questions were also answered at this time.

Experimental flights session. During the intermission, the experimental variables and estimation technique recording equipment were readied for the experimental flight tasks. The subject again boarded the simulator. A detailed timeline that the subject thereafter followed in the experimental flights is shown in Figure 7. The experimenter listened to an audio tape which guided him through the procedures during the flights.

MINUTES ELAPSED	EVENTS OCCURRING
0.0	Flight task instructions: Primary task objectives and adequate performance defined; Possible danger conditions specified.
2.0	Takeoff clearance; takeoff from airport. Climb, begin turbulence.
4.0	Reach altitude, level off, trim out. Straight and level flight; maintain altitude, heading, airspeed.
5.9	All subjects: inform of danger condition detection task start.
6.0	Danger condition detection/identification task start. Exposure to load: Straight and level flight plus danger condition task.
7.9	Secondary task subjects: inform of secondary task start.
8.0	Begin data collection period: Secondary task start for some subjects; Physiological measures start for some subjects; Primary task measures start for all subjects.
13.0	End data collection period: All tasks blanked; Simulator placed in autopilot mode, subject relaxes; Rating scale subjects: perform rating; Experimenter obtains data from computer, digital counter.
17.0	Computers reset, preparation for next flight.

Figure 7. Experimental flight task procedural timeline.

Next, the subject read a set of instructions which reiterated his objectives for the flight task (Appendix O). The specified flight parameters and associated tolerances appeared in the instructions as well as on an instrument panel placard for the subject's referral during the flights. Also, a separate sheet which indicated the particular instruments on which danger conditions could occur during the first flight was given to the subject (Appendix P).

Following takeoff, the subject again climbed to an altitude of 2000 feet, leveled off, and trimmed the aircraft to an airspeed of 100 miles-per-hour and resumed a due North heading. The simulated aircraft was flown in this straight and level configuration under turbulence for two minutes (refer to Figure 7). The subject was then informed over the cockpit speaker that the danger condition detection/identification task would begin and reminded to maintain adequate defined performance on all aspects of the primary task throughout the remainder of the flight. All subjects, regardless of their technique condition assignment, flew the simulator for two minutes while performing the danger condition task. This two-minute period was intended to expose the subject to the particular loading level <u>prior</u> to any measurement. At the end of this exposure period, subjects in the secondary task conditions were instructed to begin performing the appropriate secondary task, while maintaining adequate performance on the primary task. For subjects in the physiological measurement conditions, the measurement period also began at this point. Primary task measurement began for all subjects at this time. In all cases, the measurement or data collection period was five minutes long. The end of the two-minute exposure period was the signal for the experimenter to start the EAI-380 computer program which automatically processed data on-line for the continuous five-minute collection period. The stripchart recorder was also started at this time. Danger condition response times were obtained by the experimenter over the five-minute period using the automatic digital stopwatch circuitry. In the secondary task conditions, an associate experimenter either monitored tapping regularity data on the stripchart or recorded time estimates using the second digital stopwatch circuit.

Immediately following the data collection period, all tasks were blanked and the experimenter placed the simulator in autopilot mode. The subject was told to remove his hands from the controls and to relax in the simulator. Rating scale subjects performed a rating on the flight at this time. Final amplifier (scaled data) values were obtained from the EAI-380 computer by the experimenter and the computer was reset for the next flight. Control movements were read from the digital counter.

Following a brief rest in the simulator, the subject was given instructions for the next flight while the simulator remained on autopilot. The procedure described above and outlined in Figure 7 was followed for the remaining two flights, starting at minute 4.0 in the timeline. At the end of the third and final flight, the subject was allowed to land the simulator. After exiting the simulator, the subject was thoroughly debriefed and paid for participation. The subject was asked not to discuss any aspect of the experiment with other individuals for at least two months hence, to avoid pre-biasing the data.

RESULTS

The data analysis procedures for this experiment were divided into two separate sets, each set having different objectives. The primary analysis was that of sensitivity and the secondary analysis was that of intrusion. These analyses will be discussed individually. Computations for most statistical procedures were performed using the Statistical Analysis System (SAS) computer package as implemented on an IBM 370-165 digital computer (Barr, Goodnight, Sall, and Helwig, 1979).

Sensitivity Analysis

The objective of the sensitivity analysis was two-fold: 1) to determine the overall sensitivity of the various workload estimation techniques to changes in primary task loading, and 2) to establish the relative sensitivity among techniques to changes in loading levels, ultimately providing empirical grounds for selecting appropriate, sensitive techniques in applied workload investigations. The sensitivity analysis procedure discussed herein follows the flow diagram shown in Figure 8. Individual levels within this diagram are numbered and are referred to by numbers in the text of this report.

First (1), the raw data for each technique, such as those in the form of scaled voltages from the EAI-380

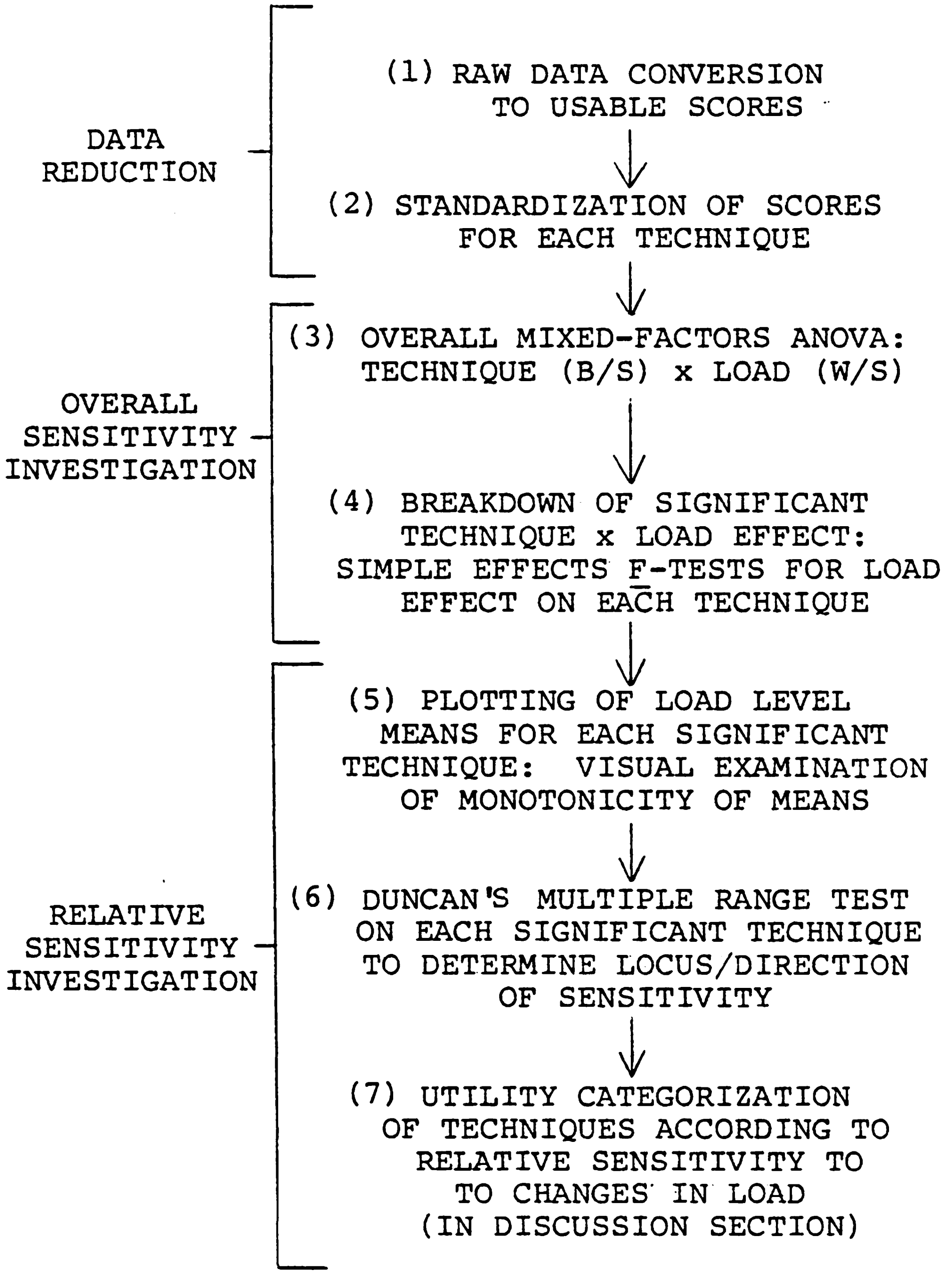

DATA REDUCTION
(1) RAW DATA CONVERSION TO USABLE SCORES
(2) STANDARDIZATION OF SCORES FOR EACH TECHNIQUE
OVERALL SENSITIVITY INVESTIGATION
(3) OVERALL MIXED-FACTORS ANOVA: TECHNIQUE (B/S) x LOAD (W/S)
(4) BREAKDOWN OF SIGNIFICANT TECHNIQUE x LOAD EFFECT: SIMPLE EFFECTS F-TESTS FOR LOAD EFFECT ON EACH TECHNIQUE
RELATIVE SENSITIVITY INVESTIGATION
(5) PLOTTING OF LOAD LEVEL MEANS FOR EACH SIGNIFICANT TECHNIQUE: VISUAL EXAMINATION OF MONOTONICITY OF MEANS
(6) DUNCAN'S MULTIPLE RANGE TEST ON EACH SIGNIFICANT TECHNIQUE TO DETERMINE LOCUS/DIRECTION OF SENSITIVITY
(7) UTILITY CATEGORIZATION OF TECHNIQUES ACCORDING TO RELATIVE SENSITIVITY TO TO CHANGES IN LOAD (IN DISCUSSION SECTION)

Figure 8. General procedure for sensitivity analysis.

amplifier circuits, were converted to a form applicable to data analysis. Next, (2) the reduced scores for each technique were standardized, i.e, converted to Z-scores, insuring that true differences in the techniques were not clouded by scaling value differences between techniques, such as breaths-per-minute versus control movements-per-second. All sensitivity-related analyses were performed on the standardized data set. This procedure has some precedent in the workload literature, having been successfully applied in driving simulation studies by Wierwille and Gutmann (1978) and Hicks and Wierwille (1979), and in aircraft simulation by Connor (1981).

Overall sensitivity ANOVA (3). After data reduction and conversion to standard units, an overall eight-by-three analysis of variance was performed, with technique as a fixed-effects, between-subjects variable and load as a fixed-effects, within-subject variable. The purpose of this ANOVA was to determine if the method of workload manipulation was effective and if the estimation techniques were differentially influenced by the loading task. The results of this ANOVA are shown in Table 4. A strong main effect of load was revealed, $\underline{F}$ (2,80) = 50.67, $\underline{p}$ = 0.0001, indicating that load manipulation was effective. The mean values of the standardized technique scores for the three load levels are shown in Figure 9. Furthermore, and of

TABLE 4

ANOVA Summary Table for Overall Technique by Load Sensitivity Analysis

Source	df	SS	F	P
Between-Subjects				
Technique (T)	7	0.0000	0.00	1.0000
Subjects (S)/T	40	68.9823		
Within-Subject				
Load (L)	2	26.2825	50.67	0.0001
L x T	14	20.0462	5.52	0.0001
L x S/T	80	20.7484		
Total	143	136.0594		

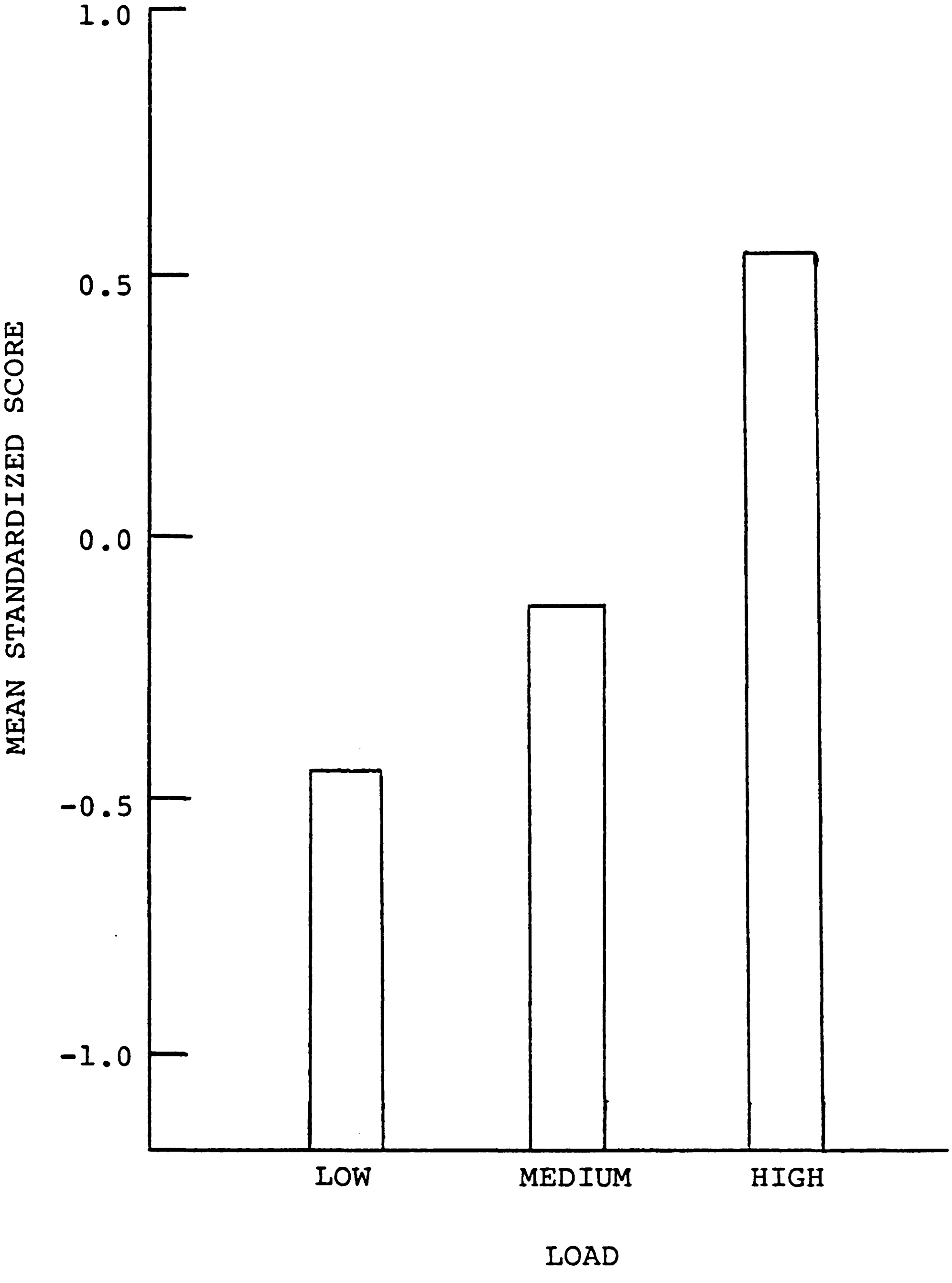


Figure 9. Effect of load on mean standardized scores collapsed across techniques.

immediate importance to subsequent analyses, a highly significant interaction of technique and load was revealed, _F_ (14, 80) = 5.52, _p_ = 0.0001. Of course, due to the standardization procedure, a main effect of technique was not possible, _F_ (7,40) = 0.00, _p_ = 1.0000.

Simple effects F-tests (_4_). The reliable technique by load effect obtained in the overall ANOVA suggested that techniques were differentially sensitive to changes in perceptual load. Therefore, the next step in the overall sensitivity analysis was to examine this interaction effect to determine which _particular_ estimation techniques were sensitive to changes in load. Simple effects _F_ -tests are usually employed in such a capacity.

The use of simple effects _F_ -tests is not warranted if treatment variances are heterogeneous. Hartley's _F_ -max test was employed to test the assumption of homogeneity of variance among the eight levels of the technique variable (Winer, 1971). The homogeneity of variance assumption was supported by the results of the _F_ -max test, _F_ -max (8,10) = 4.46, _p_ > 0.01. Therefore, simple effects _F_ -tests were performed on each technique to determine if there existed a significant effect of load on each technique. Of course, in breaking down the significant technique by load interaction from the overall ANOVA, the same error term for the interaction, load by subjects nested within technique, was required for each simple effects test. The summary tables for all eight simple effects analyses appear in Table 5.

Plotting of means for significant techniques (S). The simple effects F -tests revealed that all techniques except pulse rate standard deviation and control movements, were reliably sensitive to changes in loading on the perceptual task. These results are itemized and significant effects are plotted below.

A strong effect of load on both Modified Cooper-Harper and Multi-Descriptor ratings was revealed, with F (2,80) = 15.57, p < 0.005 and F (2,80) = 25.11, p < 0.005, respectively. The mean standardized scores obtained on these rating scales are plotted as a function of load in Figure 10 for the Modified Cooper-Harper scale and in Figure 11 for the Multi-Descriptor Scale.

