NASA CONTRACTOR REPORT 166496

Comparative Evaluation of Workload Estimation Techniques in Piloting Tasks

W. W. Wierwille

NASA GRANT NAG2-17
February 1980 - February 1983

NASA

NASA CONTRACTOR REPORT 166496

Comparative Evaluation of Workload Estimation Techniques in Piloting Tasks

Walter W. Wierwille
Department of Industrial Engineering Operations Research
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

Prepared for
Ames Research Center
under NASA Grant NAG2-17

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

FINAL REPORT

February 1, 1980 to February 1, 1983

COMPARATIVE EVALUATION OF WORKLOAD ESTIMATION TECHNIQUES IN PILOTING TASKS

NASA GRANT NAG2-17

Prepared for:
University Affairs Office
Ames Research Center
National Aeronautics and Space Administration
Moffett Field, California 94035

Submitted by:
Dr. Walter W. Wierwille
Department of Industrial Engineering and Operations Research
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

IEOR Department Report No. 8303

Acknowledgements

The work described in this document was performed under NASA Ames grant NAG2-17 during the period February 1, 1980 to February 1, 1983. Mr. Tom Wempe served as grant monitor in the early stages and Mrs. Sandra Hart served as grant monitor thereafter. Thanks are due to both grant monitors for their helpful technical suggestions.

Virginia Polytechnic Institute and State University provided cost-sharing support for the project and made laboratory equipment available for measures assessment. The Cunningham Foundation provided fellowship aid for Dr. Casali while he performed work on his dissertation. The dissertation became part of the results of this project.

Background

In January, 1980, NASA-Ames Research Center awarded a research grant to Virginia Polytechnic Institute and State University (Virginia Tech). The objective of this research was to examine the sensitivity and intrusion of a wide variety of workload assessment techniques in simulated piloting tasks. The study employed four different piloting tasks emphasizing psychomotor, perceptual, mediational, and communications aspects of piloting behaviors. An instrumented moving base general aviation aircraft simulator was used for the study. This document provides a summary of the research.

Table of Contents

I. INTRODUCTION

The increasing complexity of aircraft systems and the changing roles of pilots and other aircrew personnel have resulted in the need for techniques to measure operator workload in a wide range of situations and tasks. One need only initiate a preliminary survey of the literature on operator workload assessment techniques to discover that a mass of information has accumulated rapidly in the past two decades. However, major reviews of this literature have concluded that while workload research has advanced in both scope and technology, basic questions have gone unanswered for the practitioner who wishes to select workload measures for a given application. It has been pointed out by Wierwille and others that, in particular, there is a lack of information on relative sensitivity and intrusion of individual techniques. Without this information it is difficult to select appropriate estimation techniques for a given task.

The purpose of this research, performed at Virginia Tech, has been to help fill the need for practical information. Specifically, techniques for measurement of pilot workload have been selected and compared to determine their relative sensitivity and intrusion. Sensitivity can be defined as the relative ability of a workload estimation technique to discriminate statistically significant differences in operator loading. High sensitivity requires discriminable changes in score means as a function of load level and low variation of the scores about the means. Intrusion can be defined as an undesirable change in the task for which workload is being measured, resulting from the introduction of the workload estimation technique or apparatus.

Prior to this research study, there had been no definitive major effort aimed at sensitivity and intrusion. As a result, progress in determining which workload estimation techniques should be used in a given application was slow. The danger is that insensitive techniques may be used in a given application. These techniques would show no substantial change in comparative workload conditions, whether or not there is in fact a difference.

Four separate experiments were run under the grant. In the first study, psychomotor behavior was emphasized by having instrument-rated pilots perform manual ILS landing tasks in an aircraft simulator. In the second study, mediational behavior was emphasized by having pilots solve navigational problems presented on a display while maintaining a specified course, altitude, and airspeed. In the third study, perceptual behavior was emphasized by having pilots detect and identify instrument-indicated danger conditions while also maintaining specified course, altitude, and airspeed. And, in the fourth study, communications behavior was emphasized. In this study, pilots had to recognize and respond correctly to their own call sign and certain variations of their call sign while carrying out specific detailed commands from a simulated ground controller.

The four experiments conducted under the grant were designed to cover the major activities aircrew members perform, not so much in terms of details, as in terms of general categories of activities. In fact, the four studies were designed to emphasize the four major activities shown in the "Universal Operator Behaviors" listing of Berliner, Angell, and Shearer (1964)*, that is, psychomotor, mediational, perceptual, and communications (Table 1).

*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, 1964.

Table 1

Classification of Universal Operator Behavior Dimension
(After Berliner, Angell, and Shearer, 1964)

Processes	Activities	Specific Behavior
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

When the project was initiated, it was recognized that meaningful results could only be obtained if moderately realistic flight tasks were used. Accordingly, a Singer-Link GAT-1B general aviation flight simulator was obtained. The cost of a new device was prohibitive. So, a used one was found, purchased, installed, and refurbished. This represented a major task, but nevertheless saved a great deal of money. When the simulator refurbishment was complete, the simulated aircraft could be trimmed. With hands and feet off the controls, the simulated aircraft would very slowly drift off course, as in a real aircraft. The "drift test" is a good one to use to insure that all dynamic aspects of the simulator are correct. An imbalance in a power supply or pick-off potentiometer output will make low-drift trimming impossible. The "drift-test" also insures that the "workload" conditions imposed on the pilot are indeed those imposed by the experimenter over the trimmed aircraft. Figure 1 is a close-up view of the simulator cockpit with a subject performing a flight task while wearing physiological sensors.

In addition to the simulator itself, methods had to be devised to obtain and record a wide variety of workload estimation measures. Because Virginia Tech already had a vehicle simulation laboratory, the "raw material" for obtaining the measures already existed. After the simulator refurbishment was completed in the laboratory, the simulator was interconnected via hanging cables and slipring interconnections to an EAI-380 hybrid computer and other specific circuitry and equipment. Figure 2 shows the overall experimental setup, with the computational equipment in the foreground and the simulator in the background. Changes were made in programs and other equipment to accommodate each of the four experiments. The approach worked quite well and eliminated the need for large investments in digital equipment and software.

Figure 1. Cockpit of the simulator with pilot/subject wearing physiological sensors.

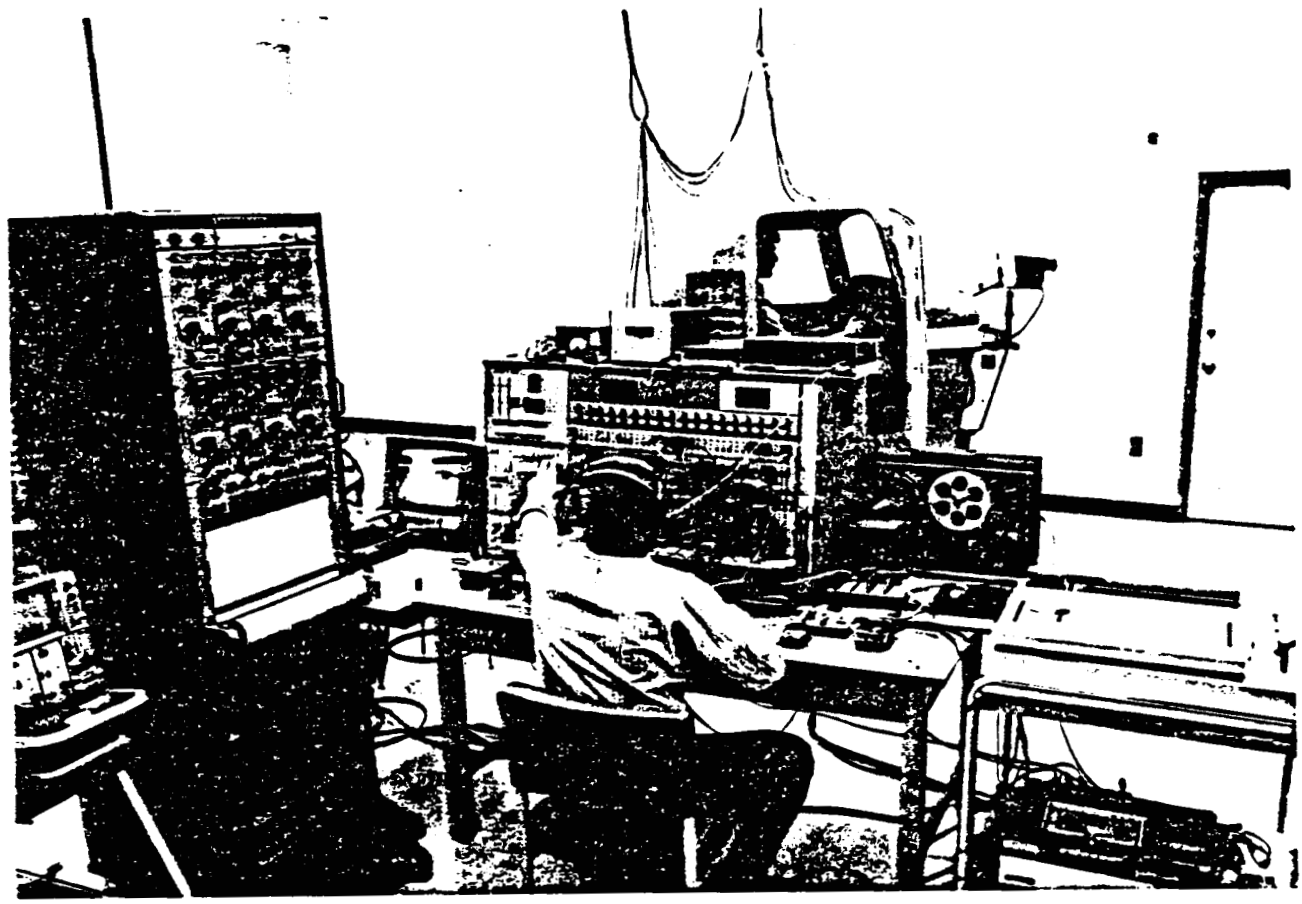

Figure 2. Overall view of the simulation facility with the experimenter's station in the foreground and the simulator in the background.

In general, the research results of the project have shown that sensitivity of various workload estimation techniques varies widely. Some twenty-five techniques were examined during the course of the investigation.

As a class, physiological techniques did not demonstrate sensitivity to any appreciable extent in our studies. Heart rate, heart rate variability, respiration rate, and pupil diameter were among the techniques tested. Heart rate, respiration rate, and pupil diameter each showed a small amount of sensitivity in only one (each) of four experiments. It should be remembered, however, that in all cases the pilot-subjects were fresh and practiced. Physiological measures might be more helpful in accessing "stress and strain" over many hours of time on task.

In the spare mental capacity category, time estimation was found to be sensitive in three out of four experiments. It appears particularly well suited to assessment of the perceptual and mediational components of workload, but is also somewhat sensitive to psychomotor load. Michon tapping regularity was found sensitive only to perceptual load. Other arithmetic logic and shadowing secondary tasks were found insensitive to psychomotor load.

Properly selected primary task measures were found sensitive in each of the four experiments. In fact, these measures are, as a group, the most sensitive. In the psychomotor experiment, the measure of control movements per unit time was highly sensitive to loading. In the perceptual experiment, time to detect and identify danger conditions was sensitive, and in the mediational experiment, a similar measure of response time was sensitive. In the communications experiment, numbers of errors of ommission and commission were sensitive. The results for primary task measures contradict the commonly held be-

lief that such measures are not sensitive to load. Our experiments show that if the measures are properly selected, they are highly sensitive to load.

In the opinion group, rating scales in general showed sensitivity in all four experiments. However, decision-tree rating scales appeared somewhat more sensitive than other types. In particular, the Cooper-Harper (CH) and Modified Cooper-Harper (MCH) scales, though simple, were as sensitive or more sensitive than others. The WCI/TE (Workload-Compensation-Interference/Technical Effectiveness) scale was sensitive in the psychomotor experiment, and the Multi-Descriptor scale was partly sensitive in the perceptual and communications experiments.

We believe that the project has been successful in determing which kinds of measures are sensitive to which kinds of load. And, since the simulator environment used for the experiments was quite realistic, the results appear to have face validity.

The remainder of this document is composed of four technical papers, one describing each of the four experiments. The papers describe the experimental conditions and the results of the sensitivity and intrusion analyses.

Following the four technical papers is a listing of publications generated during the project. Included in the list are the four above cited papers plus several others. The thesis and dissertations listed contain the most complete descriptions of each of the first three experiments. The communications experiment is most completely described by the paper included in this report. The thesis and dissertations are available through the normal library channels.

II. PSYCHOMOTOR EXPERIMENT

THE SENSITIVITY OF TWENTY MEASURES OF PILOT MENTAL WORKLOAD IN A SIMULATED ILS TASK

by

Walter W. Wierwille and Sidney A. Connor
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

ABSTRACT

Twenty workload estimation techniques were compared in terms of their sensitivity to changes in pilot loading in an ILS task. The techniques included opinion measures, spare mental capacity measures, physiological measures, eye behavior measures, and primary task measures. Loading was treated as an independent variable and had three levels: low, medium, and high. The load levels were obtained by a combined manipulation of windgust disturbance level and simulated aircraft pitch stability. Six instrumented-rated pilots flew a moving-base general aviation simulator in four sessions lasting approximately three hours each. Measures were taken between the outer and middle markers.

Two opinion measures, one spare mental capacity measure, one physiological measure, and one primary task measure demonstrated sensitivity to loading in this experiment. These measures were Cooper-Harper ratings, WCI/TE ratings, time estimation standard deviation, pulse rate mean, and control movements per unit time. The Cooper-Harper ratings, WCI/TE ratings, and control movements demonstrated sensitivity to all levels of load, whereas the time estimation measure and pulse rate mean showed sensitivity to some load levels.

The results of this experiment demonstrate the sensitivities of workload estimation techniques vary widely, and that only a few techniques appear to be sensitive in this type of ILS task, which emphasizes psychomotor behavior.

INTRODUCTION

One of the major problems in mental workload estimation is the lack of available information on the sensitivity of various workload estimation techniques [1,2]. When a researcher or human factors engineer needs to assess workload in a given experimental situation, it is not clear which technique or techniques should be used [3]. The danger is that insensitive techniques may be used. If so, experimental results will show no differences in workload when in fact there are differences.

Sensitivity in regard to workload estimation can be defined as the relative ability of a given workload estimation technique to discriminate statistically significant differences in operator loading. High sensitivity requires

discriminable changes in the score means as a function of load level and low variation of the scores about the means. When sensitivity is defined in this way, it becomes subject to experimental determination. Based on experiments that emphasize specific operator behaviors, it should be possible to predict which given techniques are sensitive.

An experiment directed at evaluating the sensitivity of workload estimation techniques in a psychomotor task has been completed and is reported briefly in this paper. An ILS piloting task was used for the evaluation. (For a more detailed discription of the experiment and results, see reference [4]).

EXPERIMENT

Subjects

Six male instrument-rated pilots served as subjects in this experiment. The flight time of the subjects ranged from 500 to 2700 hours with a mean of 1300 hours.