While the Modified Cooper-Harper ratings showed a strong sensitivity to load in the simple effects F -test (p < 0.005), one qualification should be noted at this juncture. The Modified Cooper-Harper scale yields, by design, an *ordinal* level of measurement. The simple effects F -test, as a parametric statistical procedure, assumes *interval* scaling. While there is considerable precedent in the literature for applying parametric procedures to Cooper-Harper ratings, as discussed earlier in the literature review, such practice is not entirely appropriate from a

TABLE 5

Summary of Simple Effects F-tests for Effect of Load on Each Technique

Source	df	SS	F	p
Technique: Modified Cooper-Harper Scale				
Load	2	8.0752	15.57	<0.005
Load x Subjects/Technique	80	20.7484		
Technique: Multi-Descriptor Scale				
Load	2	13.0229	25.11	<0.005
Load x Subjects/Technique	80	20.7484		
Technique: Time Estimation Standard Deviation				
Load	2	2.1281	4.10	<0.025
Load x Subjects/Technique	80	20.7484		
Technique: Tapping Regularity				
Load	2	6.7569	13.03	<0.005
Load x Subjects/Technique	80	20.7484		
Technique: Pulse Rate Standard Deviation				
Load	2	0.1055	0.20	>0.10
Load x Subjects/Technique	80	20.7484		(not sig.)
Technique: Respiration Rate (Breaths per Minute)				
Load	2	1.9256	3.71	<0.05
Load x Subjects/Technique	80	20.7484		
Technique: Danger Condition Response Time				
Load	2	13.7480	26.50	<0.005
Load x Subjects/Technique	80	20.7484		
Technique: Control Movements per Second				
Load	2	0.5665	1.09	>0.10
Load x Subjects/Technique	80	20.7484		(not sig.)

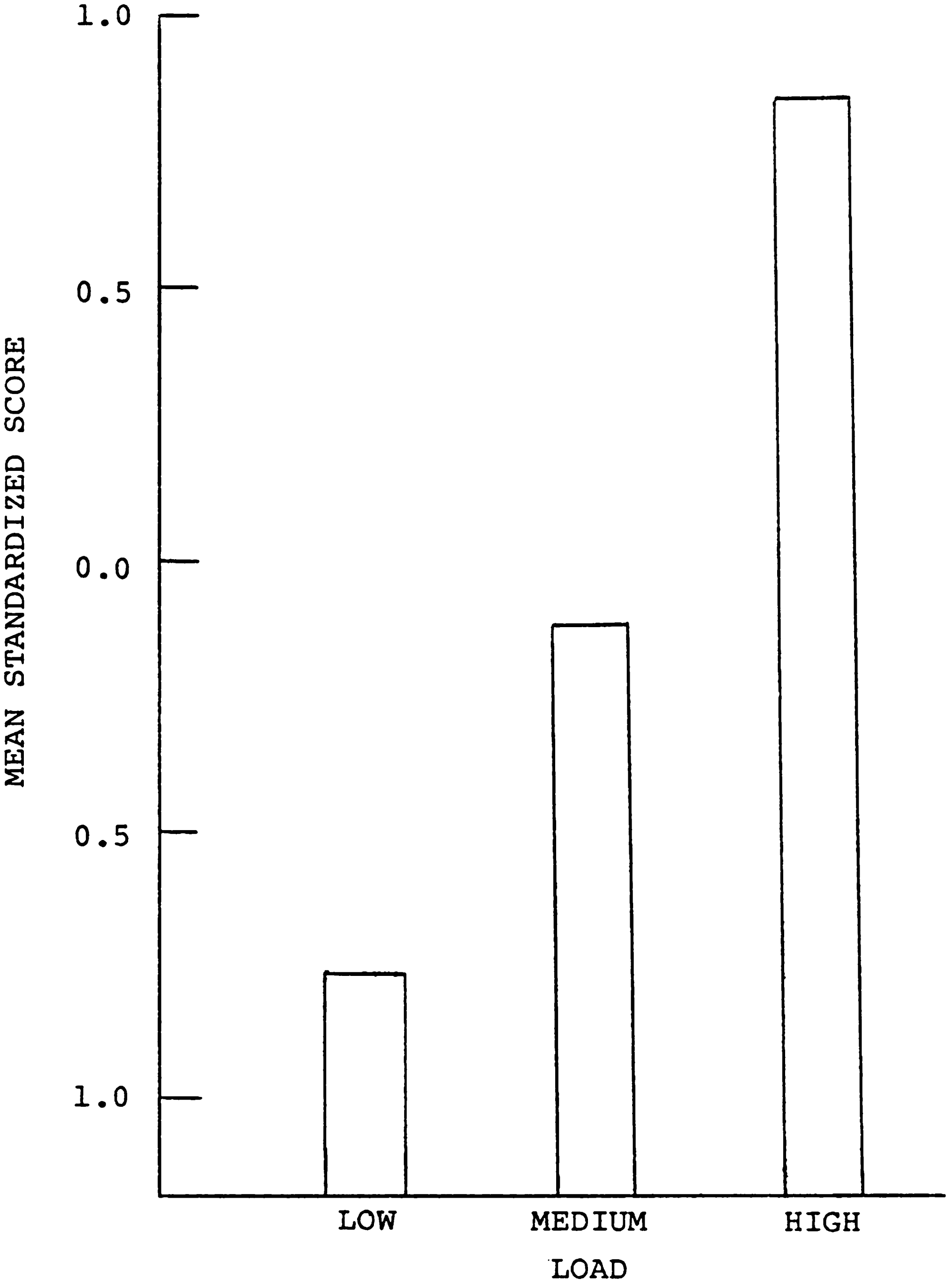


Figure 10. Effect of load on mean standardized scores for the Modified Cooper-Harper rating scale technique.

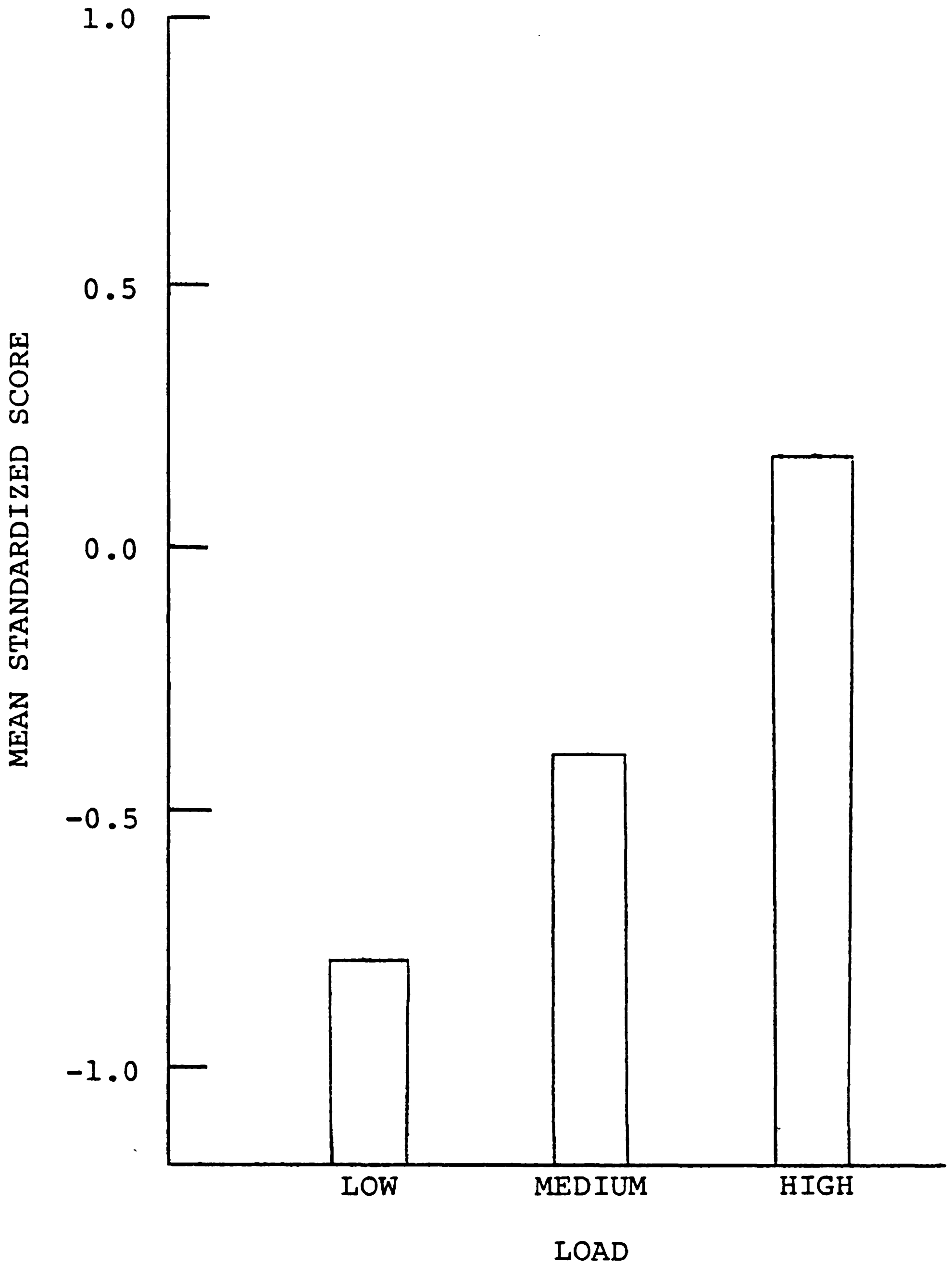


Figure 11. Effect of load on mean standardized scores for the Multi-Descriptor rating scale technique.

pure statistical standpoint. Therefore, the Modified Cooper-Harper ratings in this experiment were also subjected to a nonparametric (distribution-free) procedure prior to proceeding with subsequent sensitivity analyses. The Friedman Rank Sum test was used with a Chi-Square approximation for critical test values (Hollander and Wolfe, 1973). The Friedman test revealed that there were significant differences among loading levels for the Modified Cooper-Harper ratings, $\underline{S'}$ (2) = 10.000, $\underline{p} < 0.01$. This result agrees with that obtained previously in the simple effects $\underline{F}$ -test.

As previously mentioned, the Multi-Descriptor ratings also showed significant sensitivity to loading in the simple effects $\underline{F}$ -tests ($p < 0.005$). This scale, unlike the Modified Cooper-Harper, was designed to provide an approximate interval scale of measurement. Therefore, the use of mean Multi-Descriptor ratings, as defined in Appendix F, and $\underline{F}$ -tests was probably appropriate. However, as a second approach, a nonparametric test was performed on the _median_ ratings of the six multiple descriptors, rather than on the _mean_ ratings. Again, the Friedman rank sum test was applied to the median scale rating for each subject on each loading level. This distribution-free test yielded similar results to the simple effects $\underline{F}$ -test, demonstrating that median Multi-Descriptor ratings were sensitive to load, $\underline{S'}$ (2) = 8.8182, $\underline{p} < 0.025$.

Both secondary task measures demonstrated sensitivity to load. Time estimation standard deviation showed a significant load effect, F (2,80) = 4.10, $p < 0.025$. Tapping regularity yielded significance at F (2,80) = 13.03, $p < 0.005$. The means for these effects are plotted in Figure 12 and 13.

Of the physiological measures, pulse rate standard deviation showed no sensitivity to perceptual loading, F (2,80) = 0.20, $p > 0.10$, while respiration rate increased reliably as load increased, F (2,80) = 3.71, $p < 0.05$. The effect of load on mean respiration rate is shown in Figure 14.

The primary task measure of response time to the danger conditions yielded a particularly strong effect of load, F (2,80) = 26.50, $p < 0.005$. The mean standardized response times, as a function of load, are shown in Figure 15. The control movements measure was not sensitive to changes in perceptual load, F (2,80) = 1.09, $p > 0.10$.

Duncan's Multiple Range comparisons (*6*). The simple effects F -tests were employed to determine which particular workload estimation techniques were sensitive to load. However, the finding that certain techniques demonstrated sensitivity in the simple effects tests did not provide information concerning the *relative* sensitivity among significant techniques. Some techniques may have reliably

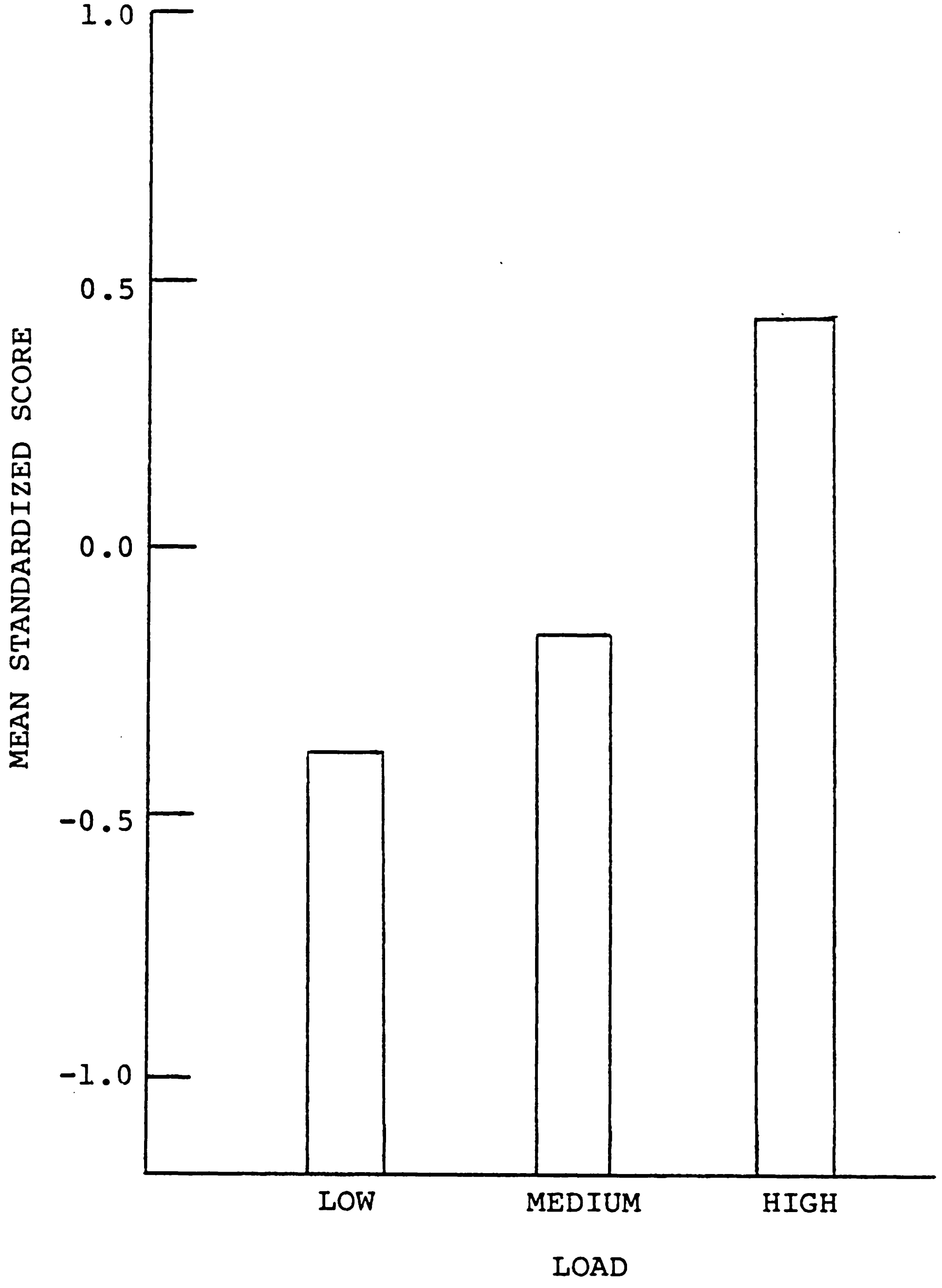


Figure 12. Effect of load onmean standardized scores for the time estimation standard deviation technique.

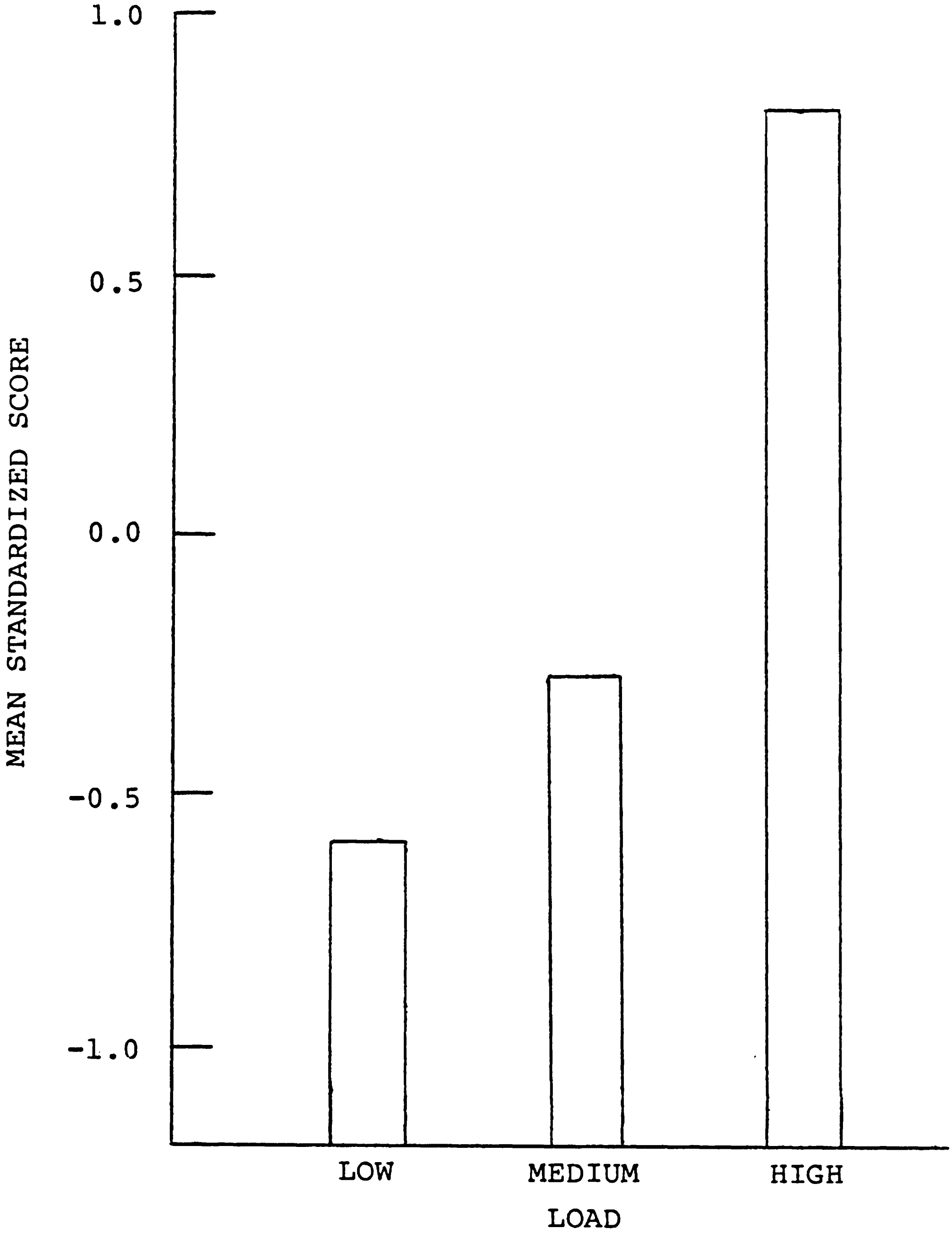


Figure 13. Effect of load on mean standardized scores for the tapping regularity technique.

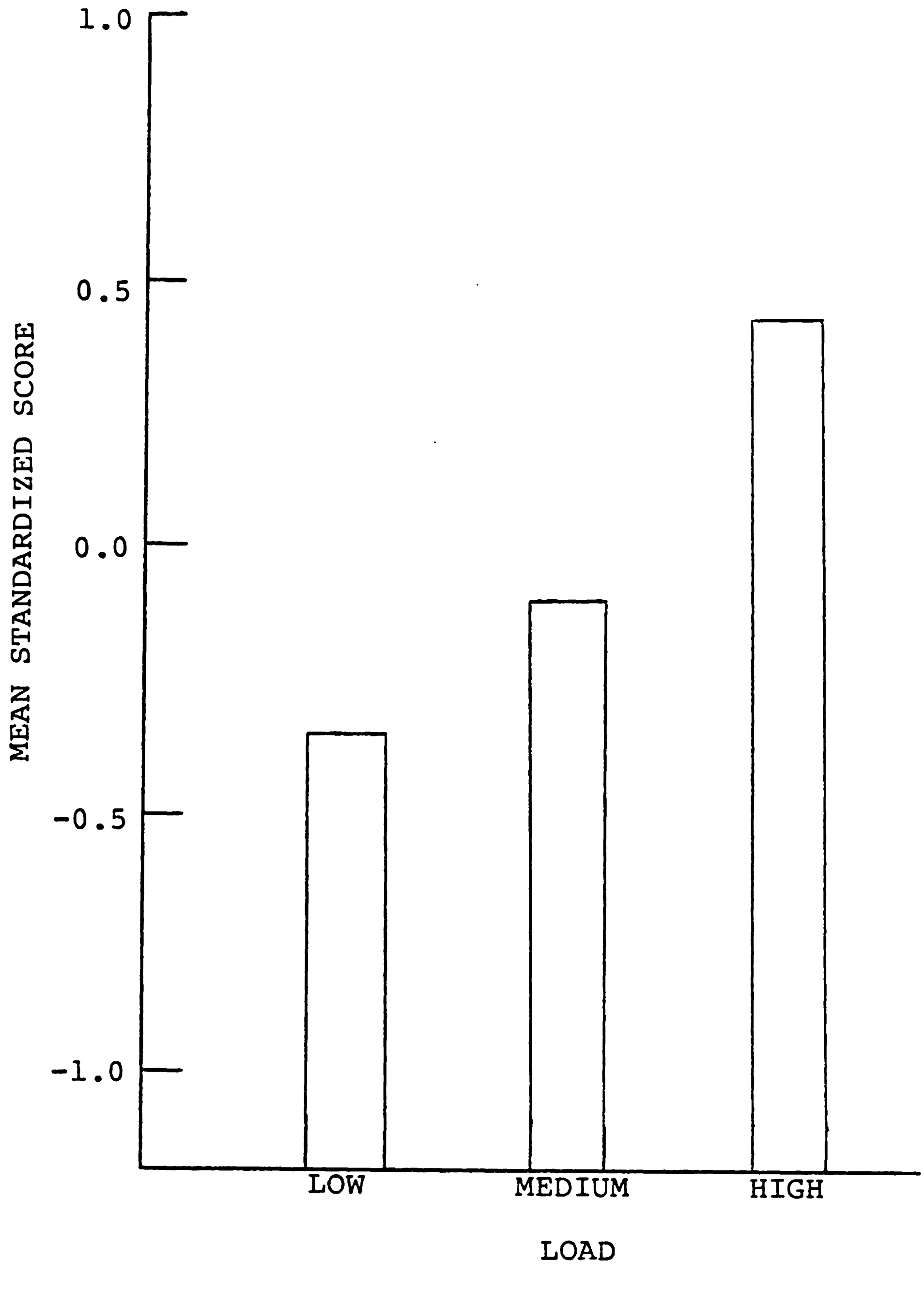


Figure 14. Effect of load on mean standardized scores for the respiration rate technique.

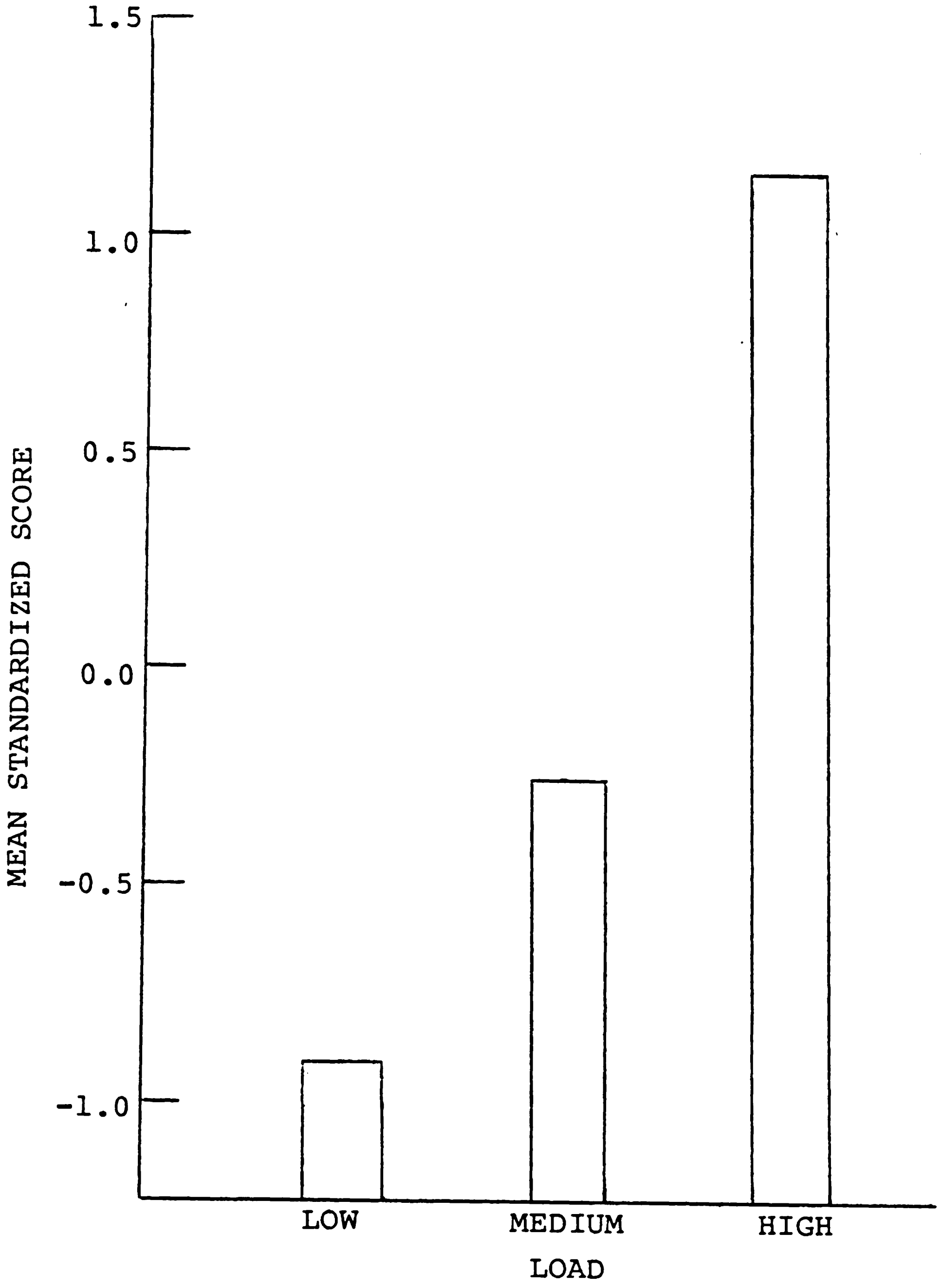


Figure 15. Effect of load on mean standardized scores for the danger condition response time technique.

discriminated between all three combinations of loading levels while others may have discriminated between two or even one combination for sensitivity to be realized in the simple effects tests. Therefore, to examine the locus and direction of the load effect on each significant technique, Duncan's Multiple Range tests were employed (Duncan, 1975). As in the simple effects _F_ -tests, the error term from the overall sensitivity ANOVA, load by subjects nested within technique, was used in the Duncan's tests.

Both the Modified Cooper-Harper and the Multi-Descriptor rating scales showed a distinct monotonic increase in mean ratings across load (Figures 10 and 11). The results of the Duncan's tests on these two sets of means are shown in Table 6. Modified Cooper-Harper ratings were significantly different for all pairs of loading levels, with $p < 0.05$. Multi-descriptor ratings reliably discriminated between low and high load, and medium and high load, but not low and medium load.

As previously shown in Figures 12 and 13, the means for both secondary task measures increased monotonically across load, with a slight compression between low and medium load levels. This increase was indicative of increased variability in performance of the secondary task (time estimation standard deviation or tapping regularity) as load level increased, especially between medium and high load levels. The Duncan's tests in Table 7 bore out this result.

TABLE 6

Results of Duncan's Multiple Range Analyses for the Rating Scale Techniques*

Technique: Modified Cooper-Harper Scale

Load Level:	LOW	MEDIUM	HIGH
Mean Value:	-0.7603	-0.1083	0.8695

Technique: Multi-Descriptor Scale

Load Level:	LOW	MEDIUM	HIGH
Mean Value:	-0.7948	-0.3897	1.1777

*means with a common line do not differ significantly at $p < 0.05$.

TABLE 7

Results of Duncan's Multiple Range Analyses for the Secondary Task Techniques*

<u>Technique: Time Estimation Standard Deviation</u>

<table>
<tr><td>Load Level:</td><td>LOW</td><td>MEDIUM</td><td>HIGH</td></tr>
<tr><td>Mean Value:</td><td>-0.3963</td><td>-0.1677</td><td>0.4200</td></tr>
<tr><td></td><td colspan="2">________________________</td><td>______</td></tr>
</table>

<u>Technique: Tapping Regularity</u>

<table>
<tr><td>Load Level:</td><td>LOW</td><td>MEDIUM</td><td>HIGH</td></tr>
<tr><td>Mean Value:</td><td>-0.5935</td><td>-0.2498</td><td>0.8435</td></tr>
<tr><td></td><td colspan="2">________________________</td><td>______</td></tr>
</table>

*means with a common line do not differ significantly at $p < 0.05$.

Time estimation standard deviation significantly increased from low to high and from medium to high loading, with $p < 0.05$. While there was a numerical increase in standard deviation from low to medium, it was not significant at the 0.05 level. The tapping regularity Duncan's test showed identical results to time estimation.

The only physiological measure to show sensitivity to load in the simple effects tests was respiration rate in breaths-per-minute. Duncan's comparisons for this measure revealed that the number of breath-per-minute significantly increased from low to high loading levels, but not from low to medium and from medium to high levels, at $p < 0.05$ (Table 8).

Finally, danger condition response times showed a strong monotonic increase as loading was increased (Figure 15). Differences between all pairs of means were statistically significant for this primary task measure, as revealed by the Duncan's test in Table 9.

The results of the multiple comparisons tests provided a basis for categorizing the various workload estimation techniques as to their relative sensitivity to changes in perceptual load (item 7 in Figure 8). The results of this categorization scheme are presented in the discussion section of this report.

TABLE 8

Results of Duncan's Multiple Range Analysis for the Respiration Rate Technique*

Load Level:	LOW	MEDIUM	HIGH
Mean Value:	-0.3517	-0.0847	0.4360

*means with a common line do not differ significantly at $p < 0.05$.

TABLE 9

Results of Duncan's Multiple Range Analysis for the Danger Condition Response Time Technique*

Load Level:	LOW	MEDIUM	HIGH
Mean Value:	-0.9192	-0.2558	1.1752

*means with a common line do not differ significantly at $p < 0.05$.

Intrusion Analysis

The objective of the intrusion analysis was to investigate the potential presence of undesirable, artificial variances in primary task performance that were attributable to the use of a particular workload estimation technique or its associated equipment. The experimental design for the intrusion analysis was structured to answer the initial and fundamental question: Did the workload estimation techniques differentially influence performance on four primary task measures: danger condition response time, control movements-per-second, pitch high-pass mean square, and roll high-pass mean square? As earlier discussed, all four primary task measures were obtained concurrently with the sensitivity dependent measure of "score" for each technique. For the intrusion analysis, "score" was deleted (because it was only applicable to the sensitivity analysis), and the four actual primary task values were used. Because there were no differences in scaling values between techniques, and furthermore because MANOVA does not require standardized scores, it was not necessary to convert the four primary task values to standard units.

Intrusion MANOVA. Due to the multiplicity of dependent measures, it was necessary to apply a multivariate analysis of variance procedure to the intrusion data set. With the

MANOVA, the Wilk's U criterion values were obtained for all independent effects. These values were subsequently converted to F -ratios, to facilitate testing for significance and interpretation of the data using common F tables. Conversion formulae appear in Rao (1965).

The results from the MANOVA shown in Table 10 are as follows. Load was the only independent effect to show significance (p = 0.0001). This effect simply demonstrated that the four primary task measures, as a group, were reliably affected by changes in load. This provided additional evidence that perceptual load was effectively manipulated in the experiment. That is, the main effect of load in the intrusion MANOVA reiterated the significant main effect of load found in the overall sensitivity ANOVA (see Table 4). The MANOVA load main effect, however, had no bearing on the interpretation of the intrusion analysis. If differential intrusion among techniques had indeed occurred, it would have been manifested as a significant technique effect or technique by load effect in the MANOVA. However, because neither a technique effect nor a technique by load effect were even close to statistical significance (p = 0.7923 and 0.2460, respectively), intrusion appeared not to have been a factor in this study.

TABLE 10

MANOVA Summary Table for Intrusion Analysis

Source	dv	df_H	df_E	U	F (converted)	p
Between-Subjects						
Technique (T)	4	7	40	0.5874	0.77	0.7923
Subjects (S)/T		(Error Term for T)				
Within-Subject						
Load (L)	4	2	80	0.0928	43.94	0.0001
L x T	4	14	80	0.4743	1.14	0.2460
L x S/T		(Error term for L, L x T)				

where: dv = number of dependent measures

df_H = degrees of freedom for treatment effect

df_E = degrees of freedom for error effect

U = Wilk's likelihood ratio statistic

$$= \frac{|E|}{|E + H|} \text{, where:}$$

$|E|$ = determinant of sum of square and cross-products matrix for error

$|E+H|$ = determinant of the sum of the sum of squares and cross-products matrix for error and the sum of squares and cross-products matrix for treatment.

Pilot Experience Level Analyses

As previously discussed, pilot experience level was used as a means for assigning pilots to technique conditions in this experiment. Six levels of experience were used. Drawing one subject from each level for each technique resulted in a cross-sectional representation of six experience levels for each technique. Because an equal number of experience levels appeared in each technique, it was possible to perform two additional, unplanned, post hoc analyses to investigate if there existed differences between the six levels of experience.