Apparatus

The primary apparatus in this experiment was a modified flight task simulator (Singer Link, Inc., General Aviation Trainer, GAT-1B). The simulator had three degrees of freedom of motion (roll, pitch, and yaw). Transulucent blinders were used to cover the windows of the simulator to reduce outside distractions and cues and to aid in the control of cockpit illumination.

Several modifications to the flight simulator were made for the experiment. These modifications permitted primary task load manipulation, secondary task operations, response measurement, and scoring. Primary task load manipulation was accomplished by changing aircraft pitch stability and random wind-gust disturbance level simultaneously. Three load conditions were developed: low, medium, and high, as shown in Table 1. Table 2 provides a list of the workload measurement techniques selected for inclusion in the present study.

Experimental Design

A complete 3 x 20 within-subject design was used for the sensitivity analysis. Load was the factor with three levels. Measurement technique (Table 2) was the factor with twenty levels.

Workload measures from different techniques were taken simultaneously on some of the data collection runs. Only those measures which were not likely to affect each other were taken simultaneously. Table 3 shows the scheme used for combining different measurement techniques for data collection. The combination of measurement techniques shown in the table was, to an extent, based on previous investigations of workload. Hicks and Wierwille's [3] study supported the combination in condition 2. The two rating scales were administered

in separate measurement conditions to prevent the ratings on one scale from biasing the ratings on the other scale. The secondary task measures were divided among several conditions because of potential intrusion and interference. Vocal measures were recorded from the two secondary tasks which required a verbal response as per Schiflett and Loikith's [5] recommendation.

It should be noted that primary task measures were recorded on all subjects and on all data collection flights for the intrusion analysis. However, only data from measurement condition 1 were used for the sensitivity analysis of the primary task measures.

General Procedure

After receiving instructions, subjects flew nine familiarization flights in the simulator. These flights were similar, but not the same as, the data collection flights. All subjects flew the familiarization flights in the same order. Steady crosswinds were introduced for each run, and subjects were given heading corrections.

After the familiarization session, the subjects participated in three data collection sessions. The familiarization session and each data collection session were held on a different day.

Each data collection session consisted of two sets of a warm-up practice flight and three data collection flights. The practice flight was the same as the first data collection flight. Since the data collection flights were counterbalanced, equal amounts of practice were provided for the low, medium, and high load conditions. The data collection flights also contained steady crosswind conditions, for which the subject was given heading corrections. The purpose of introducing steady crosswinds was to disguise the load conditions, thereby requiring subjects to fly each flight as a separate entity.

Flight Task Procedures

The flight task in this experiment was an ILS approach in the Singer Link GAT-1B aircraft simulator. Prior to the beginning of a flight, the simulated aircraft was positioned on the ground 5 miles outbound from the outer marker on the 108 degree radial, heading into the wind. When ready to begin, the experimenter informed the subject of the wind direction and speed, and gave him a heading correction for the crosswind. When contacted by the experimenter, the subject took off and climbed to 2000 feet. The subject then flew directly to the outer marker by following the localizer at 100 miles per hour until the glide slope was intercepted. Upon interception of the glide slope, the subject reduced airspeed to 80 miles per hour and proceeded down the glide slope while following the localizer to a landing. Data were recorded between the outer and middle markers. For the opinion measures, subjects gave ratings for the flight segment between the outer and middle markers immediately after landing and parking the simulated aircraft.

RESULTS

The computed scores for each technique were first converted to Z-scores (normalized scores) so that technique measure units would not affect the sensitivity analysis. Subsequently, an overall analysis of variance was performed on the scores. Since Z-scores were used, a technique main effect was not possible. A significant main effect of load was found, F (2,10) = 5.34, $p < 0.0001$, and a significant load by technique interaction was found, F (38,190) = 2.76, $p \leq 0.05$.

The load by technique interaction indicated that the measurement techniques were differentially sensitive to load. Therefore, individual ANOVAs were used to isolate the sensitive techniques.

The individual ANOVAs indicated that five of the twenty measures were sensitive. They were the Cooper-Harper scale F (2,10) = 16.39, $p = 0.0007$; the Workload-Compensation-Interference/Technical Effectiveness (SCI/TE) scale, F (2,10) = 31.15, $p < 0.0001$; the time estimation standard deviation, F (2,10) = 5.69, $p = 0.022$; the pulse rate mean, F (2,10) = 8.89, $p = 0.006$; and the control movements measure, F (2,10) = 33.34 $p < 0.0001$. The normalized means for each technique are plotted in Figures 1 through 5 as a function of load.

Newman-Keuls comparisons were then performed on the normalized means of the sensitive measures. The comparisions included low vs. medium, medium vs. high, and low vs. high load conditions. Results indicated that all differences were significant at $p < 0.05$, except for pulse-rate mean (low vs. medium and medium vs. high) and time estimation standard deviation (low vs. high).

A logical classification of techniques based on demonstrated sensitivity was generated from an examination of the Newman-Keuls comparisons, as shown in Table 4. Techniques which demonstrated sensitivity to all pairs of load conditions (i.e., low vs. medium, medium vs. high, and low vs. high) were included in class I. These measures are preferred over other techniques which demonstrated only partial sensitivity, or no sensitivity in the present study. Techniques which showed sensitivity to some differences in load conditions (but not all) were included in class II. These measures are less preferred than class I techniques, but are more preferred than class III techniques. Class III techniques did not demonstrate sensitivity to load in the present study. This class includes all techniques except those in class I and class II.

One possible reason that only five of the twenty workload assessment techniques demonstrated sensitivity in the present study is that the other techniques simply required a greater number of subjects to show a significant effect of load. It is possible to estimate the sample size required to detect a reliable load effect for a given workload assessment technique at specified levels of significance and power. These calculations were performed for techniques which did not demonstrate sensitivity in the present study, to provide an indication of the practical costs of achieving statistical significance. The procedure used for estimating the sample size required for finding sensi-

tivity is described by Bowker and Lieberman [6]. Sample sizes were estimated for a significance level of 0.05 and for a power of approximately 0.80. The results of these estimates are presented in Table 5.

CONCLUSIONS

This study has shown that five measures of workload estimation were sensitive indicators of load in a piloting task that is predominantly psychomotor in nature. Another fifteen measures, believed to be "good" measures of workload, showed no reliable effect. The main conclusion that must be drawn from the study is that few measures are sensitive to psychomotor load.

Of the five techniques demonstrating sensitivity, only three exhibited monotonic score increases with load as well as statistically reliable differences between all pairs of load levels. Consequently, only the three meet all criteria for sensitivity to psychomotor load. These class I techniques are the ones that are recommended for measurement of psychomotor load:

Cooper/Harper ratings,
WCI/TE ratings, and
Control movements per second.

The other two techniques showed sensitivity to psychomotor load, but did not discriminate between all pairs of load levels. These class II techniques are:

Time estimation standard deviation, and
Pulse rate mean.

These measures would be helpful in evaluating psychomotor load, but they should not be relied on exclusively. At least one class I technique should also be used in conjunction with these measures.

It is worth noting that only two opinion measures were taken in the present experiment, and both proved sensitive. This suggests that well-designed rating scales are among the best of techniques for evaluating psychomotor load. In regard to the primary task measures, the control movements measure alone was sensitive. However, this measure is also the only primary task measure which reflected "strategy" of the pilot. Consequently, one could speculate that selecting a primary task measure that reflects strategy will most likely result in good sensitivity.

Fifteen (techniques) measures showed no reliable change as a function of load. When these fifteen measures were subjected to a power analysis to determine sample size, the number of subjects required ranged from 12 to well over 100 (Table 5). One can only conclude that at best the fifteen measures, as taken, are much less sensitive to psychomotor load than the five appearing in Classes I and II. Of course, there is always the possibility that the measures would be sensitivite to loading along other dimensions of human performance, such as psychomotor tasks of a different nature, or mediational or cognitive tasks, for example.

In general, the results of the experiment show that there are wide variations in the sensivity of workload estimation measures. Great care must be taken in selecting measures for a given experiment. Otherwise, it is possible that no changes in workload will be found, when indeed there are changes.

REFERENCES

1. 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.

2. Wierwille, W. W. and Williges, B. H. An annotated bibliography on operator mental workload assessment. Patuxent River, Maryland: Naval Air Test Center Report No. SY-27R-80, March, 1980.

3. 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.

4. Connor, S. A. and Wierwille, W. W. Comparative evaluation of twenty pilot workload assessment measures using a psychomotor task in a moving base simulator. Moffett Field, CA: NASA-Ames Research Center, (Forthcoming report).

5. Schiflett, S. G. and Loikith, G. J. Voice stress as a measure of operator workload. Patuxent River, Maryland: Naval Air Test Center, Technical Memorandom TM 79-3 SY, December 31, 1979.

6. Bowker, A. H. and Lieberman, G. J. Engineering statistics. New Jersey: Prentice-Hall, Inc., 1959.

ACKNOWLEDGEMENTS

The authors wish to thank Mrs. Sandra Hart, NASA-Ames Research Center, for helpful technical suggestions. This work was sponsored under NASA grant NAG2-17.

TABLE 1

Primary Task Load Conditions

	LOAD CONDITION		
	Low	Medium	High
RANDOM GUST LEVEL	Low	Medium	High
Estimated Std. Dev. (mph)	0	2.7	5.9
PITCH STABILITY	High	Medium	Low
a. Control input to pitch rate output equivalent gain (degrees/s per % of control range)	0.522	3.560	7.83
b. Control input to pitch rate output equivalent time constant(s)	0.097	0.660	1.45

TABLE 2

Workload Assessment Techniques Which Were Tested in the Present Experiment

OPINION

1. Cooper-Harper Scale
2. WCI/TE Scale

SPARE MENTAL CAPACITY

3. Digit Shadowing (% errors)
4. Memory Scanning (Mean time)
5. Mental Arithmetic (% errors)
6. Time Estimation Mean (Seconds)
7. Time Estimation Standard Deviation (Seconds)
8. Time Estimation Absolute Error (Seconds)
9. Time Estimation RMS error (Seconds)

PHYSIOLOGICAL

10. Pulse Rate Mean (Pulses per minute)
11. Pulse Rate Variability (Pulses per minute)
12. Respiration Rate (Breath cycles per minute)
13. Pupil Diameter (Normalized units)
14. Voice Pattern (Digit Shadowing Task)
15. Voice Pattern (Mental Arithmetic Task)

EYE BEHAVIOR

16. Eye Transition Frequency (Transitions per minute)
17. Eye Blink Frequency (Blinks per minute)

PRIMARY TASK

18. Localizer RMS Angular Position Error (Degrees)
19. Glide Slope RMS Angular Position Error (Degrees)
20. Control Movements per second
    (Aileron + Elevator + Rudder)

TABLE 3

Combination of Measurement Techniques
for Data Collection

Measurement Condition	Measurement Techniques
1.	Cooper-Harper Scale Pupil Diameter Eye Transition Frequency Eye Blink Frequency Localizer RMS Error Glide Slope RMS Error Control Movements
2.	WCI/TE Scale Pulse Rate Mean Pulse Rate Variability Respiration Rate
3..	Digit Shadowing Voice Pattern
4.	Memory Scanning
5.	Mental Arithmetic Voice Pattern
6.	Time Estimation (Mean) (Std. Dev.) (Abs. Error) (RMS Error)

TABLE 4

Logical Classification of Techniques
Based on Demonstrated Sensitivity

Class I: Complete Sensitivity Demonstrated
- Cooper-Harper Scale
- WCI/TE Scale
- Control Movements/Unit Time

Class II: Some Sensitivity Demonstrated
- Time Estimation Standard Deviation*
- Pulse Rate Mean**

Class III: Sensitivity Not Demonstrated
- All Other Techniques (See Table 5)

*Double valued function
**Limited sensitivity

TABLE 5

Estimated Sample Sizes Required for Achieving a Significant Load Effect for Techniques not Demonstrating Sensitivity

Technique	Estimated Sample Size
SPARE MENTAL CAPACITY	
Digit Shadowing	18
Memory Scanning	>100
Mental Arithmetic	25
Time Estimation (Mean)	53
Time Estimation (Abs. Error)	>100
Time Estimation (RMS Error)	53
PHYSIOLOGICAL	
Pulse Rate Variability	45
Respiration Rate	15
Pupil Diameter	>100
Speech Pattern (D. Shadow.)	28
Speech Pattern (M. Arith.)	>100
EYE BEHAVIOR	
Eye Transition Frequency	42
Eye Blink Frequency	25
PRIMARY TASK	
Localizer RMS Error	12
Glide Slope RMS Error	41

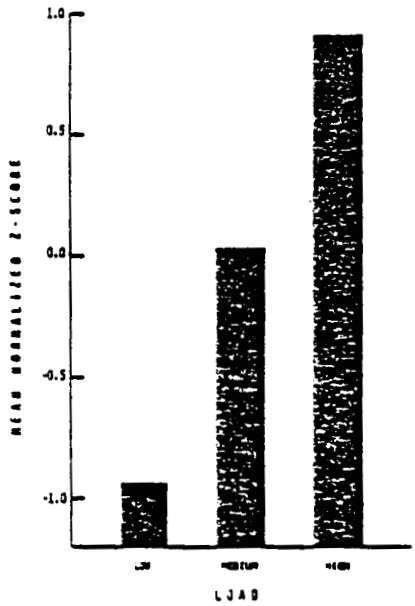

Figure 1. Mean normalized scores for the Cooper-Harper rating scale measure plotted as a function of load.

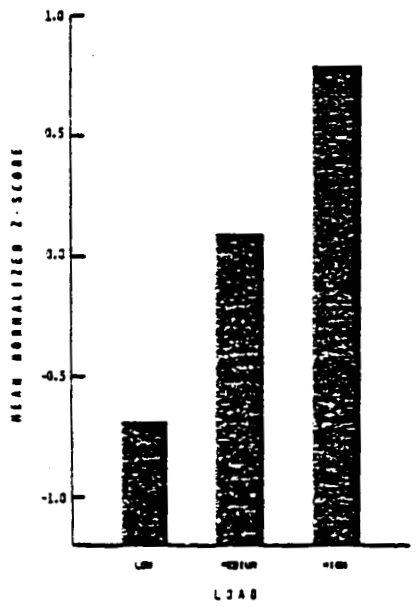

Figure 2. Mean normalized scores for the WCI/TE rating scale measure plotted as a function of load.

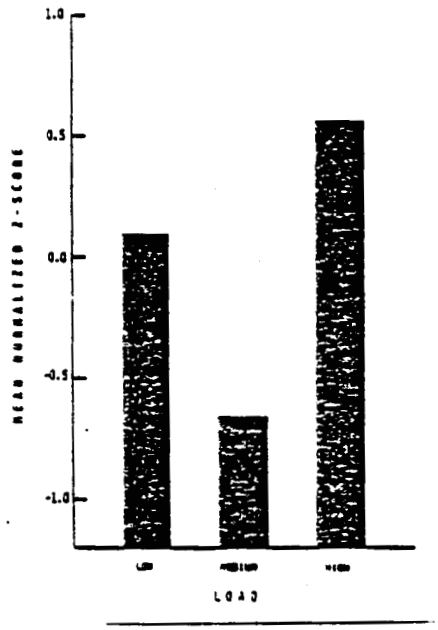

Figure 3. Mean normalized scores for the time estimation standard deviation measure plotted as a function of load.