Experience level ANOVA using sensitivity data. First, the experience level investigation was conducted on the standardized sensitivity data set. All scores were collapsed across technique and a two-way ANOVA was performed (Table 11). Experience level was the fixed-effects, between-subjects variable while load was the fixed-effects, within-subject variable. The ANOVA revealed that experience level had no main effect on the standardized workload scores, F (5,42) = 1.54, p = 0.1973. There was also no significant interaction of experience level with load, F (10,84) = 0.62, p = 0.7890. As was expected, the load main effect was retained at the same significance level as in previous analyses, F (2,84) = 29.07, p = 0.0001.

TABLE 11

ANOVA Summary Table for Experience Level Analysis Using Standardized Sensitivity Data

Source	df	SS	F	p
Between-Subjects				
Experience Level (E)	5	10.7112	1.54	0.1973
Subjects (S)/E	42	58.3070		
Within-Subject				
Load (L)	2	26.2825	29.07	0.0001
L x E	10	2.8227	0.62	0.7890
L x S/E	84	37.9720		
Total	143	136.0954		

Experience level MANOVA using intrusion data. An experience level investigation was next performed using the four primary task dependent measures in the intrusion data set. This analysis was performed to determine whether experience level had an effect on the pilots' ability to perform the navigational control task and the danger condition task. All four measures were collapsed across technique, and a two-way MANOVA was performed. Again, the load main effect was retained, at $p = 0.0001$. Experience level was demonstrated to have no significant effect on the group of primary task measures, with $p = 0.8326$. Similarly, there was no interaction of load with experience level, with $p = 0.6048$. The MANOVA statistics are presented in Table 12.

TABLE 12

MANOVA Summary Table for Experience Level Analysis Using Primary Task Intrusion Data

<u>Source</u>	<u>dv</u>	<u>df_H</u>	<u>df_E</u>	<u>U</u>	<u>F</u> (converted)	<u>p</u>
<u>Between-Subjects</u>						
Experience Level (E)	4	5	42	0.7169	0.69	0.8326
Subjects (S)/E		(Error term for E)				
<u>Within-Subject</u>						
Load (L)	4	2	84	0.1088	41.13	0.0001
L x E	4	10	84	0.6514	0.92	0.6048
L x S/E		(Error term for L, L x E)				

where: dv = number of dependent measures

df_H = degrees of freedom for treatment effect

df_E = degrees of freedom for error effect

U = Wilk's likelihood ratio statistic

$$= \frac{|E|}{|E + H|}, \text{ where:}$$

|E| = determinant of sum of squares and cross-products for error

|E+H| = determinant of the sum of the sum of squares and cross-products matrix for error and the sum of squares and cross-products matrix for treatment

DISCUSSION AND CONCLUSIONS

The primary objective of this experiment was to determine empirically the sensitivity of eight workload estimation techniques to variation in flight-related perceptual load. Both overall and relative sensitivity were addressed. A secondary, but also important, objective was to determine which, if any, estimation techniques or associated equipment intruded on primary task performance. For either of these objectives to be met, it was of fundamental importance to manipulate load on a dimension which emphasized the nearly exclusive use of perceptual behaviors, as defined by Berliner et al. (1964). The main effect of load revealed by both the overall sensitivity ANOVA (Table 4) and by the overall intrusion MANOVA (Table 10) suggested that the procedure used for manipulating perceptual load, via a danger condition detection/identification flight-related task, was indeed effective. Furthermore, the load effect on the mean standardized scores when collapsed across techniques, was characterized by a distinct, monotonically-increasing function over the three load levels (Figure 9).

The presence of a significant technique by load interaction in the initial overall sensitivity ANOVA prompted the individual sensitivity investigation for each technique. These sensitivity investigations will be discussed first, followed by a discussion of the intrusion analysis results.

Sensitivity Conclusions

Of eight techniques investigated in this experiment, six demonstrated significant sensitivity to changes in perceptual load. In all cases, the individual sensitivity analyses were performed using parametric simple effects *F* -tests or nonparametric Friedman's Rank Sum tests. Techniques which demonstrated a statistically-significant load effect included:

1) Modified Cooper-Harper scale ratings,
2) Multi-Descriptor scale ratings,
3) time estimation standard deviation,
4) tapping regularity in pels,
5) respiration rate in breaths-per-minute, and
6) danger condition response time in seconds.

Techniques which did not demonstrate reliable sensitivity to load were:

1) pulse rate standard deviation and
2) control (aileron-elevator-rudder) movements-per-second

Relative sensitivity. Because six estimation techniques did indeed show sensitivity to load, it was desirable to establish a method of categorizing techniques as to their *relative* sensitivity to load. Hopefully, this categorization scheme would be of some utility in selecting a particular technique for a workload estimation application. The major basis for the categorization was the results of the Duncan's Multiple Range tests performed on each significant technique (Tables 6-9). These tests revealed the particular pairs of loading levels between which the techniques discriminated. If a technique showed a significant difference between all three loading levels, i.e. low-medium, medium-high, and low-high, then it was assigned to category I. Techniques which showed sensitivity within two possible pairs of loading levels constituted category II. Category III techniques showed significant sensitivity to only one pair, and category IV techniques yielded no sensitivity at all. Of course, the lower the category number, the more preferable the technique was as a perceptual workload estimator. The categorization scheme is shown in Table 13 and is referred to during the remainder of the sensitivity discussion. The particular pairs of loading levels between which each technique reliably discriminated are shown in the table. Also, the p -values attained in the simple effects F -tests or Friedman's test, where applied, are shown. First, the significant techniques in Table 13 will be discussed, followed by a brief discussion of the techniques which did not show significance.

TABLE 13

Relative Sensitivity Categorization of Techniques

Workload Estimation Technique	Sensitivity Category	Load Level Pairs Discriminated at $p<0.05$ (Duncan's)			Simple effects F p-value	Friedman p-value
		low-medium	medium-high	low-high		
Modified Cooper-Harper Scale	I	X	X	X	<0.005	<0.01
Multi-Descriptor Scale	II		X	X	<0.005	<0.025
Time Estimation Standard Deviation	II		X	X	<0.025	N/A
Tapping Regularity	II		X	X	<0.005	N/A
Pulse Rate Standard Deviation	IV				>0.10 (not sig.)	N/A
Respiration Rate	III			X	<0.05	N/A
Danger Condition Response Time	I	X	X	X	<0.005	N/A
Control Movements-per-second	IV				>0.10 (not sig.)	N/A

Modified Cooper-Harper scale. This rating scale showed strong overall sensitivity to a main effect of load in both parametric and nonparametric tests. Furthermore, the measure reliably showed significant rating differences for all possible pairs of load level means. As discussed earlier in the literature review, there is previous strong research support for the sensitivity of Cooper-Harper ratings in flight-related tasks where workload was manipulated on a psychomotor, rather than a perceptual, dimension (e.g., Shultz et al., 1970; Dick, et al., 1976; Waller, 1976; Connor, 1981). These psychomotor task results should be qualified in that conclusions were not usually drawn on the basis of nonparametric statistical procedures. In the present study, the modified version of the Cooper-Harper scale was found to generalize to tasks of a perceptual nature with no loss in sensitivity from the psychomotor domain. As a category I technique, the Modified Cooper-Harper scale certainly warrants consideration as an opinion workload estimator in situations where perceptual tasks are emphasized.

Multi-Descriptor scale. The overall sensitivity of this rating scale to load was also evident in both the parametric and nonparametric tests. However, after closer scrutiny, the scale revealed not to be as sensitive to perceptual load as the Modified Cooper-Harper scale. The

Multi-Descriptor scale, a category II technique, only reliably discriminated between medium-high loading and between low-high loading. The ratings, while not significant between low-medium levels, were monotonically increasing across the full range of load. While certainly possessing demonstrated utility as a workload estimation technique, the Multi-Descriptor scale needs further development and subsequent investigation to enable it to tap small differences in perceptual workload at lower levels.

Time estimation standard deviation. Another category II measure, time estimation standard deviation, showed a monotonic increase across the three loading levels, from low to medium to high. However, the increase was significant only from medium to high and from low to high loading. Significant results using time estimation standard deviation were also obtained by Connor (1981), in a simulator-based study where psychomotor load was changed using variable turbulence and aircraft stability. However, the change in standard deviation values across Connor's psychomotor load was not monotonic as for perceptual load. Others, e.g., NASA-Ames Research Center (1975) and Hart and McPherson (1976), have found time estimation variability estimates to be sensitive indicants of changes in psychomotor load. From the results of the present study, it appeared that as perceptual load was increased by increasing the rate and number of potential danger indications, subjects' time estimates become more variable in length. Perhaps during the high loading conditions, the increased interference of the danger indications caused subjects to adopt a "retrospective" rather than an "active" strategy in producing their estimates (see Hart et al., 1977). At any rate, the time estimation standard deviation measure shows considerable promise as a perceptual workload estimation technique.

Tapping regularity. The results for Michon's (1966) tapping regularity measure closely paralleled those of time estimation standard deviation. Also a category II measure, the tapping regularity measure showed a significant decrease in regularity (pels) from medium to high and from low to high levels. In all cases, the variability plot was a monotonically-increasing function. These findings are in general agreement with Michon's 1964 and 1966 results. In these studies, tapping regularity was found to discriminate between various loading levels on different types of laboratory primary tasks which emphasized both perceptual and motor behaviors. These experiments included such tasks as maze performance, screw sorting, letter detection, and Bourdon-test performance. The tapping regularity results in the present study followed those of the Multi-Descriptor ratings; Michon (1966) and Johannsen et al. (1976) also obtained agreement between tapping results and opinion ratings of task difficulty.

Respiration rate. The breaths-per-minute measure was the sole significant physiological measure in the experiment. It also was found to be the only category III technique in the experiment. Respiration rate discriminated only between low and high loading levels in this study, although a monotonic increase in respiration frequency was found across all loading levels. This finding suggests that respiration rate should probably be limited to applications in workload assessment situations where the loading levels of interest are known beforehand to be widespread. The significant results obtained for the effect of perceptual load on respiration rate in this study are in agreement with the previous results of Ettema and Zielhuis (1971) and Spyker et al. (1971). The former found that breathing rate increased with the amount of information handled per minute in a binary choice experiment. The latter found that respiration frequency showed a positive correlation with task difficulty on a two-axis aircraft control task. Other authors have reported no reliable results with frequency measures of breathing. For instance, Gaume and White (1975) found that respiration rate did not significantly increase as workload on a digital counter monitoring task increased. Connor (1981) obtained similar results in his simulated instrument landing task where psychomotor load was varied. Other pertinent research was discussed earlier in the literature review section of this report.

Danger condition response time. This primary task measure was quite sensitive to changes in perceptual loading. Response times increased significantly between all pairs of loadings, placing this measure in category I with the Modified Cooper-Harper scale. As the rate and number of danger conditions appearing during a flight was increased, the pilots required longer to detect and identify the conditions. The response time measure, which directly reflected pilot performance on the primary loading task in this experiment, certainly merits further application in situations where the criteria for performance on the task of interest is specific to response time.

Pulse rate standard deviation. The simple effects F -tests yielded no significance due to a perceptual load effect for this measure of cardiovascular activity. Therefore it is a category IV measure. There is some previous research evidence that pulse rate variability tended to decrease with increased loading in flight-related tasks emphasizing psychomotor behaviors (e.g., Auffret et al., 1967; Stackhouse, 1973; Opmeer and Krol, 1973). However, the results of the present study coincide with the most recent results available, (Connor, 1981), where heart rate variability did not reliably change as turbulence and aircraft stability were varied to change psychomotor load. Similarly, Hicks and Wierwille (1979) found no significant change in pulse rate variability as the center of pressure of crosswind gusts was moved rearward (increasing motor load) in a simulated driving task. Because of the variations in limb movement during different levels of a flight-related (or driving-related) loading task, it is difficult to determine whether any obtained physiological effects are due to physical effort or to mental effort. The earlier studies which demonstrated significant changes in heart rate variability may be subject to such a criticism. However, in the _perceptual_ study discussed herein, motor load was held as constant as possible, while perceptual load was varied considerably. Even so, pulse rate variability showed no change over loading conditions.

Control movements-per-second. As previously discussed, the frequency of control movements is a popular dependent measure in automobile driving and flight research (e.g., Macdonald and Hoffman, 1980; Dick et al., 1976). As a primary task measure in this study, the number of control movements did not reliably reflect changes in perceptual load, placing it in category IV. One other study which emphasized perceptual loading in a simulated flight task, using aileron and elevator inputs as measures was performed by Rolfe et al. (1974). The presence of oil pressure problems was used to increase workload in this study. The lack of significance of the control inputs measure obtained by Rolfe et al. was in agreement with the results of the present study. It appears that the frequency/activity-type measure may only prove useful for estimating task difficulty in primary tasks which have a large psychomotor component, and may not be generalizable to perceptual, cognitive, or other tasks. Even within the psychomotor realm, there are some conflicting results concerning control movement measures. In his instrument landing task, Connor (1981) found that control movements reliably discriminated between all levels of psychomotor load. Control movements were found to increase with increased turbulence and decreased pitch-axis stability. On the other hand, steering reversal (movement) rates have been found to both increase and decrease as driving task demands increase (Casali and Wierwille, 1980; Macdonald and Hoffman, 1980). In conclusion, based on the present study and the Rolfe et al. (1974) study, frequency measures of control movements are not promising as primary task measures of perceptual load. Furthermore, because of the nondirectionality of results with the control movements measure in psychomotor-loading research, the measure warrants further testing prior to application to critical workload assessment endeavors.

Intrusion Conclusions

The presence of an artificial change in primary task performance, solely due to the introduction of a workload estimation technique or associated apparatus is undesirable in workload assessment practice. As discussed previously, the mere presence of intrusion contaminates measurement because primary task performance may be altered spuriously; therefore, the measured workload level may not be indicative of workload required of the primary task. Furthermore, when used in simulation or in full-scale aircraft, highly intrusive techniques may degrade pilot control performance, causing safety problems.

As evidenced by the lack of a main effect of technique or an interaction of technique with load in the intrusion MANOVA (Table 10), significant differential intrusion did not occur in the present study. Specifically, it is concluded that no differential intrusion occurred among techniques on the four primary task dependent measures. If differential intrusion had indeed occurred, there would have likely been a substantial difference between the dependent measures obtained in the primary task technique cells and in some other technique cells, such as tapping regularity, which would have surfaced as a main effect of technique.

The intrusion results of this experiment are in general agreement with those obtained in other studies which specifically addressed intrusion. The aforementioned Hicks and Wierwille (1979) study investigated the sensitivity and intrusion of five estimation techniques (digit-shadowing secondary task, visual occlusion, rating scale, pulse rate variability, and primary task measures) in a simulated driving task. The only technique found intrusive was visual occlusion, which was not investigated in the present study. Connor (1981) also reported no intrusion on primary task performance in his simulated flight task using similar measures as in the present study, but with psychomotor loading.

Conversely, Wierwille and Gutmann (1978) found significant intrusion of a digit-shadowing task, but only at high levels of psychomotor loading in a driving simulator. Loading was varied in a factorial design with two levels of vehicle response, two levels of steering ratio, and two levels of windgust disturbance. There are two plausible explanations for the lack of secondary task intrusion in the present study and the presence of it in the Wierwille and Gutmann study. First, the digit-shadowing secondary task in the Wierwille and Gutmann study required constant, dedicated visual attention for accurate performance, as did the driving task (the primary task) itself. Therefore, a distinct competition between tasks occurred on the visual input channel. In the present study, however, the primary task relied largely on visual input but the secondary tasks of time estimation and tapping regularity required auditory input and no input, respectively. In terms of input requirements, the primary and secondary tasks appeared to be mutually-noninterfering in the present study. A second explanation centers around the subject population in both studies. The subjects in the Wierwille and Gutmann study (licensed drivers) may have been more likely to sacrifice primary task performance than the pilots in the present study. Pilots are more rigorously trained to maintain accurate flight parameters at all times, and, as borne out in the present study, do not permit subsidiary, non-flight related tasks to degrade their performance on the primary flight task.

Recommendations

The initial objectives of the research described herein were, to a large degree, met. The relative sensitivity and intrusion of eight different mental workload estimation techniques were investigated in a simulated flight task emphasizing perceptual load. No differential intrusion was revealed but six of the eight techniques (at least one from each major category in Table 2) did show sensitivity to changes in perceptual load. All significant techniques displayed monotonic changes with respect to load.

Both rating scale measures proved to be quite useful. The results of this study reiterate those of others, indicating that with highly-trained populations such as pilots, rating scales are sensitive measurement instruments. Inexpensive, unobtrusive, and easily administered, rating scales are transferable to full-scale aircraft and to a wide range of tasks. This is evident with the Cooper-Harper scale, which showed category I (highest) sensitivity after being modified for assessing general piloting problems other than aircraft controllability. Similarly, the Multi-Descriptor scale could be modified for quite different tasks by substituting new descriptors based on behavior parameters exhibited in the task of interest.

The secondary task measures (time estimation and tapping regularity) also exhibited considerable sensitivity to perceptual load. However, these measures do not lend themselves to full-scale aircraft application quite as readily as the rating scales do. While no intrusion was found in the present study, the addition of unfamiliar secondary task hardware to the aircraft cockpit is a questionable practice. Furthermore, intrusion may occur if workload conditions are spread further apart than in this study. The time estimation technique, in its present form, also requires an audio input which may preclude its use in actual aircraft.