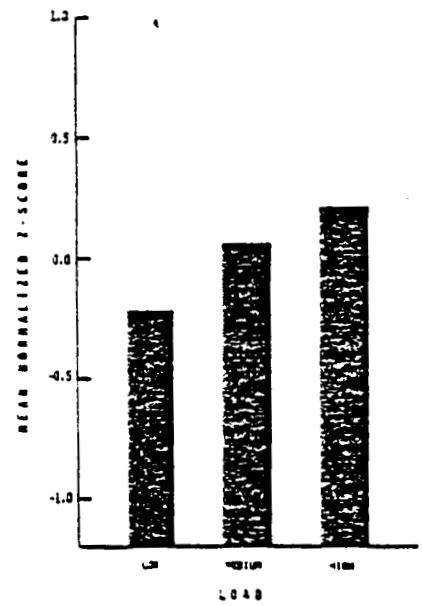

Figure 4. Mean normalized scores for the pulse rate mean measure plotted as a function of load.

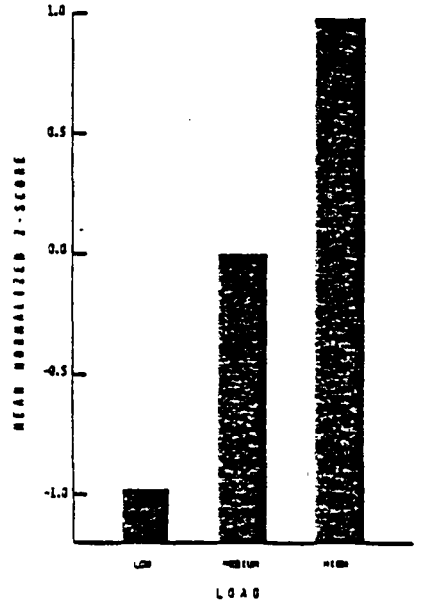

Figure 5. Mean normalized scores for the control movements measure plotted as a function of load.

III. MEDIATIONAL EXPERIMENT

EVALUATION OF THE SENSITIVITY AND INTRUSION OF WORKLOAD ESTIMATION TECHNIQUES IN PILOTING TASKS EMPHASIZING MEDIATIONAL ACTIVITY

M. Rahimi and W. W. Wierwille

Department of Industrial Engineering and Operations Research
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

ABSTRACT

In this experiment, pilots flew an instrumented moving-base simulator. Mediational loading was elicited by having them solve a variety of navigational problems. The problems were sorted into low, medium, and high load conditions based on the number and complexity of arithmetic and geometric operations required to solve them. Workload estimation techniques based on opinion, spare mental capacity, primary task performance, and physiological measures were obtained and compared. This paper describes: 1) the ability of the techniques to discriminate statistically between the three levels of loading conditions, and 2) changes in primary task performance caused by introduction of the workload technique procedures and equipment.

INTRODUCTION

Over the past thirty years, a large body of literature has been accumulated on operator mental workload estimation techniques (1,2,3,4,5, and 6). In designing new aircraft systems or modifying the existing ones, it is becoming imperative for the cockpit or system engineer to consider total mental workload demands placed on the aircrew. Therefore, accurate measurement of aircrew mental workload is a necessary part of an optimum aircrew/aircraft design. To select a technique for measuring aircrew mental workload in a given task, the following attributes should be included (6):

1. The technique should be well suited to the specific task situation.

2. The technique should accurately and reliably assess workload, that is, it should be sensitive.

3. The introduction of the technique should not significantly change the behavior of the operator/vehicle system, that is, it should be nonintrusive.

In this study a technique is said to be sensitive if it discriminates statistically between differences in operator loading requirements of an assigned task.

Also, a technique is said to be intrusive if significant primary task performance degradation is observed due to the use of the technique and associated equipment.

In the present experiment, the following eight mental workload techniques were compared with the goal of evaluating their sensitivity and intrusion to a mediationally emphasized piloting task in a moving base aircraft simulator. Two techniques from each of the four categories suggested by Wierwille and Williges (6) were included:

Opinion
1. Modified Cooper-Harper rating scale
2. Multi-descriptor rating scale

Spare Mental Capacity
Secondary task measures:
3. Time estimation
4. Tapping regularity

Primary Task Measures
5. Mediational reaction time
6. Control movements per unit time

Physiological Measures
7. Pulse rate variability
8. Pupil dilation

Opinion measures (e.g., rating scales) assume that an operator's opinion (e.g., rating) of perceived task loading is directly proportional to the operator's actual task loading. Among the different forms of opinion measures, rating scales

have been used frequently and successfully. One rating scale which has received considerable application in pilot mental workload estimation is the Cooper-Harper rating scale. The Cooper-Harper rating scale was designed for assessing handling qualities of aircraft. The authors modified the scale to match a mediationally emphasized piloting task. The second rating scale used in this study was a multi-descriptor bipolar adjective (semantic differential) rating scale. The six component scales used (attentional demand, difficulty, error level, task complexity, mental workload, and stress level) take into account the multi-dimensionality of pilot mental workload.

Spare mental capacity is a measure of the difference between the mental capacity required to perform a given task and the total capacity of the operator (7). The important assumptions associated with this concept are: (a) the operator is a single-channel system, (b) the channel has a fixed capacity, (c) the capacity has a single metric by which any task can be measured, and (d) the constituents of workload are linearly additive. Maintaining these assumptions, the operator's spare mental capacity supposedly decreases as his/her workload increases. Several approaches have been used to measure spare mental capacity. This study used two techniques from the secondary task approach. The two techniques were time estimation and tapping regularity. In theory, the degree of accuracy with which individuals perform a concurrent secondary task is a potential measure of primary task mental workload. Based on the time estimation measure, variability (standard deviation) of a subject's estimates of time (e.g., 10 second intervals) increase with an increase in primary task mental workload. Based on the tapping regularity measure, the regularity with which a subject successively moves his/her limb (e.g. tapping a finger or a foot to depress a switch) decreases with an increase in primary task mental workload.

Primary task measures are those measures which are obtained from the main or the instructed task. It is hypothesized that the increase in operator mental workload would be accompanied by a change or degradation of operator task performance. The primary task measures selected for this study were designed to reflect the mediational loading changes (reaction time to the mediational portion of the piloting task) and strategy changes of the pilots in controlling the aircraft simulator (number of control movements per unit time).

Physiological measures are those measures which are reflective of involuntary physiological changes (e.g., circulatory system changes), when the operator experiences increasing workload. The two measures selected in this group were pupil dilation and pulse rate variability. It is hypothesized that pupil diameter and pulse rate variability decrease as mediational workload increases.

METHOD

This experiment was performed in a GAT-1B moving base aircraft simulator. The simulator was modified to allow workload evaluation. An EAI-380 hybrid computer and other special purpose circuitry were interconnected with the simulator.

The primary task used for this study was to ascend in the aircraft simulator to 2000 feet, cruise at the altitude of 2000(+/-100) feet while maintaining an airspeed of 100(+/-10) mph, and hold a heading of 0(+/-10) degrees. Within the straight-and-level portion of the flight, subjects were presented with a series of slides on an Ektagraphic projector seen through the windscreen, containing navigational problems (mediational loading). These slides were presorted into low, medium, and high difficulty problems based on the number and complexity of arithmetic and geometric operations required to solve them.

The experimental design for sensitivity analysis was a complete Load by Technique factorial design. Load (low, medium, and high) was a within-subject variable and Technique (eight levels) was a between-subjects variable. Six subjects were used for each technique and the order of presentation of the three load levels was completely counterbalanced across the subjects. For intrusion analysis, five primary task dependent measures were obtained in all eight technique conditions. The five primary task measures obtained were designed to evaluate different aspects of primary task performance. They were 1. percent error of the mediational (slide) problems, 2. reaction time to the mediational problems, 3. pitch high-pass mean-squared (PHPMS) error, which is a measure of pitch control accuracy, 4. roll high-pass mean squared (RHPMS) error, which is a measure of roll control accuracy, and 5. number of control movements per second. The experimental design for intrusion analysis was the same

as the sensitivity analysis design matrix, except, five primary task measures were obtained in each cell for each technique condition.

The subjects were private pilots. Approximately equivalent cross-sections of experience levels were used for each technique. This was accomplished by selecting pilots based on preliminary questionnaire data. The average flight time per pilot was 317 hours. They flew one practice flight and three experimental flights for data collection. Most of the scores were computed on-line via the hybrid computer and special purpose circuitry. The remainder were computed shortly after all runs were completed.

Sensitivity Analysis

The raw scores in the sensitivity data matrix were standardized (z-scored) to detect true differences in the techniques, rather than scaling value differences. An overall ANOVA was performed on the z-scores. A significant main effect of load was found, $F(2,80)=20.36$, $p<0.0001$; therefore, the manipulation of load was effective. Also, a significant load by technique interaction was found, $F(14,80)=3.58$, $p<0.0001$. Therefore, scores of some techniques were more responsive to changes in load than others. To isolate those techniques which contributed to the interaction effect, eight individual ANOVAs were performed, one for each technique. In the opinion measure group, the modified Cooper-Harper scale showed a significant effect of load, $F(2,10)=14.83$, $p<0.001$ (Figure 1). It should be mentioned that, while not significant, the multi-descriptor scale did exhibit a monotonic trend, $F(2,10)=2.97$, $p=0.068$. In the spare mental capacity group, the time estimation measure showed a significant effect of load, $F(2,10)=11.39$, $p<0.001$ (Figure 2). In the primary task measure group, the mediational reaction time showed a significant effect of load, $F(2,10)=55.95$, $p<0.0001$ (Figure3). And, in the physiological measures group, neither measure showed a significant effect of load.

For techniques failing to demonstrate sensitivity, a statistical power analysis was performed to estimate the number of subjects required to detect a reliable load effect. A technique requiring a large number of subjects to demonstrate sensitivity would not be cost-effective to implement in an operational environment. The multi-descriptor rating scale required 16 subjects and the other four required more than 100 subjects.

For the sensitive techniques, Newman-Keuls multiple comparisons tests were performed to determine the locus of the effect of load on the technique z-scores. The results for the modified Cooper-Harper scale showed that there was no significant difference between low and medium loading z-scores. But, significant differences were found between low and high, and between medium and high loading z-scores. The results for the mediational reaction time measure indicated that the differences between all pairs of means were statistically significant. Finally, the results for the time estimation measure indicated that a significant difference existed between the z-scores of low compared with medium and high loading. But, no difference was detected between medium and high loading z-scores.

On the basis of these results, the mediational reaction time was classified as completely sensitive to the piloting task. The modified Cooper-Harper scale and time estimation measures were classified as partly sensitive and the remaining five techniques as nonsensitive to the mediational loading task.

Intrusion Analysis

The purpose of this analysis was to investigate possible interference of equipment and procedures used for estimating pilot mental workload with performance on the primary flight task. In this experiment, primary task performance is composed of five primary task measures: 1) percent error of the mediational problems, 2) reaction time to the mediational problems, 3) pitch high-pass mean-squared error, 4) roll high-pass mean-squared error, and 5) number of control movements per second. The eight technique measurement conditions were the same conditions used in the sensitivity analysis except for the two primary task conditions, which were called control conditions C1 and C2. It is assumed that the equipment and procedures used in the C1 and C2 conditions were not intrusive upon the primary task performance. Therefore, C1 and C2 were used as standards for comparing intrusion of the other six measurement conditions.

A MANOVA was performed to determine whether the five primary task measures (as a group) were affected by different techniques used in this experiment. The main effect of technique was significant using the Wilk's Criterion, $F(35,153)=1.72$, $p=0.0135$.

To isolate which primary task measure was particularly affected by the introduction of different techniques, five individual ANOVAs were performed, one for each primary task measure. Only two ANOVAs showed significant main effect of technique. The main effect of technique was significant for the mediational error rate measure, $F(7,40)=3.91$, $p=0.0025$; and for the mediational reaction time measure, $F(7,40)=3.36$, $p=0.0065$.

To determine which techniques were contributing to the intrusion, Duncan's multiple comparisons tests were performed on the mean scores for mediational error rate and mediational reaction time. Comparisons of the mean mediational error rate scores for the eight technique conditions showed that the scores for the time estimation technique were significantly higher than all of the other techniques. The mean mediational error rate scores for other techniques (including C1 and C2) were not significantly different. The second set of Duncan's tests was performed on mediational reaction time scores for the eight technique conditions. The results indicated that again time estimation technique had significantly larger mean scores than the other five techniques (including C1 and C2). Also, the pulse rate variability technique had significantly larger mean scores than three other techniques (including only C2).

On the basis of the results of the intrusion analysis, the time estimation technique was considered to be substantially intrusive (on two of the five primary task measures). Additionally, the pulse rate variability technique was found to be partially intrusive (on only one primary task measure).

SUMMARY

A measure of the amount of time required to solve the mediational portion of a piloting task seems to be the best measure of pilot mental workload in piloting tasks emphasizing mediational activity. Also, the technique of asking the pilots "how mentally loaded they are", in a systematic format, appears to be a desirable alternative. A modified Cooper-Harper scale was sensitive in two out of three load comparisons, and a multi-descriptor scale exhibited an increasing trend. The time estimation (standard deviation) technique appears sensitive to mediational loading, but its procedures and equipment intrude on the pilots' primary task performance. The results of this experiment demonstrate that only certain measures are sensitive to mediational loading. These results parallel those of Wierwille and Connor (8), who worked with psychomotor tasks.

ACKNOWLEDGEMENT

This research was sponsored by NASA-Ames Research Center, Moffett Field, CA. Mrs. Sandra Hart served as grant technical monitor.

REFERENCES

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

(2) W. B. Gartner, and M. R. Murphy, Pilot workload and fatigue: A critical survey of concepts and assessment techniques. Mofett Field, California: NASA Ames Research Center, NASA TN D-8365, November, 1976.

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

(4) A. H. Roscoe (Ed.), Assessing Pilot Workload. AGARD-AG-233, February, 1978.

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

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

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

(8) W. W. Wierwille and S. A. Connor, Evaluation of twenty pilot workload assessment measures using a psychomotor task in a moving-base aircraft simulator. (Submitted to Human Factors, 1982.)

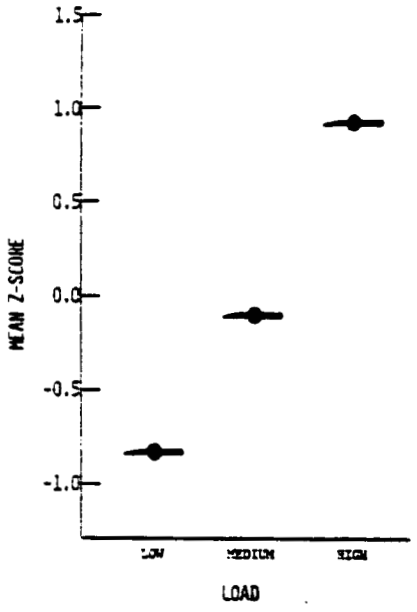

Figure 1. Mean z-scores for Modified Cooper-Harper rating scale measure vs. load.

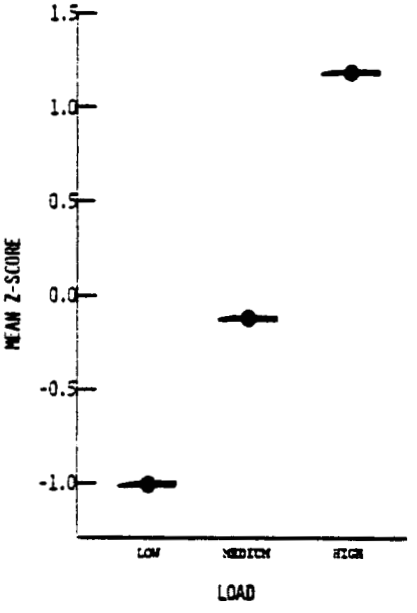

Figure 3. Mean z-scores for mediational reaction time measure vs. load.

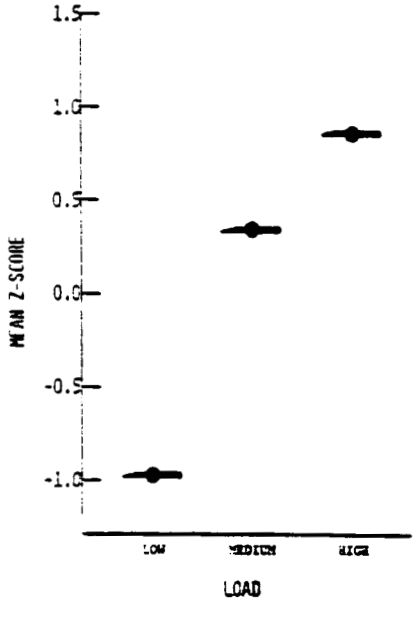

Figure 2. Mean z-scores for time estimation measure vs. load.

IV. PERCEPTUAL EXPERIMENT

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

JOHN G. CASALI and WALTER W. WIERWILLE
Department of Industrial Engineering and Operations Research
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

ABSTRACT

There are many flight task situations in which perceptual activity on the part of the pilot or aircrew member is emphasized.

Unfortunately, the sensitivity, that is, the relative ability of conventional workload estimation techniques to discriminate between perceptual load levels, is largely unknown. Similarly, the intrusion caused by employing workload techniques is unknown. Because of this lack of basic knowledge, an experiment comparing several workload techniques was conducted in an instrumented GAT-1B flight simulator.

The initial sensitivity and intrusion results of the experiment are reported in this paper, and a relative categorization of techniques is presented, based on demonstrated sensitivity.

INTRODUCTION

Numerous methods, test instruments, and analytical techniques have been purported as useful in the assessment of pilot/aircrew mental workload in flight-related tasks. This body of knowledge is the subject of several recent overviews, e.g., [1], [2], [3], [4].

A review of this literature revealed that little research effort has been directly applied to the problem of specifying a viable workload estimation technique for a given pilot/aircrew problem. Furthermore, the relative sensitivity and intrusion of most techniques has not been investigated. A sensitive workload estimation technique can be defined as one which reliably discriminates differences in mental loading requirements of a given task. Intrusion refers to an undesirable, artificial change in primary task performance, solely due to the concurrent use of an 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 measured workload level may not be representative of task requirements alone. Highly intrusive techniques may also create safety hazards.

The consequences of specifying a non-optimal 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 procedure or design.

Research Objective

In the research described herein, eight 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 processes.

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 (psychomotor, perceptual, mediational, communicative) in a single controlled experiment [5]. Therefore, this investigation concentrated on a single category: perceptual processes.

METHOD

Apparatus

All data were obtained in a Singer-Link GAT-1B moving-base (pitch, roll, yaw) flight simulator which had been modified to enable collection of workload estimation measures. Measures were obtained and processed on-line during each flight using an EAI-380 hybrid computer.

Experimental Design

Due to the dual objective of this experiment, it was most efficient to collect the two sets of data, one for sensitivity and one for intrusion, simultaneously.

Sensitivity design. A mixed, three-by-eight complete factorial design was used. Load level was the three-level within-subject variable. Workload estimation technique was the eight-level between-subjects variable. The use of six subjects per technique (a total of 48 subjects) enabled complete counterbalancing of load level presentation order across subjects. Five VFR-certified pilots and one IFR-certified pilot were assigned to each technique on the basis of their piloting experience in hours. Equivalent cross-sections of experience levels were represented in each technique.

Each pilot flew three experimental flights in the simulator. A single load level (low, medium, or high) was used in each flight. Load was manipulated by varying the rate and number of "redline" danger conditions presented on the oil pressure, oil temperature, cylinder head temperature, and fuel tank guages, and also on a carburetor ice warning LED. Subjects were instructed simply to detect the presence of a danger condition and identify it by pressing a corresponding pushbutton on the simulator instrument panel. A correct response alleviated the danger condition. No diagnosis or compensation of danger conditions was necessary, as their presence were instrument indications only and in no way affected aircraft performance. In the low load conditions, only the danger condition of carburetor icing was used. Icing occurred at an average rate of one every 50 seconds. The medium load condition was limited to left and right fuel tank problems and carburetor icing, occurring at an average rate of one failure per 10 seconds. In the high workload condition, danger indications occurred at an average rate of one per five seconds on all engine and fuel instruments, in addition to carburetor icing.

Eight workload estimation techniques were investigated in the sensitivity analysis. Included were opinion measures (Modified Cooper-Harper scale and Multi-Descriptor scale), secondary task measures (time estimation standard deviation and tapping regularity), physiological measures (pulse rate variability and respiration rate), and primary task measures (danger condition response time and aileron-elevator-rudder movements). The Modified Cooper-Harper scale was a modified version of the Cooper-Harper (1969) handling qualities rating scale [6]. The Multi-Descriptor scale tapped mental workload on a number of bipolar dimensions which were rated individually. Both rating scales and associated instructions appear in [7]. The experimental design for the sensitivity analysis was univariate, utilizing a single dependent measure called "score." "Score" represented the value obtained on each workload estimation technique. Between techniques, there were differences in scaling values for the obtained scores, such as breaths/min or control movements/s; therefore, all scores within a particular technique were converted to standard units (Z-scores) prior to statistical analysis.

Intrusion design. The experimental design for collection of intrusion data was identical to that of the sensitivity design, with the exception of the type of dependent measure used. The intrusion design was multivariate. Four primary task dependent measures were collected concurrently with the sensitivity dependent measure of "score" for each technique. The intrusion dependent measures were danger condition response time, control movements/s, and pitch and roll high-pass mean square (with filter cut-off frequency of 0.5 rad/s).

Experimental Task Procedures

Primary flight task procedure. The "primary task" refers to a particular segment of the experimental flight task during which workload level was manipulated and data were obtained. An approximate timeline showing the sequence of events during an experimental flight is shown in Figure 1.

The navigational control portion of the primary task was invariant in difficulty as load was varied solely via the danger condition task. Pilots were instructed to maintain adequate performance on all aspects of the primary task.

As shown in Figure 1, physiological, secondary task, and primary task measures were all obtained over a five-minute interval during each experimental flight task. Rating scale measurements were obtained immediately following the completion of the primary task, with the simulator in the autopilot mode.

RESULTS

Sensitivity Analysis

Sensitivity ANOVA. After data reduction and conversion of scores to stan-

dard units, an overall three-by-eight (load-by-technique) analysis of variance was performed. A significant main effect of load was revealed, $F(2,80) = 50.67$, $p = 0.0001$, indicating that the method of manipulating load level was indeed effective. Furthermore, and of immediate importance to subsequent analyses, a highly significant interaction of load-by-technique was revealed, $F(14,80) = 5.52$, $p = 0.0001$, suggesting that estimation techniques were differentially influenced by the loading task. Of course, due to the standarization procedure previously discussed, a main effect of technique was not possible, $F(7,40) = 0.00$, $p = 1.0000$.

Simple effects F-tests. The next step in the sensitivity analysis was to examine the load-by-technique effect to determine which particular workload estimation techniques were sensitive to changes in load. A Hartley's F-max test revealed that the data were homogeneous, F-max $(8,10) = 4.46$, $p > 0.01$; therefore, simple effects F-tests were performed. The simple effects tests revealed that all techniques except pulse rate variability and control movements were reliably sensitive to changes in loading on the perceptual task. Because of the large number of tests performed, only p-values for the simple effects F-tests are shown in column 4 of Table 1. Furthermore, an examination of the means revealed monotonic increases in mean values across load for all significant techniques.

Due to the ordinal nature of the two rating scales, nonparametric Friedman Rank Sum tests were performed in addition to the parametric simple effects F-tests on the scale data [8]. Again, significant differences were found among loading levels for the Modified Cooper-Harper ratings, $S'(2) = 10.000$, $p < 0.01$, and for the Multi-Descriptor ratings, $S'(2) = 8.8182$, $p < 0.025$.

Duncan's tests. These multiple-comparisons tests were employed to examine the locus and direction of the load effect on each technique found significant in the preceeding simple effects tests. The number of loading levels between which a workload estimation technique reliably disciminates is one indication of its relative sensitivity. Results of the Duncan's tests are shown in column 3 of Table 1. Categorization of the techniques' relative sensitivity to load is shown in column 2. If a technique showed a significant difference between all three loading levels 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 workload estimator.

Intrusion Analysis

Intrusion MANOVA. Due to the multiplicity of dependent measures, the intrusion data were subjected to a multivariate analysis of variance procedure. 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.

The results from the MANOVA were as follows. Load was the only independent effect to show significance, $F(2, 80) = 43.94$, $p = 0.0001$. This effect simply demonstrated that the four primary task measures, as a group, were reliably affected by changes in load. This provided evidence in addition to the sensitivity results that perceptual load was effectively manipulated in the experiment. 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 load-by-technique effect in the MANOVA. However, because neither of these effects approached statistical significance, $F(7,40) = 0.77$, $p = 0.7923$, and $F(14,80) = 1.14$, $p = 0.2460$, respectively, intrusion appeared not to have been a factor in this study.

CONCLUSIONS

In the study described herein, 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) did show sensitivity to changes in perceptual load. All significant techniques displayed monotonic increases in measured values across the three loading levels.

Both rating scale measures proved to be quite useful. These results reiterate those of others, (e.g., [9], [10], [11] indicating that with highly-trained populations such as pilots, rating scales are sensitive measurement instruments.

Both secondary task measures (time estimation and tapping regularity) exhibited considerable sensitivity to perceptual load. The results are in general agreement with those of others, such as Hart [12], and Michon [13]. However, these measures do not lend themselves to full-scale aircraft application quite as readily as the rating scales do.

From the results of the Duncan's tests, respiration rate appears sensitive to widespread changes in perceptual load. However, when comparing the results from this study with others, (e.g., [9]. [14]) it is apparent that respiration rate is a highly task-specific measure.

In contrast, pulse rate variability was not sensitive to changes in perceptual load in this study. There is some previous research evidence that pulse rate variability tends to decrease with increased loading in flight-related tasks emphasizing psychomotor behaviors (e.g., [15], [16]). However, the results of the present study coincide with the very recent results of Connor [9], who reported that heart rate variability did not reliably change as turbulence and aircraft stability were varied in a flight task emphasizing psychomotor load.

Finally, this study demonstrated that primary task measures are also quite task-specific and therefore must be selected with task objectives in mind. Control (aileron-elevator-rudder) 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.

REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] Casali, J. G. A sensitivity/intrusion comparison of mental workload estimation techniques using a simulated flight task emphasizing perceptual piloting behaviors. Doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1982.

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

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

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

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

ACKNOWLEDGEMENTS

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

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 1. Experimental flight task procedural timeline.

TABLE 1

Relative Sensitivity Categorization of Techniques

1	2	3			4	5
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

V. COMMUNICATIONS EXPERIMENT

A Comparative Evaluation of Rating Scale, Secondary Task, Physiological, and Primary Task Workload Estimation Techniques in a Simulated Flight Task Emphasizing Communications Load.

JOHN G. CASALI and WALTER W. WIERWILLE, Virginia Polytechnic Institute and State University.

Sixteen potential metrics of pilot mental workload were investigated in regard to their relative sensitivity to communication load and their differential intrusion on primary task performance. A moving-base flight simulator was used to present three cross-country flights to each of 30 subject pilots, each flight varying only in the difficulty of the inherent communications requirements. With the exception of the rating scale measures, which were obtained immediately post-flight, all measures were taken over a seven minute segment of the flight task. The results indicated that both the Modified Cooper-Harper and the workload Multi-descriptor rating scales were reliably sensitive to changes in communications load. Also, the secondary task measure of time estimation and the physiological measure of pupil diameter yielded sensitivity. As expected, those primary task measures which were direct measures of communicative performance were also sensitive to load, while aircraft control primary task measures were not, attesting to the task-specificity of such measures. Finally, the intrusion analysis revealed no differential interference between workload measures.

INTRODUCTION

The assessment of pilot and aircrew mental workload is a topic of critical importance. As a result, interest in workload measurement has surfaced in the research literature of late. Workload issues have bearing on aircraft certification, aviation safety, cockpit design, and aircraft tactical effectiveness.

Mental workload requirements placed on the pilot are quite variable and difficult to bound. If requirements are moderately excessive, there may be a measureable degradation in the performance of simple supportive tasks associated with flying, such as routine communications or engine instrument monitoring. Often, more imminent tasks will receive the focus of attention and other tasks will be time-shared or even ignored completely. Mental overload 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 and aircrew with information assimilation, processing, and action responsibilities is certainly undesirable, both from operational efficiency and safety standpoints. Recent advances in avionics and improvements in operating procedures have been directed toward easing the load on the pilot and aircrew. Despite the impact and success of these advances, the ultimate success of any flight mission is predicated on the performance of a common denominator: the human pilot. For this reason, the need for accurate empirical assessment of pilot mental workload is particularly cogent.

The recent thrust toward mental workload quantification has resulted in numerous measures, test instruments, and analytical procedures (collectively termed "workload estimation techniques") purported as useful in pilot workload assessment. Several recent summary reports have reviewed a variety of mental

workload metrics (e.g., Butterbaugh, 1978; Roscoe, 1978; Wierwille and Williges, 1978 and 1980). A review of these reports and related workload literature reveals that while several workload estimation techniques have been investigated on an individual basis in relatively simple laboratory tasks, very little effort has been directed toward comparing the differential effectiveness of a variety of techniques, in the assessment of incident workload for given situations encountered in simulated or actual flight. As a result of this research void, little basis has existed for selecting optimal estimation techniques for a given flight problem, situation, or task. With this in mind, a series of workload estimation technique comparison studies were undertaken in the Vehicle Simulation Laboratory at Virginia Tech. In each study, a simulated flight task emphasizing one of the four categories of Berliner, Angell, and Shearer's (1964) universal behaviors was used as a primary task. A variety of physiological, opinion, secondary task, and primary task workload metrics were "compared" in each study. The first three studies, reported elsewhere, emphasized psychomotor activity (Wierwille and Connor, 1983), mediational processes (Rahimi and Wierwille, 1982), and perceptual processes (Casali and Wierwille, 1982). The present study, reported herein, emphasized communicative activity in the simulated flight task.