From the results discussed previously, respiration rate appears to be sensitive to widespread changes in perceptual load. However, when comparing the results from this study with others, it is apparent that respiration rate is a highly task-specific measure. Furthermore, the transducers and sensing apparatus required for obtaining the measure may not be feasible for other than simulator-based assessment studies.

Finally, this study demonstrated that primary task measures are also quite task-specific and therefore must be selected with task objectives in mind. Control input frequency measures were not affected by incident perceptual load. However, the response time measure, which directly reflected performance on the detection/identification aspect of the primary flight task, was a most discriminating measure. Of course, for tasks which emphasize other behavioral dimensions, such as psychomotor or mediational, different primary task measures need to be devised.

This research represented a first attempt at examining the sensitivity and intrusion of workload estimation techniques in a flight task emphasizing perceptual load. In a similar manner, the application of these techniques to other tasks, such as mediational or communicative, warrants future investigation. Subsequently, the stability of the measures over time (reliability) needs research attention.

REFERENCES

Andreassi, J. L. Psychophysiology. New York: Oxford University Press, 1980.

Auffret, R., Seris, H., Berthoz, A. and Fatras, B. Estimate of the perceptive load by variability of rate of heartbeat: Application to a piloting task. Le Travail Humain, 1967, 80, 309-310.

Baker, D. L. and Intano, G. P. Helicopter yaw axis augmentation investigation - CDG-PFH-4. Randolph AFB, Texas: USAF Instrument Flight Center, IFC Test Plan 74-11, December, 1974.

Barr, A. J., Goodnight, J. H., Sall, J. P., and Helwig, J. T. A user's guide to SAS-79. Raleigh, N.C.: SAS Institute, 1979.

Bateman, S. C., Goldsmith R. Jackson, K. F., Ruffell-Smith, H. P., and Mottodes, V. S. Heart rate of training captains engaged in different activities. Aerospace Medicine, 1970, 41, 425-429.

Berliner, C., Angell, D., and Shearer, D. J. Behaviors, measures, and instruments for performance evaluation in simulated environments. Paper presented at the Symposium and Workshop on the Quantification of Human Performance, Albuquerque, New Mexico, August, 1964.

Boyce, P. R. Sinus arrhythmia as a measure of mental load. Ergonomics, 1974, 17, 177-183.

Butterbaugh, L. C. Crew workload-technology review and problem assessment. Wright-Patterson Air Force Base, Ohio: Flight Dynamics Laboratory, Technical Memorandum, AFFDL-TM-78-74-FGR, January, 1978.

Cantrell, G. K. and Hartman, B. O. Application of time and workload analysis techniques to transport flyers. Brooks AFB, Texas: USAF School of Aviation Medicine, Technical Report SAM-TR-67-71, August, 1967.

Casali, J. G. and Wierwille, W. W. Investigation of the effects of various design alternatives on moving-base driving simulator discomfort. Human Factors, 1980, 22, 741-756.

Chiles, W. D. Objective methods. In A. H. Roscoe (Ed.) Assessing pilot workload. AGARD-AG-233, February, 1978, 54-77.

Chiles, W. D. and Alluisi, E. A. On the specification of operator or occupational workload with performance-measurement methods. Human Factors, 1979, 21, 515-528.

Clement, W. Annotated bibliography of procedures which assess primary task performance in some manner as the basic element of a workload measurement procedure. Hawthorne, Calif.: Systems Technology, Inc., Technical Report No. 1104-2, January, 1978.

Connor, S. A. A comparison of pilot workload assessment techniques using a psychomotor task in a moving-base aircraft simulator. Unpublished master's thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1981.

Cooper, G. E. and Harper, R. P., Jr. The use of pilot rating in the evaluation of aircraft handling qualities. Moffett Field, California: National Aeronautics and Space Administration, Ames Research Center, NASA TN-D-5153, April 1969.

Crabtree, M. S. Human factors evaluation of several control system configurations, including workload sharing with force wheel steering during approach and flare. Wright-Patterson AFB, Ohio: USAF Flight Dynamics Laboratory, AFFDL-TR-75-43, April, 1975.

Dick, A. O., Brown, J. L. and Bailey, G. Statistical evaluation of control inputs and eye movements in the use of instrument clusters during aircraft landing. Rochester, New York: University of Rochester, Center for Visual Science, Technical Report 4-76, 1976.

Donnell, M. L. The application of decision-analytic techniques to the test and evaluation phase of the acquisition of a major air system: Phase III. McLean, Virginia: Decisions and Designs, Technical Report PR79-6-91, May, 1979.

Donnell, M. L. and O'Connor, M. F. The application of decision analytic techniques to the test and evaluation phase of the acquisition of a major air system: Phase II. McLean, Virginia: Decisions and Designs, Technical Report TR 78-3-25, April, 1978.

Duncan, D. B. t-tests and intervals for comparisons suggested by the data. Biometrics, 1975, 31, 339-359.

Dyer, R. F., Matthews, J. J., Wright, C. E., and Yurdawitch, K. L. Questionnaire construction manual. Fort Hood, Texas: U.S. Army Research Institute for the Behavioral and Social Sciences, Field Unit, Technical Report p-77-1, July, 1976.

Edwards, A. L. Techniques of attitude scale construction. New York: Appleton Century Crofts, 1957.

Edwards, L. R., Pilette, S. S., Biggs, B. E., and Martinek, H. The effect of workload on performance of operators monitoring unattended ground sensors. Alexandria, Virginia: U.S. Army Research Institute for the Behavioral and Social Sciences, Technical Paper 321, September, 1978.

Ellis, G. A. Subjective assessment. In A. H. Roscoe (Ed.) Assessing pilot workload. AGARD-AG-233, February, 1978, 11-22.

Ettema, J. H. and Zielhuis, R. L. Physiological parameters of mental load. Ergonomics, 1971, 14, 137-144.

Gartner, W. B. and Murphy, M. R. Pilot workload and fatigue: A critical survey of concepts and assessment techniques. Moffett Field, California: National Aeronautical and Space Administration Ames Research Center, NASA TN D-8365, November, 1976.

Gaume, J. G. and White, R. T. Mental workload assessment, II. Physiological correlates of mental workload: Report of three preliminary laboratory tests. St. Louis, Missouri: McDonnell Douglas Corporation, Report MDC J7023/01, December, 1975.

Geer, C. W. User's guide for the test and evaluation sections of MIL-H-46855. Seattle, Washington: Boeing Aerospace Company, Technical Report D194-10006-1, June, 1977.

Geiselhart, R., Schiffler, R. J. and Ivey, L. J. A study of task loading using a three-man crew on a KC-135 aircraft. Wright-Patterson AFB, Ohio: Aeronautical Systems Division, ASD-TR-76-19, October, 1976.

Gerathewohl, S. J. Definition and measurement of perceptual and mental workload in aircrews and operators of Air Force weapon systems: A status report. In B. O. Hartman (Ed.) _Higher Mental Functioning in Operational Environments_, AGARD Conference Proceedings No. 181, April, 1976, C1-1 - C2-7.

Gunning, D. Time estimation as a technique to measure workload. _Proceedings of the 22nd Annual Meeting of the Human Factors Society_, Detroit, Michigan, October 16-19, 1978, 41-45.

Hall, T. J., Passey, G. E. and Meighan, T. W. Peformance of vigilance and monitoring tasks as a function of workload. Wright-Patterson AFB, Ohio: Aerospace Medical Research Laboratories, AMRL-TR-65-22, March, 1965.

Hart, S. G. A cognitive model of time perception. Paper presented at the 56th Annual Meeting of the Western Psychological Association, Los Angeles, California, April, 1976.

Hart, S. G. Pilot workload during final approach in congested airspace. _Proceedings of the 1978 IEEE Conference on Decision and Control_, San Diego, California, January 10-12, 1979, 1345-1349.

Hart, S. G. and Bird, K. L. Effects of feedback, counting, and tracking on time estimation and production. Paper presented at the 16th Annual Conference on Manual Control, Massachusetts Institute of Technology, Cambridge, Massachusetts, May, 1980.

Hart, S. G., Childress, M.E., and Hauser, J. W. Individual definitions of the term "workload." Symposium on Psychology in the Department of Defense, U.S. Air Force Academy, Colorado, 1982.

Hart, S. G. and McPherson, D. Airline pilot time estimation during concurrent activity including simulated flight. Paper presented at the 47th Annual Meeting of the Aerospace Medical Association, Bal Harbour, Florida, May, 1976.

Hart, S. G., McPherson, D., Kreifeldt, J. and Wempe, T. E. Multiple curved descending approaches and the air traffic control problem. Moffett Field, California: National Aeronautical and Space Administration, Ames Research Center, NASA TM-78, 430, August, 1977.

Hart, S. G., McPherson, D., and Loomis, L. L. Time estimation as a secondary task to measure workload: summary of research. Proceedings of the Fourteenth Annual Conference on Manual Control, April 25-27, 1978. University of Southern California, Los Angeles, 693-712. (NASA Conference publication 2060, Ames Research Center, Moffett Field, California.)

Hart, S. G. and Simpson, C. A. Effects of linguistic redundancy on synthesized cockpit warning message comprehension and concurrent time estimation. Proceedings of the 12th Annual NASA-University Conference on Manual Control, University of Illinois, May, 1976, 309-321. (NASA TM X-73 170).

Hartman, B. O., and McKenzie, R. E. (Eds.) Survey of methods to assess workload. AGARD-AG-246, August, 1979.

Hawkins, H. L., Church, M., and deLemos, S. Time-sharing is not a unitary ability. Eugene, Oregon: University of

Oregon, Center for Cognitive and Perceptual Research, Technical Report No. 2, June 30, 1978.

Helander, M. G. Physiological reactions of drivers as indicators of road traffic demand. In Driver performance studies: Transportation Research Record 530. Washington, D.C.: U.S. Transportation Research Board, Technical Report TRB/TRR-530, 1975, 1-17.

Helm, W. R. Human factors test and evaluation, functional description inventory as a test and evaluation tool development and initial validation study. Volume I and II. Patuxent River, Maryland: U.S. Naval Air Test Center, SY-77R-75, September, 1975.

Helm, W. R. Human factors evaluation of model P-3C UPDATE I airplane: Third interim report. Patuxent River, Maryland: U.S. Naval Air Test Center, SY-122R-75, February 1976.

Hicks, T. G. and Wierwille, W. W. Comparison of five mental workload assessment procedures in a moving-base driving simulator. Human Factors, 1979, 21, 129-143.

Hollander, M. and Wolfe, D. A. Nonparametric statistical methods. New York: Wiley, 1973.

Jahns, D. W. Operator workload: What is it and how should it be measured? In K. D. Cross and J. J. McGrath (Eds.) Crew system design. Santa Barbara, California: Anacapa Sciences, July, 1973(a).

Jahns, D. W. A concept of operator workload in manual vehicle operations. Meckenheim, Germany: Forschungsinstitue fur Anthropotechnik, Report No. 14, 1973(b).

Jex, H. R. and Allen, R. W. Research on a new human dynamic response test battery. Part II. Psychophysiological correlates. Proceedings of the 6th Annual NASA-University Conference on Manual Control, Wright-Patterson AFB, Ohio, April, 1970, 767-777.

Johannsen, G., Pfendler, C. and Stein, W. Human performance and workload in simulated landing-approaches with autopilot-failures. In T. B. Sheridan and G. Johannsen (Eds.) Monitoring behavior and supervisory control. New York: Plenum, 1976, 83-95.

Kahneman, D. Attention and effort. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1973.

Kalsbeek, J. W. H. Sinus arrhythmia and the dual task method in measuring mental load. In W. T. Singleton, Fox, J. G., and Whitfield, D. (Eds.) Measurement of Man at Work. London: Taylor and Francis, 1973, 101-113.

Kalsbeek, J. W. H. and Sykes, R. N. Objective measurement of mental load. Acta Psychologica, 1967, 27, 253-261.

Kelley, C. R. and Wargo, M. J. Cross-adaptive operator loading tasks. Human Factors, 1967, 9, 395-404.

Knowles, W. B. Operator loading tasks. Human Factors, 1963, 5, 155-161.

Kreifeldt, J., Parkin, L. and Rothschild, P. Implications of a mixture of aircraft with and without traffic situation displays for air traffice management. Proceedings of the 12th Annual NASA-University Conference on Manual Control, University of Illinois, May, 1976, 179-200. (NASA TM X-73 170).

Laurell, H. and Lisper, H. O. A validation of subsidiary reaction time against detection of roadside obstacles during prolonged driving. Ergonomics, 1978, 21, 81-88.

Laurig, W. and Phillip, U. Changes in the pulse frequency rhythm in relation to the workload. (veranderungen der Pulsfrequenzarrhythmie in Abhangigkeit von der Arbeitsschwere). Arbeitsmedizin Sozialmedizin, Arbeitshygiene, 1970, 5, 184-188. (Royal Aircraft Establishment Library Translation 1586).

Lebacqz, J. V. and Aiken, E. W. A flight investigation of control, display, and guidance requirements for decelerating descending VTOL instrument transitions using the X-22A variable stability aircraft. Volume I. Buffalo, New York: Calspan Corporation, AK-5336-F-1, September, 1975.

Levine, J. M., Ogden, G. D. and Eisner, E. J. Measurement of workload by secondary tasks: A review and annotated bibliography. Washington, D.C.: Advanced Research Resources Organization, Contract No. NAS2-9637, January, 1978.

Lisper, H. O., Laurell, H. and Stening, G. Effects of experience of the driver on heart-rate, respiration-rate, and subsidiary reaction time in a three hour continuous driving task. Ergonomics, 1973, 16, 501-506.

Luczak, H. and Laurig, W. An analysis of heart rate variability. Ergonomics, 1973, 16, 85-97.

Macdonald, W. A. and Hoffman, E. R. Review of relationships between steering wheel reversal rate and driving task demand. Human Factors, 1980, 22, 733-739.

Madero, R. P., Sexton, G. A., Gunning, D., and Moss, R. Total aircrew workload study for the AMST, Volume I, Results. Wright-Patterson Air Force Base, Ohio: Flight Dynamics Laboratory, Final Technical Report, AFFDL-TR-79-3080, Vol. I, February, 1979.

Markeiwicz, L., Koradecka, D., and Konarska, M. Retention of selected physiological indicators in pilots in the course of agricultural flights. Washington, D.C.: National Aeronautics and Space Administration, Technical Translation TT-F-17441, August, 1977.

McCauley, M. E., Kennedy, R. S., and Bittner, A. C., Jr. Development of performance evaluation tests for environmental research (PETER): Time estimation test. Proceedings of the 23rd Annual Meeting of the Human Factors Society, Boston, Massachusetts, October 29-November 1, 1979, 513-517.

McLean, J. R. and Hoffmann, E. R. Steering reversals as a measure of driver performance and steering task difficulty. Human Factors, 1975, 17, 248-256.

Michon, J. A. A note on the measurement of perceptual motor load. Ergonomics, 1964, 7, 461-463.

Michon, J. A. Tapping regularity as a measure of perceptual motor load. Ergonomics, 1966, 9, 401-412.

Michon, J. A. and van Doorne, H. Equipment note: A semi-portable apparatus for the measurement of perceptual motor load. Ergonomics, 1967, 10, 67-72.

Mobbs, R. F., David, G. C. and Thomas, J. M. An evaluation of the use of heart rate irregularity as a measure of mental workload in the steel industry. London, England: British Steel Corporation, BISRA, OR/HF/25/71, August, 1971.

Moray, N. Models and measures of mental workload. In N. Moray (Ed.) Mental workload: its theory and measurement. New York: Plenum Press, 1979, 13-21.

Mulder, G. and Mulder-Hajoinides van der Meulen, W. R. E. H. Mental load and the measurement of heart rate variability. Ergonomics, 1973, 16, 69-83.

Murphy, J. V. and Gurman, B. S. The integrated cockpit procedure for identifying control and display requirements of aircraft in advanced time periods. Proceedings of the AGARD Conference on Guidance and Control Displays, AGARD-CP-96, 4-1 - 4-7.

NASA-Ames Research Center. Secondary task for full-flight simulation incorporating tasks that commonly cause pilot error: Time estimation. Moffett Field, California. NASA-TM-X-74153, October, 1975.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

O'Connor, M. F. and Buede, B. M. The application of decision analytic techniques to the test and evaluation phase of the acquisition of a major air system. McLean, Virginia: Decisions and Designs, Technical Report 77-3, April, 1977.

O'Connell, J. T. NASA CH-47B flight evaluation of helicopter/VSTOL control/display parameters. NATC Trip Report, September, 1978.

Ogden, G. D., Levine, J. M., and Eisner, E. J. Measurement of workload by secondary tasks. *Human Factors*, 1979, *21*, 529-548.

Opmeer, C. H. J. M. and Krol, J. P. Towards an objective assessment of cockpit workload. I-physiological variables during different flight phases. *Aerospace Medicine*, 1973, *44*, 527-532.

Perelli, L. P. Physiologic aspects of workload/fatigue/stress. In Hartman and McKenzie (Eds.) *Survey of Methods to Assess Workload*. AGARD-AG-246, August, 1979, 13-18.

Philipp, U., Reiche, D. and Kirchner, J. H. The use of subjective rating. *Ergonomics*, 1971, *14*, 611-616.

Price, D. L. The effects of certain gimbal orders on target acquisition and workload. *Human Factors*, 1975, *17*, 571-576.

Rao, C. R. *Linear statistical inference and its applications*. New York: Wiley, 1965.

Reising, J. M. The definition and measurement of pilot workload. Wright-Patterson AFB, Ohio: USAF Flight Dynamics Laboratory, AFFDL-TM-72-4-FGR, February, 1972.

Rohmert, W. Determination of stress and strain of air traffic control officers. *Proceedings of the AGARD Conference on Methods to Assess Workload*, AGARD-CPP-216, April, 1977, A6-1 - A6-8.

Rohmert, W., Laurig, W., Philipp, U. and Luczak, H. Heart rate variability and work-load measurement. Ergonomics, 1973, 16, 33-44.

Rolfe, J. M. The secondary task as a measure of mental load. In W. T. Singleton, J. G. Fox, and D. Whitfield (Eds.) Measurement of Man at Work. London: Taylor and Francis, 1973, 135-148.

Rolfe, J. M. The measurement of human response in man-vehicle control situations. In T. B. Sheridan and G. Johannsen (Eds.) Monitoring behavior and supervisory control. New York: Plenum Press, 1976, 125-137.

Rolfe, J. M., Chappelow, J. W., Evans, R. L., Lindsay, S. J. E. and Browning, A. C. Evaluating measures of workload using a flight simulator. Proceedings of the AGARD Conference on Simulation and Study of High Workload Operations, AGARD-CP-146, April, 1974, A4-1 - A4-13.

Rolfe, J. M. and Lindsay, S. J. E. Flight deck environment and pilot workload: Biological measures of workload. Applied Ergonomics, 1973, 4, 199-206.

Roscoe, A. H. Heart rate monitoring of pilots during steep-gradient approaches. Aviation, Space, and Environmental Medicine, 1975, 46, 1410-1413.

Roscoe, A. H. Use of pilot heart rate measurement in flight evaluation. Aviation, Space and Environmental Medicine, 1976, 47, 86-90.

Roscoe, A. H. (Ed.) Assessing pilot workload. AGARD-AG-233, February, 1978 (a).

Roscoe, A. H. Physiological methods. In A. H. Roscoe (Ed.) Assessing pilot workload. AGARD-AG-233, February, 1978, 23-51.(b)

Sanders, M. G., Burden, R. T., Jr., Simmons, R. R., Lees, M. A., and Kimball, K. A. An evaluation of perceptual-motor workload during a helicopter hover maneuver. Ft. Rucker, Alabama: U. S. Army Aeromedical Research Laboratory, 78-14, May, 1978.

Sanderson, J. Aviation fundamentals. Denver: Author, 1978.

Schiflett, S. G. Operator workload: An annotated bibliography. Patuxent River, Maryland: US Naval Air Test Center, SY-257R-76, December, 1976.

Schultz, W. C., Newell, F. D. and Whitbeck, R. F. A study of relationships between aircraft system performance and pilot ratings. Proceedings of the 6th Annual NASA-University Conference on Manual Control, Wright-Patterson AFB, Ohio, April 7-9, 1970, 339-340.

Senders, J. W. The human operator as a monitor and controller of multidegree of freedom systems. IEEE Transactions on Human Factors in Electronics, 1964, HFE-5, 2-5.

Senders, J. W. The estimation of operator workload in complex systems. In K. B. Degreene (Ed.) Systems Psychology. New York: McGraw-Hill, 1970, 207-216.

Siegel, A. I. and Wolfe, J. J. Man-machine simulation models: Psychosocial and performance interaction. New York: Wiley, 1969.

Simpson, C. A. and Hart, S. G. Required attention for synthesized speech perception for two levels of linguistic redundancy. Paper presented at the 93rd meeting of the Acoustical Society of America, State College, Pennsylvania, June 7-10, 1977.

Singer (Link Division). Link general aviation trainer, GAT-1, operation and maintenance manual. Binghamton, N. Y., Author, 1973.

Smit, J. and Wewerinke, P. H. An analysis of helicopter pilot control behavior and workload during instrument flying tasks. In AGARD Operational Helicopter Aviation Medicine. AGARD-CP-255, May, 1978, 30-1 - 31-11.

Soliday, S. M. Effects of task loading on pilot performance during simulated low-altitude high-speed flight. Fort Eustis, Virginia: U. S. Army Transportation Research Center, USATRECOM 64-69, February, 1965.

Soutendam, J. Instruments and methodology for the assessment of physiological cost of performance in stressful continuous operations - the air traffic services tower environment. Proceedings of the AGARD Conference on Methods to Assess Workload, AGARD-CPP-216, April, 1977, A7-1 - A7-32.

Spyker, D. A., Stackhouse, S. P., Khalafalla, A. S. and McLane, R. C. Development of techniques for measuring pilot workload. Washington, D.C.: National Aeronautics and Space Administration, Contractors' Report NASA CR-1888, November, 1971.

Stackhouse, S. Workload evaluation of LLNO display. Minneapolis, Minnesota: Honeywell, 7201-3408, October, 1973.

Stackhouse, S. P. The measurement of pilot workload in manual control systems. Minneapolis, Minnesota: Honeywell, Inc., F0398 FR1, January, 1976.

Steininger, K. Subjective ratings of flying qualities and pilot workload in the operation of a short haul jet transport aircraft. Proceedings of AGARD Conference on Studies on Pilot Workload, AGARD-CPP-217, April, 1977, B11-1.

Stern, R. M., Ray, W. J. and Davis, C. M. Psychophysiological recording. New York: Oxford University Press, 1980.

Sun, P. B., Keane, W. P. and Stackhouse, S. P. The measurement of pilot workload in manual control systems. Proceedings of Aviation Electronics Symposium, Fort Monmouth, New Jersey, April, 1976.

Trumbo, D. and Noble, M. Response uncertainty in dual-task performance. Organizational Behavior and Human Performance, 1972, 7, 203-215.

Ursin, H. and Ursin, R. Physiological indicators of mental load. In N. Moray (Ed.) Mental workload: its theory and measurement. New York: Plenum Press, 1979, 340-365.

Waller, M. C. An investigation of correlation between pilot scanning behavior and workload using stepwise regression analysis. Hampton, Virginia: NASA Langley Research Center, NASA TM X-3344, March, 1976.

Whitaker, L. A., Dual-task interference as a function of cognitive load processing. Acta Psychologica, 1979, 43, 71-84.

Wickens, C. D. The effect of time sharing on the performance of information processing tasks: A feedback control analysis. Ann Arbor, Michigan: The University of Michigan, Human Performance Center, Technical Report No. 51, August, 1974.

Wickens, C. D. and Tsang, P. Attention allocation in dynamic environments. University of Illinois (Urbana-Champaign) Engineering Psychology Research Laboratory, Technical Report EPL-79-3/AFOSR-79-3, June, 1979.

Wierwille, W. W. Physiological measures of aircrew mental workload. Human Factors, 1979, 21, 575-593.

Wierwille, W. W. and Gutmann, J. C. Comparison of primary and secondary task measures as a function of simulated vehicle dynamics and driving conditions. Human Factors, 1978, 20, 233-244.

Wierwille, W. W. and Williges, R. C. Survey and analysis of operator workload assessment techniques. Blacksburg, Virginia: Systemetrics, Inc. Report No. S-78-101, September, 1978.

Wierwille, W. W. and Williges, B. H. An annotated bibliography on operator mental workload assessment. Patuxent River, Maryland: Naval Air Test Center, Technical Report SY-27R-80, March, 1980.

Wierwille, W. W., Williges, R. C., and Schiflett, S. G. Aircrew workload assessment techniques. In B. O. Hartman and R. E. McKenzie (Eds.) Survey of methods to assess workload. AGARD-AG-246, August, 1979, 19-53.

Williges, R. C. and Wierwille, W. W. Behavioral measures of aircrew mental workload. Human Factors, 1979, 21, 549-574.

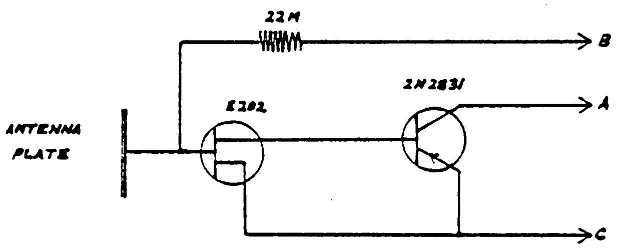
Winer, B. J. Statistical principles in experimental design, second edition. New York: McGraw-Hill, 1971.

Zwaga, H. J. G. Psychophysiological reactions to mental tasks: Effort or stress? Ergonomics, 1973, 16, 61-67.

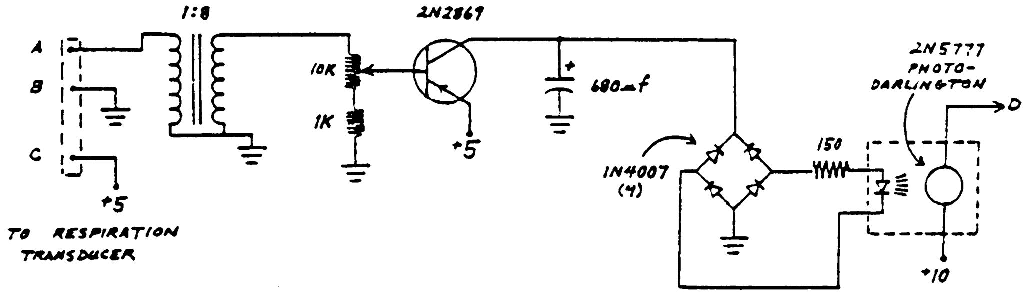
APPENDIX A

RESPIRATION MONITORING EQUIPMENT CIRCUIT DIAGRAMS

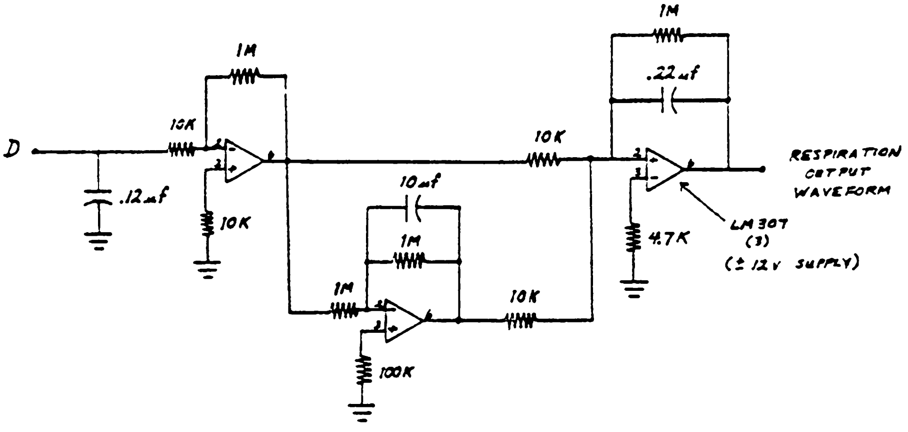
RESPIRATION TRANSDUCER:



DETECTOR:

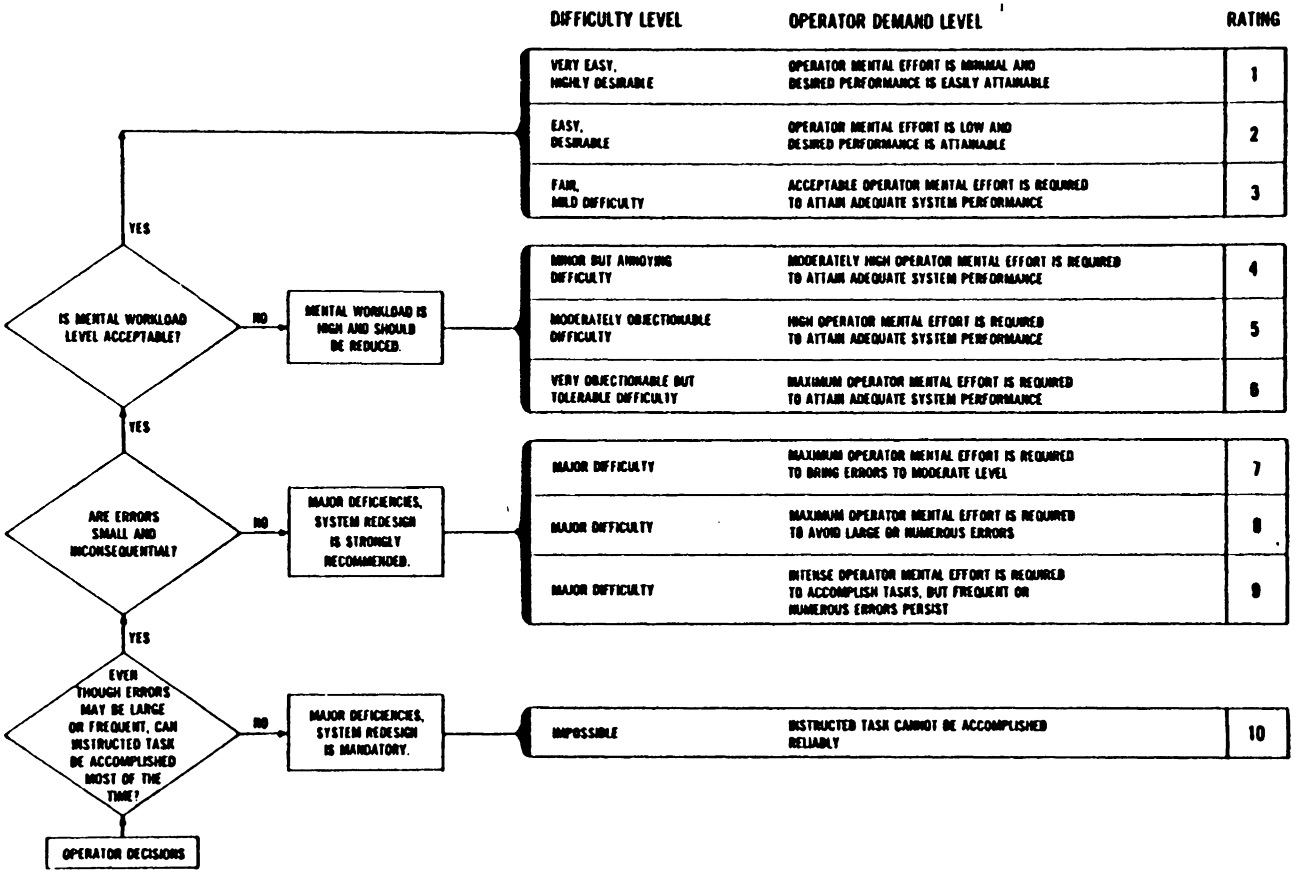


AMPLIFIER-SIGNAL CONDITIONER:



APPENDIX B

MODIFIED COOPER-HARPER RATING SCALE


IS MENTAL WORKLOAD LEVEL ACCEPTABLE?
YES
NO
MENTAL WORKLOAD IS HIGH AND SHOULD BE REDUCED.
YES
ARE ERRORS SMALL AND INCONSEQUENTIAL?
NO
MAJOR DEFICIENCIES, SYSTEM REDESIGN IS STRONGLY RECOMMENDED.
YES
EVEN THOUGH ERRORS MAY BE LARGE OR FREQUENT, CAN INSTRUCTED TASK BE ACCOMPLISHED MOST OF THE TIME?
NO
MAJOR DEFICIENCIES, SYSTEM REDESIGN IS MANDATORY.
OPERATOR DECISIONS

DIFFICULTY LEVEL	OPERATOR DEMAND LEVEL	RATING
VERY EASY, HIGHLY DESIRABLE	OPERATOR MENTAL EFFORT IS MINIMAL AND DESIRED PERFORMANCE IS EASILY ATTAINABLE	1
EASY, DESIRABLE	OPERATOR MENTAL EFFORT IS LOW AND DESIRED PERFORMANCE IS ATTAINABLE	2
FAIR, MILD DIFFICULTY	ACCEPTABLE OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE	3
MINOR BUT ANNOYING DIFFICULTY	MODERATELY HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE	4
MODERATELY OBJECTIONABLE DIFFICULTY	HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE	5
VERY OBJECTIONABLE BUT TOLERABLE DIFFICULTY	MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE	6
MAJOR DIFFICULTY	MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO BRING ERRORS TO MODERATE LEVEL	7
MAJOR DIFFICULTY	MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO AVOID LARGE OR NUMEROUS ERRORS	8
MAJOR DIFFICULTY	INTENSE OPERATOR MENTAL EFFORT IS REQUIRED TO ACCOMPLISH TASKS, BUT FREQUENT OR NUMEROUS ERRORS PERSIST	9
IMPOSSIBLE	INSTRUCTED TASK CANNOT BE ACCOMPLISHED RELIABLY	10

APPENDIX C

MODIFIED COOPER-HARPER RATING SCALE INSTRUCTIONS

Overview

After each of the following flights, you will be asked to give a rating on the Modified Cooper-Harper Scale for workload. This rating scale and important definitions for using the scale are shown on the sample which I have given you. Before you make any flights, we will review:

1. The definitions of the terms used in the scale,
2. The steps you should follow in making your ratings on the scale, and
3. How you should think of the ratings.