There are many flight situations in which the aircraft pilot or crew are subjected to high communications loads. In high-traffic terminal areas, workload requirements due to tasks such as holding pattern maintainence, chart reading, and traffic avoidance may already be quite high. The communications tasks in these areas are particularly frequent, difficult, and of high priority, further contributing to the total workload. Pilots must listen for messages and instructions referenced only by their aircraft registration number, often under conditions of dense communication traffic and poor radio

reception. A message intended for a particular aircraft is often embedded in a near-continuous string of extraneous messages. Therefore, while the pilot may not be involved in constant, active two-way conversation, he/she nevertheless must devote constant attention to the communications aspect of the flight task to detect messages and instructions intended specifically for his/her aircraft.

Research Objective

In the experiment described here, sixteen different workload estimation techniques were comparatively evaluated in a simulated flight task in which pilot loading level was varied only on a communications dimension. The major objective of the comparison process was to establish the relative sensitivity and the differential intrusion of each technique in application to a realistic flight task which stressed the use of speech communication skills and behaviors.

Sensitivity and intrusion are two factors which have direct bearing on the utility and viability of any mental workload estimation technique. A sensitive workload estimation technique is defined as one which can reliably discriminate between different levels of operator loading. The load (workload) levels or changes in loading to be investigated should reflect alternatives of interest in the test situation. Ultimately, sensitive techniques may be applicable to workload assessment problems in which the levels to be compared are dictated by such factors as alternative configurations of on-board equipment, alternative procedures, differing mission objectives and goals, and alternative crew sizes. Changes in mental workload influenced by these factors may be large enough to cause degradations in aircraft control, increase attentional demands to the point that time-sharing and omission of tasks occurs, and reduce pilot ability in accomplishing mission objectives. It is these

substantial changes in mental workload impositions that state-of-the-art workload measurement techniques attempt to assess.

Intrusion refers to an undesirable, artificial variance in instructed task performance, solely attributable to the concurrent use of a workload estimation technique, related procedure, or associated apparatus. Workload assessment is contaminated by the presence of intrusion. Primary task performance is altered artifactually and the indicated workload level may not be representative of task requirements alone. Furthermore, techniques which prove highly intrusive may degrade pilot control performance and create safety hazards if applied in actual flight.

Prior to the selection of a particular estimation technique for a given application, both the sensitivity and the intrusion properties of the technique should be known. The consequences of selecting a nonoptimal technique are quite serious. First, an insensitive technique which does not reliably detect shifts in mental loading in a particular process may mask true differences in workload. Also, intrusive techniques may alter performance of a fundamental flight-related task, invalidating workload assessment results, because the pilot behaves in an unusual manner. Such biases could ultimately lead to acceptance of a hazardous procedure or design, especially if overload or near-overload conditions are present.

EXPERIMENTAL METHOD

Subjects

A total of 30 pilots (29 males and one female) were used as subjects. All were volunteers and were paid a nominal gratuity for participating. Only subjects with a minimum of a VFR (visual flight rules) private pilot's license were permitted to participate. They ranged in piloting experience from 70 hours to 2500 hours with a mean of 379 hours.

In an effort to control for individual subject differences due to variations in logged flight time, assignment of subjects to experimental conditions was done on the basis of piloting experience in single-engine, general aviation aircraft in hours. After the number of piloting hours for each subject was determined, a rank ordering of all experience levels (in hours) was performed. This ranking was then divided into sextiles, with six subjects per sextile. One subject was then selected at random from each sextile and assigned to the first workload technique condition. This procedure continued for all five technique conditions, resulting in a cross-sectional representation of six subjects, one from each experience level sextile, for each technique condition. The only additional stipulation was that five of the subjects in each condition were VFR-certified and the sixth subject was IFR (instrument flight rules)-certified.

Fundamental Apparatus

All data were obtained in a Singer-Link model GAT-1B flight simulator located in the Virginia Tech Vehicle Simulation Laboratory. This moving-base (pitch, roll, yaw) simulator was extensively modified to enable collection of workload-related measures. Necessary signals were obtained and processed on line during each flight using an EAI-380 hybrid computer which was interfaced

with the simulator's computational dynamics system. Measures were also displayed on a Sanborn model 350 stripchart recorder for permanent record.

In all conditions, the simulator was operated in the instrument flight mode under flourescent room lighting, with translucent blinders over each cockpit window. This prevented subject distraction from irrelevant laboratory cues and helped maintain a constant level of cockpit illumination.

Pilots in the GAT-1B communicated with the experimenter via a lapel microphone and cockpit speaker system. The pilot's microphone was actuated by a push-to-talk switch mounted on the control yoke of the simulator. For the communications aspect of each simulated flight task, a tape-recorded message consisting of communications stimuli and flight-related instructions was presented over the cockpit speaker. A BIC model T-1 dual-channel cassette recorder was used for playback of the communications tapes. Pilots' verbal responses and signals representing push-to-talk switch actuations were recorded on separate channels of tape on a Sony model TC-270 reel-to-reel recorder. These tapes were later used in the analysis of responses to the communications task. A full description of the communications stimuli appears in the experimental design section.

Workload Estimation Measures and Apparatus

The sixteen workload estimation techniques investigated in this experiment included some from each of the four categories described by the Wierwille and Williges (1978) classification scheme: (1) opinion (rating scale) measures, (2) spare mental capacity (secondary task) measures, (3) physiological measures, and (4) primary task measures. Each of the measures (abbreviations shown in parentheses) will be briefly discussed below.

Modified Cooper-Harper scale (MCH) - Opinion. A modified version of the Cooper-Harper (1969) aircraft handling qualities rating scale was used. The

original Cooper-Harper scale is directed at aircraft handling qualities rating and as such does not lend itself well to ratings of more general workload dimensions. Therefore, a modified version of the Cooper-Harper scale was developed by the experimenters. The flow diagram of the original scale was retained, but the verbal descriptors were changed. The modified scale, while still ordinal, is applicable to a wider variety of task workload applications, including tasks with a communications emphasis. Immediately after each data flight the actual measure obtained was the rated scale value, 1-10, given by the pilot. The Modified Cooper-Harper scale and related instructions appear in Casali (1982) and in Rahimi (1982).

Multi-descriptor scale (MD) - Opinion. The workload Multi-descriptor rating scale, also developed at Virginia Tech, consisted of seven workload "descriptors" which were each rated independently of each other immediately after a flight (Casali, 1982). This scale is to an extent based on the bipolar rating scale research conducted by Bird (1982). The ratings were done on a linear equal-appearing interval scale of 11 discrete steps including a center point. Descriptors included attentional demand, error level, difficulty, task complexity, mental workload, stress level, and overload level. For each flight, a single rating for the Multi-descriptor scale was computed as the arithmetic mean of the obtained ratings on the seven descriptors.

Time estimation standard deviation (TE) - Secondary task. The method of production was used for the time estimation task. Subjects performed this task while they performed the primary flight task (Hart, 1976). Subjects were prompted to begin mental production of a 10-second time interval by a one second tape-recorded tone at 750 Hz played back over the cockpit speaker. These prompt tones were so sequenced that they were not superimposed on

(transmitted over) the recorded message for the communications task. Each pre-recorded prompt tone was separated by approximately 21 seconds on tape. After hearing the tone, the subject was instructed to press a yoke-mounted microswitch once to start the interval and once again to signal a 10-second lapse. These switch depressions provided start and stop signals for an electronic timing circuit which provided a display of interval length.

It should be noted that the switch used for time estimation (and for tapping regularity) was actuated by the right thumb. The transmit switch used for communications was actuated by the left thumb. The control yoke was clearly labeled so that confusion of the two switches would be minimized.

For each flight, the standard deviation (in seconds) of the subject's time interval estimates was computed. On some trials, the subject did not initiate the beginning of an interval after the prompt. These trials were unusable and were deleted from the computation. Trials on which the subject initiated the interval estimate but did not signal the end before the occurrence of the next prompt were scored as 20 seconds in length. Since 20 seconds was the time between the end of one tone and the beginning of the next, the 20 second score represented the <u>minimum</u> possible (most conservative) length of the subject's unfinished estimate.

<u>Tapping regularity (TR) - Secondary task.</u> Subjects in the tapping regularity condition were instructed to tap (depress) the yoke-mounted microswitch as regularly or rhythmically as possible at a rate of one tap every two seconds. Again, this secondary task was performed concurrently with the primary flight task and represented a variation of Michon's (1966) tapping task procedure. The signals (taps) from the microswitch were inputted to a program on the EAI-380 computer and the program output was analyzed for the length of time between successive taps. These first difference values were

applied to the tapping regularity formula presented below, and a single regularity value was obtained for each flight. Intervals between taps of greater than five seconds in length were mapped to a value of five seconds in the the computations, because longer intervals were not representative of the uninstructed tapping rate and would heavily bias the formula.

The computational formula used for tapping regularity was

$$TR = \frac{\frac{1}{N_e}\sum_{i=1}^{N_e} \Delta t_{ie} - \frac{1}{N_b}\sum_{i=1}^{N_b} \Delta t_{ib}}{\frac{1}{N_b}\sum_{i=1}^{N_b} \Delta t_{ib}}$$

where Δt_i = the time between consecutive taps (the i^{th} interval)

N = the number of intervals

b = the subscript associated with the baseline run, in which tapping was performed alone.

e = the subscript associated with a data run in which tapping was performed simultaneously with the flight tasks.

Respiration rate (RR) - Physiological. Subjects' respiration rates in breaths-per-minute were obtained surreptitiously using a transducer fabricated in the Vehicle Simulation Laboratory (Casali and Wierwille, 1980). A strip-chart trace representing the subject's respiratory behavior, obtained during a flight, was later analyzed for this frequency measure.

Heart rate mean (HRM) and standard deviation (HRSD) - Physiological. These cardiovascular measures were sensed using a Hewlett-Packard plethysmograph -patient monitor system, model 7807C, and processed using the EAI-380 computer. At the end of a seven-minute data-recording period, values corresponding to heart rate mean and heart rate mean square (over the seven-minute

period) were read from the computer. From these scaled values, heart rate means and standard deviations in beats-per-minute were computed. It should be noted that heart rate mean values were obtained from a different set of subjects than heart rate standard deviations values so that a between-subjects comparison of the two measures could be performed.

Pupil diameter (PD) - Physiological. For continuous recording of subjects' pupil diameter size during a flight, a closed-circuit television (CCTV) system was used. Ambient illumination was held constant during a flight. A close-up shot of the subject's right eye was obtained with a Panasonic model PK-700 color video camera with 6:1 servo-controlled zoom lens. The camera signal was recorded on a Panasonic model NV-8310 videocassette recorder at a tape speed of 3.33 cm-per-second. After data collection, the videocassette recorder's freeze-frame capability was used to sample pupil-iris ratio during playback on a Setchell-Carlson model 12M918 monitor. The actual measure obtained for each flight consisted of the mean of all samples of pupil size divided by iris size (a dimensionless quantity), taken approximately every ten seconds during a run. As an extra precaution pupil diameter was measured only when the subject fixated on the attitude indicator.

Eyeblinks (EB) - Physiological. Utilizing the CCTV system, a frequency count of the number of eyeblinks (right eye) was obtained during a flight. The actual measure used was blinks-per-minute.

Eye fixations (EF) - Physiological. Again with the CCTV system, the number of subject eye fixations (fixations-per-minute) on any point on the simulator instrument panel was obtained during each flight.

Control movements (CM) - Primary task. During each experimental flight, the total number of elevator, aileron, and rudder inputs was tabulated on a Fluke model 1900-A digital counter, after being processed on the EAI-380

hybrid computer. This count was then divided by 420 seconds (seven 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 four percent of full movement range-per-second, after the derivative of control position passed through zero.

Pitch high pass mean square (PHPMS) and roll high pass mean square (RHPMS)- Primary task. The signals for these measures of flight task performance were obtained directly from the GAT-1B dynamics and inputted to high-pass filtering and mean square computational programs on the hybrid computer. Final values for each measure were in units of $(\text{radians})^2$. Low-frequency deviations were filtered out to insure that differences in aircraft trim level between pilots would not mask true deviations in heading and attitude control, as reflected in the pitch and roll angular excursions of the simulated aircraft. Filter cut-off frequency for both pitch and roll was 0.05 radian-per-second.

Errors of omission (ERRO) - Primary task. During a flight, an error of omission was counted when the subject failed to respond to the aircraft's designated call sign, "One-Four-India-Echo," or certain specified variations of that call sign, including "One-India-Four-Echo," "One-India-Echo-Four," "One-Echo-Four-India," "One-Echo-India-Four," and "One-Four-Echo-India." On the communications portion of the primary flight task, subjects were instructed to respond to the above "target" call signs by pressing the push-to-talk switch and verbalizing the word "now" to acknowledge detection. If a target call sign was missed by the subject, an error of omission was scored. The total number of errors of omission was obtained for each experimental flight.

Errors of commission (ERRC) - Primary task. An error of commission was said to occur if the subject pressed the push-to-talk switch and responded "now" in response to a call sign other than one of the six target signs. These call signs were extraneous. Again, a total count of commission errors was obtained during each flight.

Communications response time (CRT) - Primary task. For each correct response to the target call signs on the communications task, the response time in seconds was obtained. A mean response time was then computed for each data run. Response times were extracted after the completion of the experiment from the magnetic tape record of the subject's transmissions during flight. Response time was measured from the end of the spoken message (call sign) to the beginning of the subject's "now" response.

Experimental Design

Due to the dual objective of this study, it was most efficient, on the basis of conserving pilot resources, to collect two sets of data, one for the sensitivity analysis and one for the intrusion analysis simultaneously. A mixed three-by-five complete factorial design, shown in Table 1, was used for data collection. Communications load level was a fixed-effects, within-

(Insert Table 1 about here)

subject variable. Workload estimation technique group was a fixed-effects, between-subjects variable. Using six subjects (random-effects) per technique group, it was possible to completely counterbalance the presentation order of load levels across subjects to protect against habituation or practice effects.

Technique group independent variable. In three of the five technique groups, more than one workload estimation technique was obtained from each

subject. The selection of individual workload measures for inclusion in each group was done on the basis of their apparent independence and mutual lack of intrusion. The investigators' operational experience with these techniques, combined with the intrusion results from three other studies (Casali and Wierwille, 1982; Rahimi and Wierwille, 1982; Wierwille and Connor, 1983) were exercised in making the judgments for the groupings.

The second objective of the study, the intrusion investigation, provided another constant for assigning estimation techniques to groups. The experimental design for the intrusion investigation was structured to answer the fundamental question: Does the use of certain workload estimation techniques result in differential influence on known measures of primary task performance? Therefore, to enable comparison of workload measures as to their intrusion level, it was necessary to assign measures of interest to different technique groups. In particular, the following groups of measures were of interest to the differential intrusion investigation: eye behavior measures, respiration rate, heart rate standard deviation, tapping regularity, and time estimation. The experimental design for collection of intrusion data, concurrent with sensitivity data, was identical to that of the design matrix shown in Table 1. The major difference between the sensitivity design and the intrusion design was in the type of dependent measure used. The sensitivity design was univariate, in that a single dependent measure called "score" was obtained for each technique. The intrusion design was multivariate, in that five primary task dependent measures, including RHPMS, PHRMS, CM, ERRO, and ERRC were unobtrusively obtained in each workload technique group. Note that in groups one, two, and five, primary task measures were considered in the sensitivity analysis, while in all groups, the five primary task measures were obtained but considered only in the intrusion analysis.