If you have any questions as we review these points please ask me.

Important Definitions

To understand and use the modified Cooper/Harper scale properly, it is important that you understand the terms used on the scale and how they apply in the context of this experiment.

First, "primary task" is the flight task you have been assigned to perform in this experiment. It includes flying the aircraft within specified levels of accuracy and detecting and identifying danger conditions that will be presented on the engine, fuel, and carburetor icing instruments. It will be described in more detail later.

Second, the operator in this situation is you. Because the scale can be used in different situations, the person performing the ratings is called an operator. You will be operating the system and then using the rating scale to quantify your experience.

Third, the system is the complete group of equipment you will be using in performing the primary task. Together you and the system make up the operator/system. (For the present experiment, the system is composed of the aircraft simulator, its instruments, controls, and the pushbutton panel for identifying danger conditions).

Fourth, errors include any of the following: mistakes, incorrect actions or responses, blunders, omissions, and incompletions. In other words, errors are any appreciable deviation from desired operator/ system performance.

Finally, mental workload is the integrated mental effort required to perform the primary task. It includes such factors as level of attention, depth of thinking, and level of concentration required by the primary task.

Rating Scale Steps

On the Modified Cooper-Harper scale you will notice that there is a series of decisions which follow a predetermined logical sequence. This logic sequence is designed to help you make more consistent and accurate ratings. Thus, you should follow the logic sequence on the scale for each of your ratings in this experiment.

The steps which you will follow in using the rating scale logic are as follows:

1. First you will decide if the primary task can be accomplished most of the time; if not, then your rating is a 10 and you should circle the 10 on the rating scale.
2. Second, you will decide if adequate performance is attainable. Adequate performance means that the errors are small and inconsequential in performing the primary task. Adequate performance will be defined in the set of instructions you will be given later for the flight task. If they are not, then there are major deficiencies in the system and you should proceed to the right. By reading the descriptions associated with the numbers 7, 8, and 9, you should be able to select the one that best describes the situation you have experienced. You would then circle the most appropriate number.
3. If adequate performance _is_ attainable your next decision is whether or not your mental workload for the primary task is tolerable. If it is not tolerable, you should select a rating of 4, 5, or 6. One of these three ratings should describe the situation you have experienced, and you would circle the most appropriate number.

4. If mental workload is tolerable, you should then move to one of the top three descriptions on the scale. You would read and carefully select the rating 1, 2, or 3 based on the corresponding description that best describes the situation you have experienced. You would circle the most appropriate number.

Remember you are to circle only one number, and the number should be arrived at by following the logic of the scale. You should always begin at the lower level and follow the logic path until you have decided on a rating. In particular, do not skip any steps in the logic. Otherwise your rating may not be valid and reliable.

How You Should Think of the Rating

Before you begin making ratings there are several points that need to be emphasized . First, be sure to try to perform the primary task as instructed and make all your evaluations within the context of the primary task. Try to maintain adequate performance as specified for your task.

Second, the rating scale is not a test of your personal skill. On all of your ratings, you will be evaluating the system for a general user population, not yourself. You may assume you are an experienced member of that population. You should make the assumption that problems you encounter are not problems you created. They are problems created by the system and the instructed primary task. In other words, don't blame yourself if the system is deficient, blame the system.

Third, try to avoid the problem of nit picking an especially good system, and of saying that a system which is difficult to use is not difficult to use at all. These problems can result in similar ratings for systems with quite different characteristics. Also, try not to overreact to small changes in the system. This can result in ratings which are extremely different when the systems themselves are quite similar. Thus, to avoid any problems, just always try to "tell it like it is" in making your ratings.

If you have any questions, please ask the experimenter at this time.

APPENDIX D

MENTAL WORKLOAD MULTI-DESCRIPTOR RATING SCALE

ATTENTIONAL DEMAND refers to the portion of your total time required (or the amount of attention required) to perform the primary task.

YOUR RATING OF ATTENTIONAL DEMAND

A B C D E F G H I J K

NONE MODERATE EXTREMELY HIGH

ERROR LEVEL refers to the magnitude and frequency of mistakes, omissions, incorrect procedures, and incompletions you made in performing the primary task.

YOUR ERROR LEVEL RATING

A	B	C	D	E	F	G	H	I	J	K
NO ERRORS					MODERATE LEVEL					NUMEROUS OR LARGE ERRORS

DIFFICULTY refers to how hard or difficult you found the primary task.

YOUR DIFFICULTY RATING

A B C D E F G H I J K

EXTREMELY EASY — MODERATE — EXTREMELY DIFFICULT

TASK COMPLEXITY refers to how complicated or complex you found the primary task.

YOUR TASK COMPLEXITY RATING

A B C D E F G H I J K

EXTREMELY EASY

MODERATE

EXTREMELY COMPLEX

MENTAL WORKLOAD refers to the integrated mental effort required to perform the primary task. It includes such factors as depth of thinking and level of concentration required by the primary task.

YOUR MENTAL WORKLOAD RATING

A B C D E F G H I J K

EXTREMELY LOW — MODERATE — EXTREMELY HIGH

STRESS LEVEL refers to your emotional reaction while you performed the primary task. Stress may be considered as your feeling of anxiety, concern, uneasiness and uncertainty brought on as a direct result of performing the primary task.

YOUR STRESS LEVEL RATING

A	B	C	D	E	F	G	H	I	J	K
ABSOLUTE CALM					MODERATE STRESS					EXTREME STRESS

APPENDIX E

MENTAL WORKLOAD MULTI-DESCRIPTOR RATING SCALE INSTRUCTIONS

Overview

After each of the following flights, you will be asked to give a rating on several descriptors associated with the primary flight task. The descriptors are attentional demand, difficulty, error level, task complexity, mental workload, and stress level. Each of these descriptors is defined on the sheet on which your rating will be made, so that you will not become confused by the terms. Before you make any flights, we will review:

1. A full set of the rating scale sheets and the definitions of terms used,
2. The steps you should follow in making your ratings on the scale, and,
3. How you should think of the ratings.

If you have any questions as we review these points, please ask me. You will have one practice trial before we begin the experimental trials.

Important Definitions

Before examining the rating scale sheets, I want to provide you with an important definition that you should keep in mind while performing the ratings. The words "primary task" appear on every rating sheet. "Primary task" is the task you have been assigned in this experiment. It includes flying the aircraft within specified levels of accuracy and detecting and identifying danger conditions that will be presented on the engine, fuel, and carburetor icing instruments. It will be described in more detail later. Thus your ratings should be based on the total of these duties during each experimental run.

Also, later we will use the word "system". The system is the complete group of equipment you will be using in performing the primary task. (For the present experiment, the system is composed of the aircraft simulator, its instruments, controls, and the pushbutton module for identifying danger conditions.)

Now, let's proceed to the rating sheets. The first of them deals with Attentional Demand. As the sheet indicates, attentional demand refers to the portion of your total time required (or the amount of attention required) to perform the primary task. Below the definition on the sheet is a rating scale. After an experimental run you are to circle the one letter which best describes the attentional demand associated with the primary task. If the task took up most of your attention, your rating should be to the right of center, that is, one of the letters G through K. If the task took little of your attention, your rating should be to the left of center.

The letters themselves have no meaning implied. They are simply used to designate equal intervals on a continuum of attentional demand from none to extremely high. The letters allow us to extract information from the sheet with little likelihood of error.

We would like for you to rate your attentional demand in performing the primary task as accurately as possible. Please take as much time as necessary and remember, that as you move to the right on the scale, you are indicating higher attentional demand, and as you move to the left on the scale, you are indicating lower attentional demand. A rating of C represents less attentional demand than a rating of D.

The remaining five rating scale sheets each contain a single descriptor definition and a single rating scale associated with that descriptor. Please read over the five remaining definitions and examine their corresponding scales. The descriptions are not necessarily mutually exclusive (non-overlapping). Therefore, it is possible that your ratings on two or more scales may be similar. However, on each scale, please rate the descriptor based only on the definition given. If there are any questions, please ask.

Rating Scale Steps

After you have completed an experimental run, you will be given a full set of the rating scale sheets. Read the descriptor definition at the top of the first sheet. Then carefully select a rating on the scale which best describes the level of the descriptor you experienced.

Remember, please do not hurry in making your rating. Read the definition first, then carefully make the rating by circling the letter which is most appropriate.

How You Should Think of the Rating

Before making ratings there are several points that need to be emphasized. First, be sure to try to perform the primary task as instructed and make all your evaluations within the context of the instructed task. Try to maintain adequate performance as specified for your primary task. Second, the rating sheets are not a test of your personal skill. On all of your ratings, you will be evaluating the system for a general user population, not yourself. You may assume you are an experienced member of that population. You should make the assumption that problems you encounter are not problems you created. They are problems created by the system and the instructed task. In other words, don't blame yourself if the system is deficient, blame the system.

Third, try to avoid the problems of nit picking an especially good system and of saying that a system which is difficult to use is not difficult to use at all. These problems can result in similar ratings for systems with quite different characteristics. Also, try not to overreact to small changes in the system. This can result in ratings which are extremely different when the systems themselves are quite similar. Thus, to avoid any problems, just always "tell it like it is in making your ratings".

If you have any questions, please ask the experimenter at this time.

APPENDIX F

COMPUTATIONAL PROCEDURES FOR INDIVIDUAL WORKLOAD ESTIMATION TECHNIQUES

Subjective Opinion Techniques

Modified Cooper-Harper scale rating. The actual ratings given by the subject pilot, from 1 to 10, were obtained for each of the three experimental flights. No conversion calculation was necessary for this rating.

Multi-Descriptor scale rating. For each experimental flight, a single rating was computed as the arithmetic mean of the ratings on the six descriptors. The letter "A" corresponded to a rating of zero, "B" corresponded to one, and so on, up to "K," which corresponded to 10.

Secondary Task Techniques

Time estimation standard deviation. For each experimental flight, the standard deviation of the subjects' time interval estimates was computed. Over the five-minute data collection period, the subject had the opportunity to produce 14 time estimates. On some trials, the subject pilot did not initiate the beginning of an interval after the word "now." These trials were not included in the standard deviation computations. Trials on which the subject initiated the interval estimate but did not signal the end before the next "ready, now" signal were conservatively scored as 20 seconds in length. Twenty seconds was the length between the trial delineators, so the 20 second score represented the minimum possible (most conservative) length of the subject's unfinished estimate.

Tapping regularity. For each experimental flight, Michon's (1966) measure of tapping regularity was computed over the five-minute data recording period. This measure, in units of "pels," is based on the summation of the absolute values of the differences in time between successive taps of the microswitch. It takes into account the regularity of tapping performance on both the "loaded" tapping task (performed during one of the experimental flights) and on the baseline tapping task (performed by itself), in the following general formula:

$$\frac{\text{loaded tapping regularity - baseline tapping regularity}}{\text{baseline tapping regularity}}$$

The computational formula, in Michon's (1966) terms and adapted for this experiment is:

$$\text{PML}_k\text{(in pels)} = \frac{\frac{N_k}{T_k} \sum_{1}^{N_k-1} |\Delta t_k| - \frac{N_o}{T_o} \sum_{1}^{N_o-1} |\Delta t_o|}{\frac{N_o}{T_o} \sum_{1}^{N_o-1} |\Delta t_o|}$$

where: t = length of interval between two successive taps

T = Total period of measurement in seconds, during which N intervals are produced

N = number of intervals produced in T seconds

$|\Delta t| = |t_n - t_{n+1}|$

subscript n = rank order of interval t

subscript o = index of data related to baseline tapping

subscript k = index of data related to loaded tapping; in this case, low, medium, or high loading

PML = "perceptual-motor load"

Physiological Measurement Techniques

Pulse rate variability. This measure was computed as the standard deviation of the instantaneous pulse rate obtained from the Hewlett-Packard Model 7807C patient monitor-plethysmograph system. The measure was computed on-line over the five-minute data collection period using the EAI-380 hybrid computer. A scaled voltage value was read from the digital voltmeter.

Respiration rate. Respiration rate was obtained by counting the number of breaths-per-minute from the stripchart for each five-minute data collection period. One breath corresponded to one inhalation/exhalation cycle.

Primary Task Techniques

Danger condition response time. For each five-minute data period flight, this primary task measure was computed as the arithmetic mean of the subject's response times in seconds to the carburetor icing LED. On those trials in which the subject did not respond by pulling the carburetor heat knob before the LED went out automatically, the response time was scored as 15 seconds. (Fifteen seconds was the maximum duration that the LED was left on, so this was a conservative correction.)

Control movements. During each experimental flight, the total number of elevator, aileron, and rudder inputs during the five-minute data collection period was tabulated on the Heathkit Model 1600A digital counter. This count was then divided by 300 seconds (five minutes) to obtain the number of control movements-per-second. A single movement was said to occur whenever the particular control movement rate attained a velocity greater than two percent of full range per second after the time derivative of control position passed through zero.

Pitch high-pass mean square. This primary task measure, used only in the intrusion analysis, was obtained for each flight as follows. A "raw" pitch position signal was obtained directly from the GAT-1B dynamics computer and input to high-pass filtering and mean square computational programs on the EAI-380 computer. A running pitch mean square value was computed on-line during the five-minute data collection period. The final mean square value was displayed as a scaled voltage value on the digital voltmeter of the computer. Low-frequency deviations were removed with a high-pass filter having a cutoff frequency of 0.05 radians-per-second.

Roll high-pass mean square. This measure, also used only in the intrusion analysis, was obtained in a manner identical to that of the pitch high-pass mean square measure above.

APPENDIX G

WORKLOAD EXPERIMENT - GENERAL DESCRIPTION

The purpose of this experiment is to investigate the responses of pilots under conditions of varying difficulty. During the experiment you will be asked to fly several short cross-country flights in an aircraft simulator. The initial flight will be for practice only. In three subsequent flights data will be collected. The simulator that you will fly, a GAT-1B, is a standard moving-base flight simulator that is commonly used for flight training in the United States. During all flights, you should fly the simulator as you would an actual aircraft. At no time during the experiment will adverse conditions or extreme turbulence occur. Also, at any time during the course of the experiment, even after reading more detailed instructions, you may decline to continue participation for any reason. If so, you will be paid for the portion of time that you have spent in the experiment. Obviously, the experimenters would like for you to complete the experiment.

For the flight session you will be assigned to a particular experimental condition. Some conditions may require that you perform simple additional tasks while flying the aircraft. You may also be asked to wear special physiological sensors. These sensors are completely safe and do not apply any electrical current to your body. They are comfortable to wear and you will be familiarized with them prior to application. The sensors have the sole purpose of checking your body's reaction to the tasks. Finally, you may be asked for opinion ratings after each flight concerning the tasks you performed in the flight. Because this experiment will be comparing different conditions, as a participant you will receive only a subset of the items mentioned above.

Prior to the flight tasks you will be given explicit instructions concerning the flight profile you are to fly. After the instructions, a member of the research team will answer questions you may have. However, if a particular question may influence the outcome of the experiment, the investigator may delay a detailed answer until after your data flights.

Some of the tasks you will encounter during the flights may seem difficult to perform. In thse instances, you are simply to do your best at all times. The tasks are not designed to test your skill. Also, the data that you produce in this experiment will be treated in a strictly anonymous fashion.

The duration of the flight session will be approximately two hours. You will be paid $5.00 per hour for your participation.

Following the experiment, you are asked to refrain from discussing the experiment with other persons. If a future subject were to learn of any aspect of the experiment beforehand, the data might be biased and contamination of the results of the project would then occur. After May 15, 1982 all data will have been obtained and you may discuss the experiment with anyone.

The research team consists of Mr. John G. Casali, Ph.D. candidate, IEOR Department, and Dr. Walter W. Wierwille, Professor, IEOR Department. These individuals may be contacted at the address and phone below. The research project is sponsored by NASA, Ames Research Center, Moffett Field, California.

Vehicle Simulation Laboratory
Department of IEOR
Room 155 Whittemore Hall
Virginia Tech
Blacksburg, VA 24060
Phone (703) 961-5358

APPENDIX H

PARTICIPANT'S CONSENT

As a participant in this experiment, you have certain rights. The purpose of this sheet is to describe these rights to you and to obtain your written consent to participate.

1. You have the right to discontinue participating in the study at any time for any reason. If you decide to terminate the experiment, inform a member of the research team and he will pay you for the portion of time you have spent.

2. You have the right to inspect your data and to withdraw it from the experiment if you feel that you should. In general, data are processed and analyzed after all subjects have completed the experiment. In this experiment, the investigators can provide you with some qualitative information immediately following the experiment. Subsequently, all data are treated with anonymity. Therefore, if you wish to withdraw your data, you must do so immediately after your participation is completed.

3. You have the right to be informed as to the overall results of the experiment. If you wish to receive a synopsis of the results, include your address (four months hence) with your signature below. If after receiving the synopsis, you would then like further information, please contact the Vehicle Simulation Laboratory (address and phone below) and a full report will be made available to you.

The faculty and graduate student members of the research team sincerely appreciate your participation. They hope that you will find the experiment a pleasant and interesting experience. If you have any questions about the experiment or your rights as a participant, please do not hesitate to ask. The investigators will try to answer them, subject only to the constraint that the results will not be pre-biased by a detailed answer.

Your signature below indicates that you have read and understood your above stated rights as a participant, and that you consent to participate.

____

Signature

____

____

____

printed name and address for a synopsis of results

Vehicle Simulation
Laboratory
Department of IEOR
Room 155 Whittemore
Hall
Virginia Tech
Blacksburg, VA 24061
Phone (703) 961-5358

APPENDIX I

TIME ESTIMATION (SECONDARY TASK) INSTRUCTIONS

In each of your four flights (one practice flight and three experimental flights) you will be asked to perform a secondary time estimation task while flying the aircraft simulator. You are to perform the time estimation task simultaneously with the integrated primary task. The primary task consists of flying the simulator within specified levels of accuracy and detecting danger conditions presented on the engine, fuel, and carburetor icing instruments. The primary task will be described in more detail later.

For the time estimation task, you will perform mental estimates of 10-second intervals, indicating the beginning and end of each interval by pressing the microswitch mounted on the control yoke of the simulator. Follow the procedure outlined below for each trial.

1. The beginning of each trial will be indicated by the word "ready" over the cockpit speaker. This is to signal you to prepare to estimate a 10-second interval.
2. Soon after the word "ready", the word "now" will be presented over the speaker. "Now" signals you to press the microswitch once to indicate the beginning of the 10-second interval.

3. When you feel that 10 seconds have elapsed since you first pressed the microswitch, press it again (once) to indicate the end of the 10-second interval.
4. After you have pressed the microswitch a second time to signal the end of the interval, the current trial is over. Wait for the next "ready" signal to begin the next trial.

During your time estimates, you are to continue flying straight and level, maintaining specified altitude, airspeed, and heading, and detecting and identifying danger conditions. Try not to count or to tap to aid yourself in making time estimates. Merely press the microswitch to indicate the beginning and end of an interval which you perceive to be of 10-seconds duration. Also, please remember that your primary task is to fly the simulator and to detect and identify danger conditions. Try not to allow your time estimations to interfere with the performance of your primary task.

Prior to performing the time estimation task during the three experimental flights, you will be given two opportunities to practice the time estimation task.

First, while seated in the stationary simulator ("on the ground") you will be given several practice trials to perform the time estimation task by itself. The procedure for the practice trials will be identical to the procedure discussed above. When you hear "ready", prepare to press the microswitch. When you hear "now", press the microswitch to indicate the beginning of a 10-second interval. After you feel that 10 seconds have elapsed, press the microswitch again to indicate the end of the interval. Also, you will have an opportunity to practice the time estimation task again while flying the simulator during the practice flight. Again, your signal to prepare to begin an interval estimate will be the word "ready" over the cockpit speaker. If you have any questions concerning the time estimation task procedures, please ask the experimenter at this time.

APPENDIX J

TAPPING REGULARITY (SECONDARY TASK) INSTRUCTIONS

In each of your four flights (one practice flight and three experimental flights) you will be asked to perform a secondary tapping task while flying the aircraft simulator. You are to perform the tapping task simultaneously with the integrated primary task. The primary task consists of flying the simulator within specified levels of accuracy and detecting danger conditions as presented on the engine, fuel, and carburetor icing instruments. The primary task will be described in more detail later.

For the tapping task, you are to tap (depress) the control yoke-mounted microswitch at a rate of 1 tap (depression) every 2 seconds. Try your best to tap the microswitch at this rate as regularly or as rhythmically as possible. Because you are trying to tap at a rate of 1 tap per 2 seconds, each microswitch depression will indicate that 2 seconds (in your opinion) have elapsed since your last depression. The measured interval will be from one switch depression to the next depression. You do not have to "flick" or hit the microswitch and let it return immediately after each tap. Instead, it is probably better to tap the microswitch with a smooth, stroking motion to maintain rhythm.

Also, please remember that your primary task is to fly the simulator and to detect and identify danger conditions. Try not to allow your tapping to interfere with performance of your primary task.

Prior to performing the tapping task during the three experimental flights, you will be given two opportunities to practice the tapping regularity task.

First, while seated in the stationary simulator ("on the ground") you will be instructed to tap the microswitch at the rate of 1 tap per 2 seconds for a few minutes of practice. Later, you will again perform the tapping task while actually flying the simulator during the practice flight.

During all flights, the experimenter will instruct you over the cockpit speaker when to begin tapping and when to stop tapping. If you have any questions concerning the tapping task procedures, please ask the experimenter at this time.

APPENDIX K

PHYSIOLOGICAL SENSOR DESCRIPTION

Ear Plethysmograph

During your flights, you will be requested to wear a special physiological sensor while you fly the simulator. This sensor will not harm you in any way. It does not emit harmful radiation, and it does not make any electrical contact with your body. There are no electrodes involved. The sole purpose of the sensor is to check you body's reaction to the flight task.

The sensor is the ear plethysmograph. It is worn on the top of your right ear like this (experimenter demonstrates on his own ear).

Torso Belt

During your flights, you will be requested to wear a special physiological sensor while you fly the simulator. This sensor will not harm you in any way. It does not emit harmful radiation, and it does not make any electrical contact with your body. There are no electrodes involved. The sole purpose of the sensor is to check your body's reaction to the flight task.

The sensor is called the torso belt. It is worn around your waist like this (experimenter demonstrates on his own waist.)

APPENDIX L

DANGER CONDITION RATINGS (PART 1)

Several danger conditions can be indicated on the flight simulator instrument panel. These conditions are listed below. Rank order them in terms of how critical or severe you think that they are to successful completion of a routine flight. Assume that the danger conditions appear at the halfway point in a daylight VFR cross-country flight of three-hours duration in a single-engine propeller-driven plane. Also assume that never more than one danger condition appears on one flight. Assume that both fuel tanks are full upon takeoff. Give a ranking of 1 to the most severe condition, 2 to the next, and so on. If you feel that two or more danger conditions are equally severe, rank them with the same number.

Indicated Danger Condition	Ranking
Fuel left tank extreme low	
Fuel right tank extreme low	
Alternator amperes extreme low	
Alternator amperes extreme high	
Oil pressure extreme low	
Oil pressure extreme high	
Cylinder head temperature extreme low	
Cylinder head temperature extreme high	
Oil temperature extreme low	
Oil temperature extreme high	
Carburetor icing (assume heat-equipped)	

DANGER CONDITION RATINGS (PART 2)

Now that you have ranked the danger conditions that may appear on the flight simulator instrument panel, the research team will know which conditions you feel are most severe. However, for the purposes of this experiment you are to regard all danger conditions as equally critical or severe. You are to assume that each indicated danger condition will have an equal impact on the aircraft's performance. In the next set of instructions, you will be told to respond to indicated danger conditions (during a flight) by pressing a corresponding pushbutton for each danger condition with equal attentiveness for all danger conditions during your scans of the instrument panel. In other words, try to pay equal attention to each instrument on which a danger condition may be displayed.

APPENDIX M

DANGER CONDITION DETECTION/IDENTIFICATION INSTRUCTIONS

Several instruments and a single warning light on the flight simulator instrument panel will be used at various intervals in your flights to indicate impending danger conditions in several of the aircraft's systems. These danger conditions will be indicated as follows.

I. ENGINE/FUEL INSTRUMENTS (located on the lower left of the panel in a group of six small instruments):

1. Fuel left/fuel right (LOW only) - a low fuel condition in either the left or right wing fuel tanks will be indicated by the needle pointer swinging to the far left position and remaining there. Separate instruments are provided for the left and right tanks. Normal fuel level will be at approximately one-quarter full in each tank.
2. Alternator amperes (LOW or HIGH) - a sudden failure in the aircraft electrical charging system will be indicated by the needle pointer swinging to the far left (low) or far right (high) and remaining there. If low, a full discharge is indicated. If high, the system is overcharging. Normal system performance is indicated by a near center pointer position.

3. Oil pressure (LOW or HIGH) - engine oil pressure may suddenly go low (far left pointer position) or high (far right pointer position). In either case, abnormal lubrication system operation is implied. Normal oil pressure is indicated by a near center pointer position.

4. Cylinder head temperature (LOW or HIGH) - extreme low or high cylinder head temperatures will be indicated by far left (low) or far right (high) pointer positions on the cylinder head temperature instrument. Normal cylinder head temperature will be indicated by a pointer position slightly right of center.

5. Oil temperature (LOW or HIGH) - extreme low oil temperature will be indicated by a far left pointer position. Likewise, extreme high oil temperature will be indicated by a far right pointer position. Normal oil temperature is near center.

II. CARBURETOR ICING WARNING LIGHT (LED) (located on the lower right of the instrument panel).

Carburetor icing conditions will be indicated by a glowing red LED on the lower right of the instrument panel.

On each of the 6 instruments mentioned above, danger conditions will be indicated as "all or none." That is, on each individual instrument, the pointer will be at the extreme right or left to indicate a problem, or otherwise will be near the midpoint to indicate "normal." No "in between" conditions will occur. As for the carburetor icing warning light, the light will either be "on" to indicate icing or "off" for normal conditions.

As part of the task of flying the simulator, you will be required to monitor all, some, or none of the six instruments and carburetor icing light for indications of danger conditions. Your job, while flying the plane at the specified altitude, heading, and airspeed, will be to detect and identify which danger conditions exist as quickly and as accurately as you can.

For the fuel left, fuel right, alternator amperes, oil pressure, cylinder head temperature, and oil temperature instruments, you are to indicate a danger condition by pressing the corresponding pushbutton on the panel to your left. For each danger condition, press the correct button only one time. The condition should correct itself shortly thereafter. The push buttons are labeled with the names of the instrument to which they correspond. Black buttons are for low conditions while red buttons are for high conditions. For the carburetor icing light, you are to pull out the red carburetor heat knob (once) upon detecting that the LED is "on." The knob will immediately return by itself. You will be familiarized with the pushbutton panel and the carburetor heat knob after entering the simulator.

Danger conditions will occur only during the straight and level portion of a flight. (No danger conditions will occur during take-off and landing.) More than one danger condition may be indicated at once. Also, one danger condition may recur during a single flight. A danger condition should return to normal soon after you have detected and identified it. Even if you miss a danger condition it will eventually return to normal.

It is extremely important that you detect and identify the danger conditions as quickly and as accurately as possible. You will be scored for response time as well as accuracy of response. Despite any bias you may have, try to regard all danger conditions as equally important and attend to them as such. Do not attempt to diagnose what may be causing danger conditions, just detect them and indicate their presence. Furthermore, do not attempt, in any way, to correct a danger condition by altering your flying procedure such as reducing airspeed. The indicated danger conditions will in no way affect performance of the simulated aircraft.

REMEMBER

1. Detect danger conditions and identify them quickly and accurately using the appropriate pushbutton or carburetor heat knob.
2. Regard all danger conditions as being equal in criticality.

3. Do not try to diagnose, compensate for, or correct the danger condition -- just detect and identify it.

APPENDIX N

PRACTICE FLIGHT TASK INSTRUCTIONS

The procedure for your practice flight is as follows.

1. Assume that the airplane is situated on a typical runway. The field elevation is 180 feet and barometric pressure is 29.92 inches Hg. (This will not change during the experiment). You will be positioned at a heading of 0 degrees (due North) for takeoff purposes and to follow during the flight.
2. The experimenter will inform you over the intercom when you are cleared for takeoff. Takeoff speed in the flight simulator is approximately 75 mph assuming zero wind conditions. Remember to release the parking brake and check flaps position prior to each takeoff.
3. After takeoff, you should climb at about 72 mph, the speed at which the maximum rate of climb is realized. Climb to 2000 feet, level off, and fly at a constant speed of 100 mph. After leveling off at 2000 feet and attaining an airspeed of 100 mph you should trim the aircraft using the thumbwheel on the lower right portion of the dash panel.
4. You should continue straight and level flight until the experimenter instructs you to relax. During the straight and level portion of your flight, you should detect and identify <u>all</u> danger conditions which may be

displayed on the engine/fuel instruments and the carburetor ice warning light. Also, it is imperative that you maintain adequate flight performance at all times. Adequate flight performance is maintaining the specified heading of 0 degrees within ± 10 degrees, holding an altitude of 2000 feet within ± 100 feet, and sustaining an airspeed of 100 mph within ± 10 mph. You should strive to maintain wings level at all times. You will encounter mild turbulence during your flight.

5. The experimenter will inform you when to land the aircraft. No danger conditions will occur during descent. After landing, set the parking brake but do not shut down the engine. Remain in the simulator after landing for further instruction.

If you have any questions, please ask the experimenter at this time.

APPENDIX O

EXPERIMENTAL FLIGHT TASK INSTRUCTIONS

The procedures for this experimental flight are as follows.

1. Assume that the airplane is situated on a typical runway. The field elevation is 180 feet and the barometric pressure is 29.92 inches Hg. (This will not change during the experiment). You will be positioned at a heading of 0 degrees (due North) for takeoff purposes and to follow during the flight.
2. The experimenter will inform you over the intercom when you are cleared for takeoff. Takeoff speed in the flight simulator is approximately 75 mph assuming zero wind conditions. Remember to release the parking brake and check flaps position prior to each takeoff.
3. After takeoff, you should climb at about 72 mph, the speed at which the maximum rate of climb is realized. Climb to 2000 feet, level off, and fly at a constant speed of 100 mph. After leveling off at 2000 feet and attaining an airspeed of 100 mph you should trim the aircraft using the thumbwheel on the lower right portion of the dash panel.
4. You should continue straight and level flight until the experimenter instructs you to relax. During the straight and level portion of your flight, you should

detect and identify all danger conditions which may be displayed on the engine/fuel instruments and the carburetor ice warning light. Also, it is imperative that you maintain adequate flight performance at all times. Adequate flight performance is maintaining a 0 degree (due North) heading within ± 10 degrees, holding an altitude of 2000 feet within ± 100 feet, and sustaining an airspeed of 100 mph within ± 10 mph. You should strive to maintain wings level at all times. You will encounter mild turbulence during your flight.

5. The experimenter will inform you when to land the aircraft. No danger conditions will occur during descent. After landing, set the parking brake but do not shut down the engine. Remain in the simulator after landing for further instructions.

REMEMBER

Your fundamental responsibility as a pilot in this experiment is:

1. To fly straight and level, maintaining:
    a) 0 degrees (due North) heading within ± 10 degrees
    b) 2000 feet altitude within ± 100 feet, and
    c) 100 mph airspeed within ± 10 mph
2. To detect and identify all danger conditions presented on the engine/fuel instruments and carburetor ice warning light.

If you have any questions concerning these instructions, please ask the experimenter at this time.

APPENDIX P

DANGER CONDITION SPECIFICATION FOR EACH FLIGHT

(Low Loading)

During this flight, the only danger condition that will appear is carburetor icing. Again, this will be indicated by the LED on the lower right portion of the instrument panel. Respond to the lighted carb ice warning LED by pulling out the carb heat knob (once) and releasing it. Respond as quickly as you can to the ice warning light. No other danger conditions will appear on the engine/fuel instruments during this flight, so you will not be using the pushbutton panel to your left. You will only need to respond to the ice warning light using the carb heat knob.

(Medium Loading)

During this flight, the only danger conditions that will appear are carburetor icing and extreme low fuel in the left and right tanks. Again, carburetor icing will be indicated by the LED on the lower right portion of the instrument panel`and extreme low fuel in the left/right tank will be indicated by a pointer position below the last indicator mark on the left/right tank fuel gauge. Respond to the lighted carb ice warning LED by pulling out the carb heat knob (once) and releasing it. Respond to extreme low fuel in the left/right tank by pressing the pushbutton corresponding to this condition. No other danger conditions will ever appear on the oil pressure, oil temperature, and cylinder head temperature gauges during the flight. You will only need to respond to the ice warning LED and the fuel left/right gauge. Remember to respond as quickly as you can.

(High Loading)

During this flight, danger conditions may appear on any or all of the engine and fuel tank gauges as well as the carb ice warning LED. Respond to the engine/fuel danger conditions by pressing the appropriate pushbutton on the panel to your left and respond to the lighted carb ice warning LED by pulling out the carb heat knob (once) and releasing it. Remember to respond as quickly and as accurately as you can.

The two page vita has been removed from the scanned document. Page 1 of 2

A SENSITIVITY/INTRUSION COMPARISON OF MENTAL WORKLOAD ESTIMATION TECHNIQUES USING A SIMULATED FLIGHT TASK EMPHASIZING PERCEPTUAL PILOTING BEHAVIORS

by

John Gordon Casali

(ABSTRACT)

Forty-eight licensed pilots flew three cross-country flights in which certain aspects of perceptual workload were varied by altering the rate and number of instrument-displayed incipient danger conditions. A moving-base simulation of a single-engine general aviation aircraft was used. The sensitivity of eight mental workload estimation techniques to changes in perceptual workload was investigated within a univariate factorial design. Concurrently, the differential intrusion of the eight techniques on four primary task measures was investigated using multivariate analysis.

Of the eight techniques, six displayed statistically-significant sensitivity to load level. These included two opinion rating scales, secondary task measures of time estimation standard deviation and tapping regularity, respiration rate, and a primary task measure of danger condition detection/identification time. No intrusion effect was found. Recommendations for applying the various

techniques, based on the relative sensitivity of those showing significance, are discussed.