Loading independent variable. The second independent variable shown in Table 1 varied with respect to the loading level associated with the flight task (primary task). Each subject pilot flew three experimental flights in the simulator. A single communications loading level was used in each flight.

The primary task in the experiment consisted of two major aspects: the aircraft control aspect and the communications aspect. Subjects were told to strive to maintain adequate (specified) performance on all aspects of the primary task.

For the aircraft control aspect, which was invariant in difficulty across the three flights, subjects were instructed to fly straight and level while maintaining the assigned altitude within $\pm$ 100 feet (30.5 meters), the assigned heading within $\pm$ 5 degrees (0.087 radians), and the assigned airspeed within $\pm$ 10 miles-per-hour (16.1 km-per-hour).

In addition to the control aspect of the flight task, the subject was required to attend to an eight-minute tape-recorded message presented over the cockpit speaker during flight. This was the communications aspect of the primary task. Part of the message consisted of a series of flight-related instructions to carry out certain commands including maneuvers, adjustments, and radio transmissions. These instructions varied only in their sequence of presentation across flights, and consisted of "tower" commands to change heading, change altitude, report present altitude and heading, report aircraft model (Cessna 150) and call sign (Cessna-One-Four-India-Echo), change altimeter setting, change radio frequency, and report airspeed. Between instructions, a series of abbreviated aircraft call signs were presented over the speaker. Each call sign or stimulus consisted of two single-digit numbers and two phonetic letters, combined in any order, such as Six-Alpha-Niner-Foxtrot and One-Seven-Bravo-Zulu. Subjects were instructed to depress the push-

to-talk switch and utter "now" whenever a target call sign was transmitted, and not to respond to extraneous call signs. A placard located on the simulator instrument panel was used to remind subjects of the particular target call signs.

Lists of call signs were randomly constructed from the numbers 0-9 and the international phonetic alphabet. The only exceptions to the random stimuli construction process was the selection of target call signs (e.g., One-India-Four-Echo) and presentations of target call sign alphanumerics (e.g., India-One-Four-Echo). The call sign recognition portion of the communications task was used to manipulate workload level between flights. In the low condition, call signs were presented at the rate of one every 12 seconds on average, and none of the extraneous (non-target) call signs were permutations of alphanumerics used in target call signs. The medium condition rate was one call sign every five seconds on average and 30 percent of the extraneous call signs were permutations. In the high workload condition, call signs were presented at an average rate of one every two seconds, and 40 percent of non-target call signs were target permutations.

On the communications task, subjects were instructed that adequate performance consisted of correctly carrying out all instructed commands and correctly identifying all target call signs as quickly as possible.

Procedure

After reading a description of the experiment and signing a consent form, subjects boarded the simulator and received instructions pertaining to their particular workload technique group condition. For instance, subjects in group 1 first received instructions concerning the use of the Modified Cooper-Harper scale. All subjects in groups 3 and 4 received practice on the particular secondary task to which they had been assigned.

The next set of instructions given to all subjects, regardless of their assigned technique group, specified the objectives, procedures, and adequate performance parameters of the primary task. Subjects then flew an eight-minute practice flight which gave them experience on all aspects of the primary task. Secondary task subjects (groups 3 and 4) again received practice on their particular secondary task, which was presented during the last three minutes of the practice flight concurrently with the communications task. Next, each subject flew three consecutive individual experimental (data) flights.

The three flights did not differ in terms of the aircraft control demands placed on the subject. A mild turbulence, having amplitude peaks of approximately 6 miles-per-hour (9.7 km/hr), was applied over the duration of each flight using the GAT-1B random gust generator. This turbulence forced the subject to scan the basic flight instruments and make control inputs, even when there were no commanded flight path changes taking place. Again, the flight task emphasized the speech communications aspects of pilot behavior.

In each of the three data flights, subjects were exposed to the loading level of the communications task one minute prior to data collection. All physiological, secondary task, and primary task measures were obtained over a continuous seven-minute interval, immediately following the initial one-minute exposure. Immediately after the data collection interval, rating scale subjects performed their ratings while the simulator was under autopilot control. Following the third data flight, the subject landed the simulated aircraft, was debriefed and paid for participation, and then was dismissed.

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 are treated individually.

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 communications loading, and (2) to establish the relative sensitivity among techniques to changes in loading levels, ultimately providing empirical grounds for selecting or rejecting a technique for applied workload investigations.

Sensitivity data reduction. First, the raw "scores" for each workload estimation technique, such as those in the form of scaled values from the EAI-380 amplifier circuits, were converted to numerical values applicable to data analysis. Next, the "reduced" scores for each technique were standardized across all three loading levels, i.e., converted to Z-scores, to insure that true sensitivity differences among techniques were not clouded by scaling measurement differences among various techniques, such as breaths-per-minute versus Modified Cooper-Harper scale values. This standardization procedure, having some precedent in the workload literature, is further discussed in Wierwille and Gutmann (1978) and Hicks and Wierwille (1979). All of the following sensitivity-related analyses were performed on the standardized data set using a SAS computer package (SAS Institute, 1979).

Overall senstivity ANOVA. The standardized data set was initially subjected to a three-by-sixteen analysis of various procedure (Table 2). All

(Insert Table 2 about here)

sixteen workload estimation techniques were considered as between-subjects measures in this analysis, even though those techniques within a single technique group were obtained from the same subject. However, it should be recalled that techniques within a single group were believed to be mutually independent and mutually unintrusive. Also, the purpose of this initial ANOVA was simply to determine (1) if the method of workload manipulation was effective (as evidenced by a load main effect) and, (2) if the estimation techniques were differentially influenced by the loading task (as evidenced by a load-by-technique interaction effect). In the latter case, the presence of the significant interaction, F (30,160) = 3.70, p = 0.0001, allowed the investigation of relative technique sensitivity to proceed. In the former case, the strong main effect of load, F (2,160) = 14.13, p = 0.0001, demonstrated that the method of manipulating communications load was indeed effective. Of course, due to the standardization of technique scores across loading levels, a main effect of technique was not possible, F (15,80) = 0.00, p = 1.0000.

Individual technique sensitivity ANOVAs. The reliable load-by-technique interaction obtained in the overall ANOVA suggested that the techniques were differentially sensitive to communications load. Therefore, the next step in the relative sensitivity analysis was to examine this interaction effect to determine which particular techniques varied with respect to load and with what degree of discriminability among loading levels. Simple-effects F-tests would typically be used in such a capacity. However, because these tests normally consider the L x S/T interaction from the overall ANOVA as the denominator (error) term in the F-ratios, their application is implicitly predicated on the assumption that variances among the various techniques are

homogeneous in nature. To test this assumption, a Hartley's F-max test was performed (Winer, 1971). This test revealed significant heterogeneity of variance among techniques, F-max (16,10) = 103.37, $p < 0.001$. Therefore, individual ANOVAs were applied to each of the 16 techniques versus load. These ANOVAs did not require homogenity of variance among techniques because each ANOVA utilized a unique error term (subject-by-load) for each technique. The summary tables for the individual ANOVAs appear in Table 3. As shown in the table, load had a significant effect on at least one technique from each

(Insert Table 3 about here)

of the four major technique categories. The remainder of the discussion of sensitivity results will be presented by workload technique category.

Opinion measures' sensitivity. Both opinion measures were highly sensitive to changes in load: the Modified Cooper-Harper scale at $F(2,10) = 13.57$, $p = 0.0014$ and the Multi-descriptor scale at $F(2,10) = 5.73$, $p = 0.0219$. Due to the apparent ordinality of the rating scale measurements, especially of the Modified Cooper-Harper scale, the ANOVA results were corroborated with the results of nonparametric Friedman rank-sum tests (Hollander and Wolfe, 1973) applied to each rating scale. The Freidman results were in agreement with those yielded by the ANOVAs, with $S'(2) = 11.14$, $p < 0.005$ for the Modified Cooper-Harper scale and $S'(2) = 9.00$, $p < 0.025$ for the Multi-descriptor scale.

The mean standardized scores obtained on the Modified Cooper-Harper scale are plotted as a function of load in Figure 1. To examine the locus and

(Insert Figure 1 about here)

direction of the load effect on the Modifed Cooper-Harper technique, a

Duncan's Multiple Range Test was applied (Duncan, 1975). The results of this test, designated by the letters within the graph in Figure 1, showed that the Modified Cooper-Harper scale ratings reliably increased in value between low and medium load, between low and high load, but not between medium and high load, (at a $p < 0.05$ criterion level).

The Multi-descriptor scale was not as sensitive to load (at $p < 0.05$) as the Modified Cooper-Harper scale (Figure 2). According to the Duncan's test

(Insert Figure 2 about here)

results, Multi-descriptor ratings reliably increased only between low and high loading levels at the 0.05 level.

Secondary task measures' sensitivity. The tapping regularity measure did not exhibit sensitivity at the 0.05 level (Table 3). However, the time estimation secondary task was quite sensitive to changes in communications load, $F(2,10) = 9.27$, $p = 0.0053$. The variability of subjects' time estimates increased as communications load increased. Specifically the Duncan's test revealed that time estimation standard deviation scores reliably differed between all loading levels with the exception of medium and high. This effect is plotted in Figure 3.

(Insert Figure 3 about here)

Physiological measures' sensitivity. The sole physiological measure to display sensitivity to changes in communications load was the pupil diameter measure, $F(2,10) = 5.90$, $p = 0.0203$. No other physiological measures approached significance in the individual ANOVAs (all at $p > 0.30$). The pupil diameter measure reliably discriminated low and medium, low and high, but not medium and high loading levels, according to the Duncan's test (Figure 4).

(Insert Figure 4 about here)

Primary task measures sensitivity. Each of the three primary task measures which directly reflected subjects' performance on the communications aspect of the primary task were significantly affected by communications load: the errors of omission measure at $F(2,10) = 9.79$, $p = 0.0044$, the errors of commission measure at $F(2,10) = 20.90$, $p = 0.0003$, and the communications response time measure at $F(2,10) = 4.15$, $p = 0.0486$. As shown in Table 3, none of the aircraft control primary task measures (control movements, pitch high-pass mean square, and roll high-pass mean square) were reliably affected by changes in communications load (all at $p > 0.40$). There were more errors of omission and more errors of commission as the communications burden increased, as shown in Figures 5 and 6. The omission measure reliably discriminated between low and medium, and between low and high, but not between

(Insert Figure 5 and 6 about here)

medium and high loading levels, according to the Duncan's test (Figure 5). Commission errors were significantly greater in number under high load conditions than under medium or low load conditions (Figure 6). But the commission errors under medium load were not significantly greater than under low load. The response time measure showed a monotonic decrease from low load to medium load to high load; however, the decrease was only significant between the low and high levels (Figure 7).

(Insert Figure 7 about here)

Intrusion Analysis

The objective of the intrusion analysis was to investigate the potential presence of undesirable, artificial changes in primary task performance that were attributable to the introduction of a particular group of workload estimation techniques and associated apparatus. As previously discussed, the intrusion analysis was applied to five primary task measures which were obtained in all cells of the experimental design concurrently with the sensitivity data (Table 1). This design was structured to answer the initial multivariate question: Were the workload technique groups differentially responsible for changes in performance on the primary task measures of errors of omission, errors of commission, control movements, pitch high pass mean square, and roll high pass mean square?

Intrusion MANOVA. Due to the multiplicity of dependent measures, it was necessary to apply a multivariate analysis of variance (MANOVA) to the intrusion data set (Table 4). With the MANOVA, the Wilk's U criterion values were obtained for all independent effects shown in the summary table (Cramer, 1972). The MANOVA demonstrated that load was the only significant independent effect, U (2,50) = 0.1640, $p < 0.01$. This result simply corroborated the significant main effect of load found in the overall sensitivity ANOVA (Table 2), demonstrating that communications load was indeed effectively manipulated in the experiment. The MANOVA load main effect, however, had no bearing on the interpretation of the intrusion analysis. If differential intrusion among technique groups had occurred, it would have been manifested as a significant technique effect or embedded in a significant load-by-technique effect. Because technique demonstrated no main effect, U (4,25) = 0.34, $p > 0.05$, and because there was no significant interaction of load with technique, U (8,50) = 0.38, $p > 0.05$, differential intrusion appeared not to have been a factor in this study.

CONCLUSIONS

Sensitivity Conclusions

Of the 16 workload measures investigated in this study, seven of the measures yielded some sensitivity to communications load while nine did not. The insensitive measures will be discussed first.

Insensitive techniques. As shown in Table 3, eight of the nine insensitive measures were not close to statistical significance (all at $p > 0.36$). However, the tapping regularity measure approached significance at $p < 0.0781$, but did not meet the selected $p < 0.05$ cutoff level chosen by the investigators. The relatively conservative $p < 0.05$ level was adhered to because of the probability of committing a Type I error due to chance, considering the large number (16) of techniques investigated. In any case, an estimate of the sample size required for obtaining significance at the $p < 0.05$ level was computed for all insensitive techniques using a power test (Keppel, 1973). This estimate of the number of subject required provides a means of comparing the discrimination power of the insensitive techniques with that of the sensitive techniques, which each yielded significance at least at the $p < 0.05$ level with only six subjects. The results of the sample size estimates are presented in Table 5.

(Insert Table 5 about here)

Interestingly, at least one technique from each category exhibited a reliable effect of load. Each category will be discussed separately.

Opinion techniques. Both of the rating scales demonstrated distinct monotonic increases in pilots' ratings of workload as a function of communications load. The Modified Cooper-Harper scale showed discrimination ability within each pair of loading levels except medium-high. However, the

medium-high pair also showed a similar compression effect with two other measures which displayed equal discriminability to the Modified Cooper-Harper scale, including time estimation and errors of omission. Despite the lack of significance of the scale within the medium-high pair, Modified Cooper-Harper scale values did follow a monotonically-increasing trend (Figure 1). The Multi-descriptor scale was revealed less sensitive to adjacent workload levels (low-medium and medium-high) than the Modified Cooper-Harper scale. From these findings, it appears that the Multi-descriptor scale, in its present form, would only be applicable to workload assessment or comparison situations in which differences in the communications burden between alternative situations are known beforehand to be widespread. Perhaps due to the specificity of communications tasks, new workload descriptors need to be incorporated into the scale prior to further investigation of the scale's sensitivity to communications-type load.

The rating scale results of the present study are in close agreement with those of other recent studies. The Multi-descriptor and especially the Modified Cooper-Harper scale were found to be among the most sensitive of techniques in simulated flight tasks emphasizing psychomotor load (Wierwille and Connor, 1983), perceptual load (Casali and Wierwille, 1982), and mediational load (Rahimi and Wierwille, 1982).

Secondary task techniques. The variability of pilot's time estimates reliably and monotonically increased with increase in communications load (Figure 3). This increase was significant for all but the medium-high pair, where compression again occurred. One explanation for the increased variability of time estimates in higher communications workload conditions is based on the strategy that the "estimator" adopts (Hart and McPherson, 1976). In low or moderate primary task load situations, the subject may be able to

make a conscious, sustained effort to monitor the passage of time—an active estimate. However, as primary task load increases, resulting in competition between concurrent activities, the subject's attention is diverted from the active mode of time estimation. The subject may begin to estimate elapsed time by memory, basing the estimates on the recall of events that occurred during the period since the end of the last estimate. In this "retrospective" mode, estimates would be expected to become more variable in length due to the interference of concurrent activity. Similar results for the time estimation standard deviation measure were obtained in the psychomotor, perceptual, and mediational studies mentioned above.

Physiological techniques. Pupil diameter was the sole physiological measure to reflect changes in communications load reliably. However, this effect must be considered somewhat suspect due to the double-values function (non-monotonic) shown in the plot of pupil diameter versus load (Figure 4). Also, the apparatus for obtaining the pupil diameter measurements, the CCTV system, was not the most sophisticated available. The CCTV system was originally intended for use in obtaining eye behavior frequency-type measures, which did not require as high degree of resolution as pupillary measurement, such as eyeblink count and eye fixations per unit time.

Primary task techniques. The introduction of high levels of communications load apparently did not degrade the subject pilots' ability to control the aircraft, as evidenced by the distinct lack of significance for the primary task measures of aileron-elevator-rudder movements, pitch deviation, and roll deviation. However, the primary task measures which reflected communications task performance each did show changes with load. In particular, both the number of omission errors and commmission errors on the call sign recognition task reliably and monotonically increased as load increased.

Interestingly, the response time measure showed a reverse trend. Response times to target call signs decreased as communications load increased, and the amount of decrease was significant only between low and high loading levels (Figure 7). At first glance this result may appear spurious, but at least one explanation may exist. It was particularly apparent in low loading conditions that because there were long periods of silence between the presentations of call signs, subjects tended to forget the presence of the communications part of the primary task and concentrate on maintaining instructed flight parameters. When a call sign did occur following a period of silence, the subject pilot may not have been "primed" for the response. This would have resulted in the longer response times for the low condition. In the high condition the call signs were presented almost continuously with few periods of silence. Therefore, the subject was constantly reminded of the imminence of a target call sign simply by the ongoing taped presentation of extraneous call signs. Perhaps the continuity of the high workload call sign recognition task raised the subject's level of awareness and increased the preparedness to respond, resulting in shorter response time.

Intrusion Conclusions

Workload measurement appeared not to be contaminated by intrusion in this experiment. Due to the duality of the experimental design for investigating sensitivity as well as intrusion, only differential intrusion among technique groups could be assessed. The finding of no significant intrusion in the present study was in general agreement with the three earlier studies on psychomotor, perceptual, and mediational load with one exception. The Rahimi and Wierwille (1982) study revealed that time estimation and heart rate mean measures had an intrusive effect on mediational primary task measures. The heart rate measure's effect is difficult to explain because of the

unobtrusive nature of the plethysmograph system. The time estimation measure, due to the fact that it in itself creates an increase in concurrent activity, certainly would be expected to interfere or at least compete with primary task performance. One explanation for the difference in time estimation findings for the Rahimi and Wierwille study and the present study is as follows. In the former study, the time estimation "prompt" stimuli and the mediational task stimuli were presented to the subject through separate sensory input channels: the prompts were auditorially-presented and the mediational stimuli were visually-presented. It appeared that subjects may have attempted to time-share between inputs, causing the intrusion. In the present study, both the time estimation prompts and the communications task were presented auditorially. Subjects were observed to disregard the secondary task at times when the communications burden was high. Perhaps the log-jamming of two tasks on one sensory channel made time-sharing more difficult, and subjects tended to "blank" the time estimation task in favor of the more critical communications task. As a result, the secondary task did not interfere substantially with primary task performance.

SUMMARY

The results of this study suggest that with highly-trained populations such as pilots, rating scales are sensitive measurement instruments. Rating scales are particularly attractive also because they are inexpensive, unobtrusive, easily administered, and readily transferable to full-scale aircraft and to a wide range of tasks. One secondary task measure, that of time standard deviation, also exhibited considerable sensitivity to communications load. However, this measure is not so easy to implement in full-scale aircraft as a rating scale and also may intrude on primary task performance if workload conditions are spread further apart than in this study, or if other sensory input modalities are overburdened, resulting in time-sharing. Physiological measures, as a whole, were not sensitive metrics of communications load nor were aircraft control-related primary task measures. However, primary task measures which directly reflected instructed performance on the communications task were most discriminating. Of course, the task specificity of these measures restricts their utility for application to other primary tasks.

ACKNOWLEDGEMENTS

The authors wish to thank Ms. Julie H. Skipper for her valuable help with the scores computations and statistical analyses. The authors also wish to thank Ms. Sandra Hart, NASA Ames Research Center, who served as grant technical monitor for this project (NAG-217).

REFERENCES

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.

Bird, K. L. Subjective rating scales as a workload assessment technique. Proceedings of the Seventeenth Annual Conference on Manual Control, Pasadena, CA: Jet Propulsion Laboratory, Publ. No. 81-95, June, 1982, 33-39.

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.

Casali, J. G. A sensitivity/intrusion comparison of mental workload estimation techniques using a simulated flight task emphasizing perceptual piloting behaviors. Doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1982.

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.

Casali, J. G. and Wierwille, W. W. A sensitivity/intrusion comparison of mental workload estimation techniques using a flight task emphasizing perceptual piloting activities. Proceedings of the 1982 International IEEE Conference on Cybernetics and Society, Seattle, Washington, October, 1982, 598-602.

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.

Cramer, C. Y. A first course in methods of multivariate analysis. Blacksburg, Virginia: Author, 1972.

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

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. 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.

Hicks, T. G. and Wierwille, W. W. Comparison of five mental workload assessment procedures in a moving-base driving simulator. Human Factors, 1979, 21, 192-143.

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

Keppel, G. Design and analysis: A researcher's handbook. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1973.

Michon, J. A. Tapping regularity as a measure of perceptual motor load. Ergonomics, 1966, 9, 402-412.

Rahimi, M. Evaluation of workload estimation techniques in simulated piloting tasks emphasizing mediational activity. Doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1982.

Rahimi, M. and Wierwille, W. W. Evaluation of the sensitivity and intrusion of workload estimation techniques in piloting tasks emphasizing mediational activity. Proceedings of the 1982 International Conference on Cybernetics and Society, Seattle, Washington, October, 1982, 593-597.

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

SAS Institute, Inc. SAS User's guide; 1979 edition. Cary, North Carolina, Author, 1979.

Wierwille, W. W. and Connor, S. A. Evaluation of twenty pilot workload assessment measures using a psychomotor task in a moving-base aircraft simulator. Human Factors, 1983, 25 (In press).

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.

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

TABLE 1

Experimental Design Matrix

COMMUNICATIONS LOAD (W/S)

Group Number (B/S)	LOW	MEDIUM	HIGH
1	$S_1 - S_6$	$S_1 - S_6$	$S_1 - S_6$
2	$S_7 - S_{12}$	$S_7 - S_{12}$	$S_7 - S_{12}$
3	$S_{13} - S_{18}$	$S_{13} - S_{18}$	$S_{13} - S_{18}$
4	$S_{19} - S_{24}$	$S_{19} - S_{24}$	$S_{19} - S_{24}$
5	$S_{25} - S_{30}$	$S_{25} - S_{30}$	$S_{25} - S_{30}$

WORKLOAD TECHNIQUE GROUPS:

Group Number	Techniques Included
1	MCH, HRM, PD, EB, EF, RHPMS, CRT
2	MD, HRSD, PHPMS, ERRC
3	TR
4	TE
5	RR, ERRO, CM

TABLE 2

ANOVA Summary Table for Overall Technique by Load Sensitivity Analysis

Source	df	SS	F	P
Between-Subjects				
Technique (T)	15	0.0000	0.00	1.0000
Subjects (S)/T	80	193.1008		
Within-Subjects				
Load (L)	2	7.4653	14.13	0.0001
L x T	30	29.3224	3.70	0.0001
L x S/T	160	42.2537		
Total	287	272.1429		

TABLE 3

Summary Tables for the Individual Technique ANOVAS

Source	df	SS	F	P
Technique: Modified Cooper-Harper Scale				
Load	2	5.6003	13.57	0.0014
Subject x Load	10	2.0635		
Technique: Multi-Descriptor Scale				
Load	2	2.9157	5.73	0.0219
Subject x Load	10	2.5429		
Technique: Time Estimation				
Load	2	3.9619	9.27	0.0053
Subject x Load	10	2.1362		
Technique: Tapping Regularity				
Load	2	0.1954	3.33	0.0781
Subject x Load	10	0.2936		
Technique: Respiration Rate				
Load	2	0.0696	0.17	0.8436
Subject x Load	10	2.0112		
Technique: Heart Rate Mean				
Load	2	0.0472	0.74	0.5037
Subject x Load	10	0.3210		
Technique: Heart Rate Standard Deviation				
Load	2	1.7718	0.74	0.5022
Subject x Load	10	11.9982		
Technique: Pupil Diameter				
Load	2	0.1369	5.90	0.0203
Subject x Load	10	0.1161		
Technique: Eye Blinks				
Load	2	0.0273	0.28	0.7588
Subject x Load	10	0.4814		

TABLE 3 Continued

Source	df	SS	F	P
Technique: Eye Fixations				
Load	2	0.2742	1.13	0.3619
Subject x Load	10	1.2166		
Technique: Control Movements				
Load	2	0.4540	0.69	0.5252
Subject x Load	10	3.3030		
Technique: Pitch High Pass Mean Square				
Load	2	1.3113	0.95	0.4182
Subject x Load	10	6.8846		
Technique: Roll High Pass Mean Square				
Load	2	0.0327	0.26	0.7771
Subject x Load	10	0.6318		
Technique: Errors of Omission				
Load	2	6.6719	9.79	0.0044
Subject x Load	10	3.4087		
Technique: Errors of Commission				
Load	2	11.5964	20.90	0.0003
Subject x Load	10	2.7738		
Technique: Communications Response Time				
Load	2	1.7211	4.15	0.0486
Subject x Load	10	2.0713		

TABLE 4

MANOVA Summary Table for Intrusion Analysis

Source	dv	df_H	df_E	U	P
Between-Subjects					
Technique (T)	5	4	25	0.34	>0.05
Subjects (S)/T		(Error Term for T)			
Within-Subject					
Load (L)	5	2	50	0.16	<0.01
L x T	5	8	50	0.38	>0.05
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

TABLE 5

Estimated Sample Sizes for Obtaining a Significant Load Effect for the Insensitive Techniques*

Technique	Required Sample Size
Opinion	
(all significant)	--
Secondary Task	
Tapping Regularity	15
Physiological	
Respiration Rate	38
Heart Rate Mean	>100
Heart Rate Standard Deviation	>100
Eye Blinks	42
Eye Fixations	>100
Primary Task	
Control Movements	98
Pitch High Pass Mean Square	>100
Roll High Pass Mean Square	42

*At $p < 0.05$ and with power > 0.80. Sample sizes greater than 100 are designated by ">100".

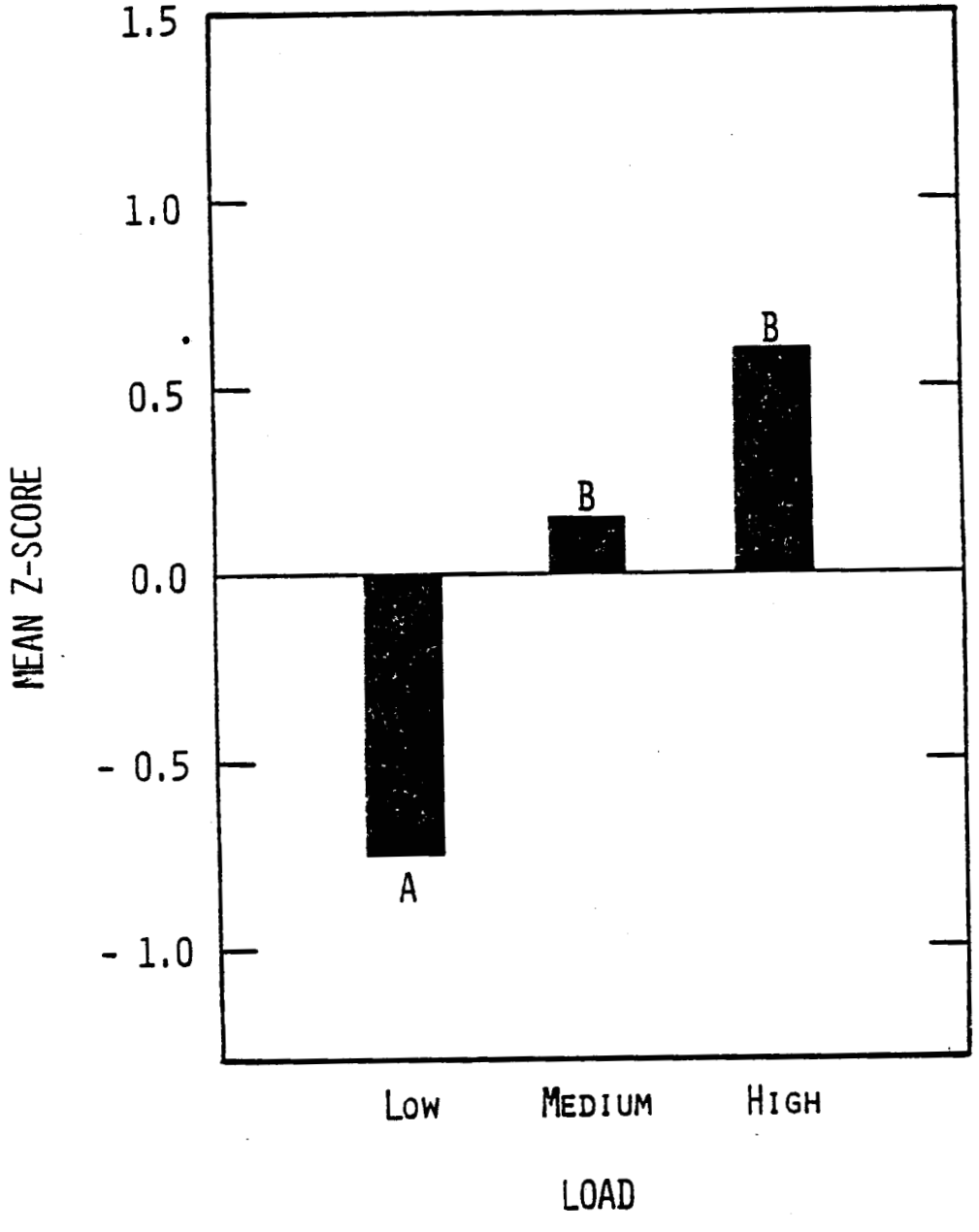

Figure 1. Effect of load on mean standardized scores for the Modified Cooper-Harper rating scale technique. (Means with different letters are significantly different, $p < 0.05$).

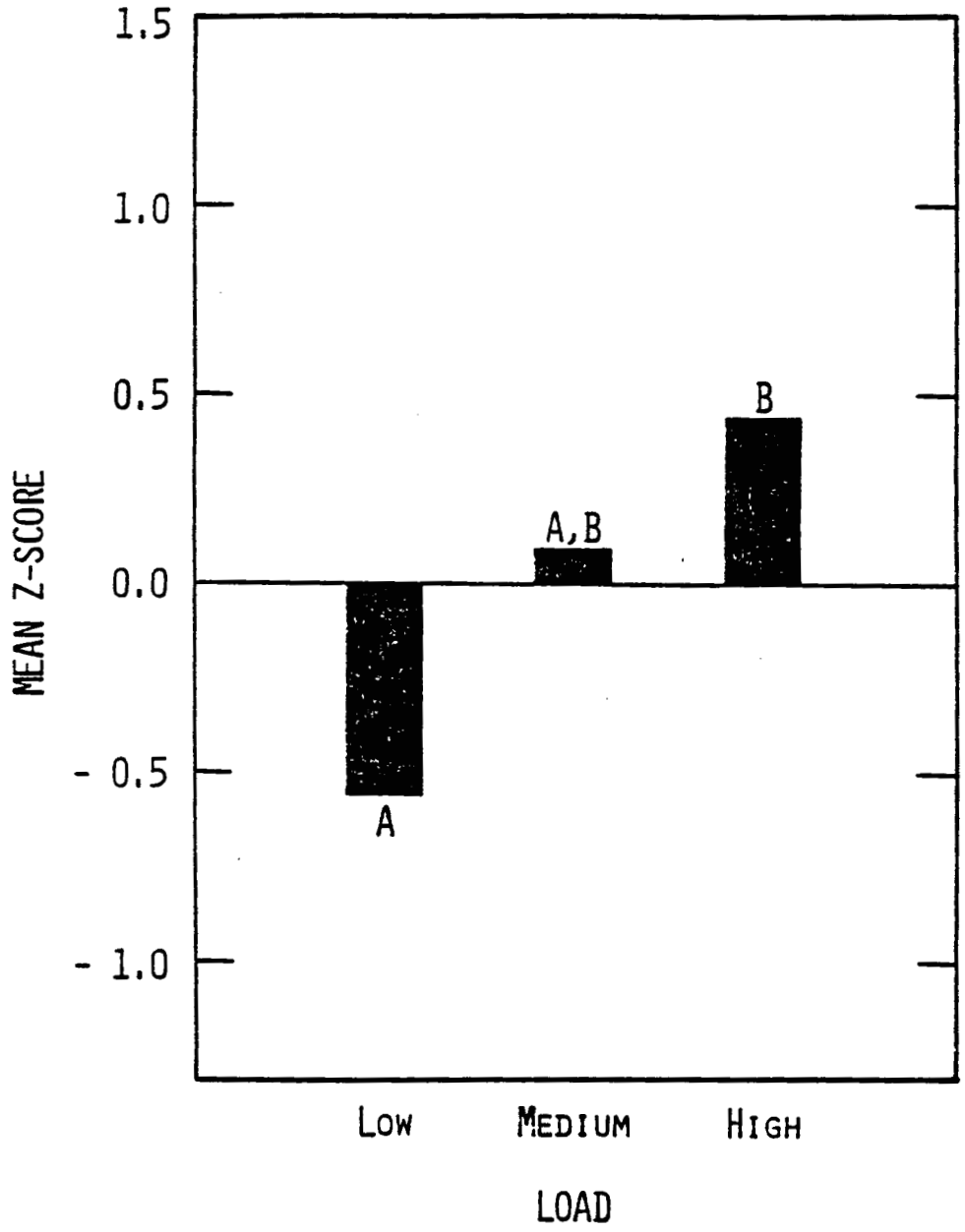

Figure 2. Effect of load on mean standardized score for the Multi-descriptor rating scale technique. (Means with different letters are significantly different, $p < 0.05$).

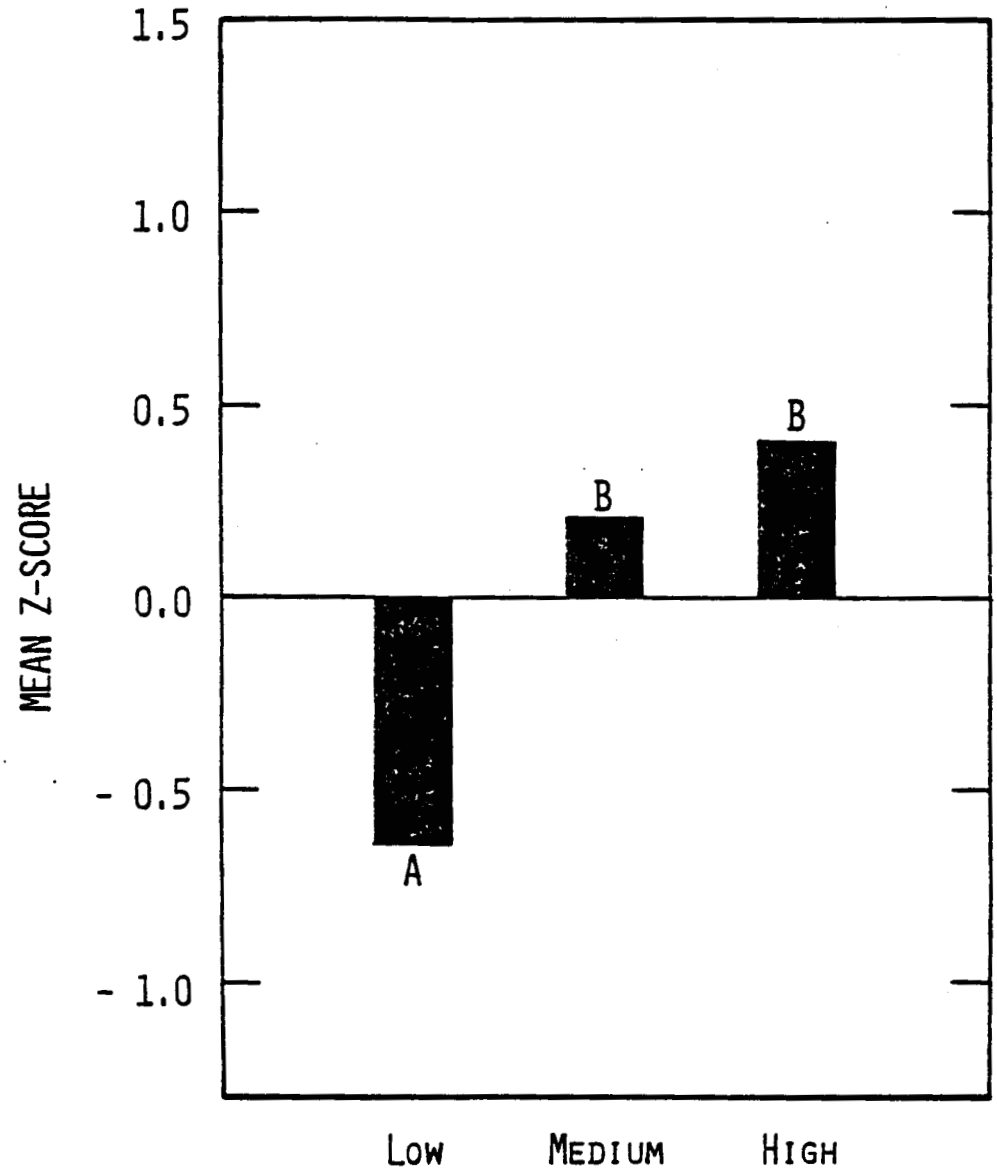

Figure 3. Effect of load on mean standardized scores for the time estimation standard deviation technique. (Means with different letters are significantly different, $p < 0.05$).

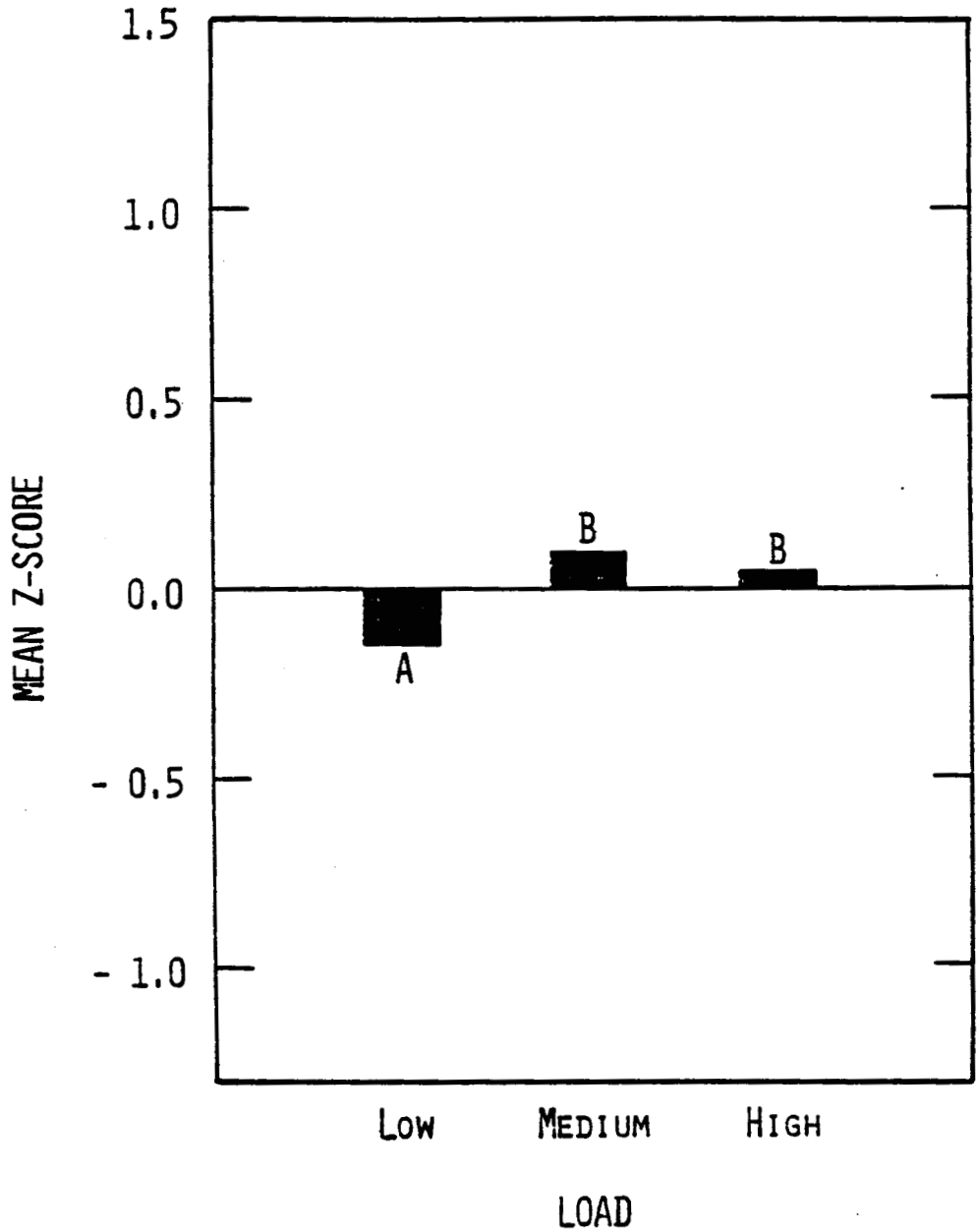

Figure 4. Effect of load on mean standarized scores for the pupil diameter technique. (Means with different letters are significantly different, $p < 0.05$).

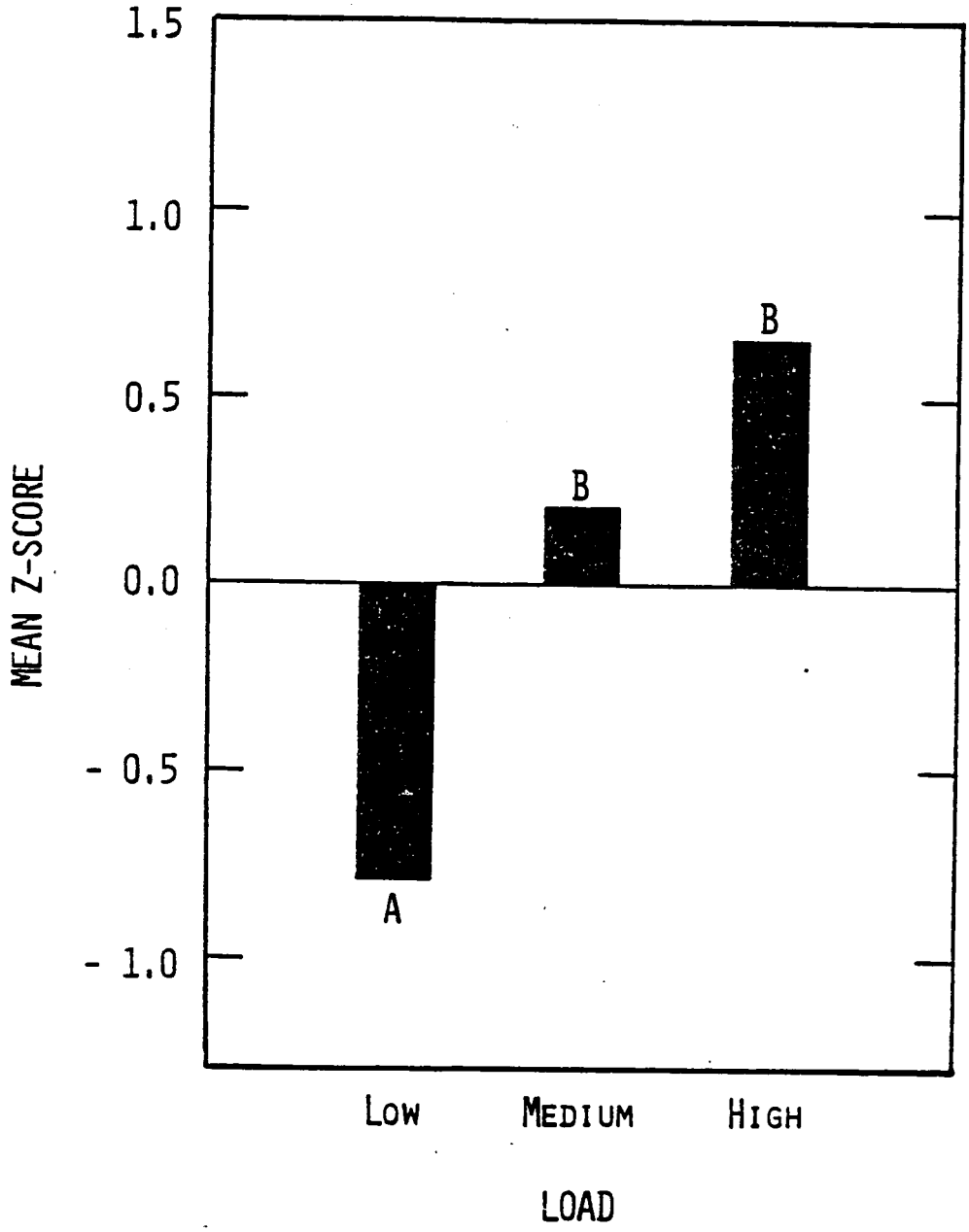

Figure 5. Effect of load on mean standardized scores for the errors of omission technique. (Means with different letters are significantly different, $p < 0.05$).

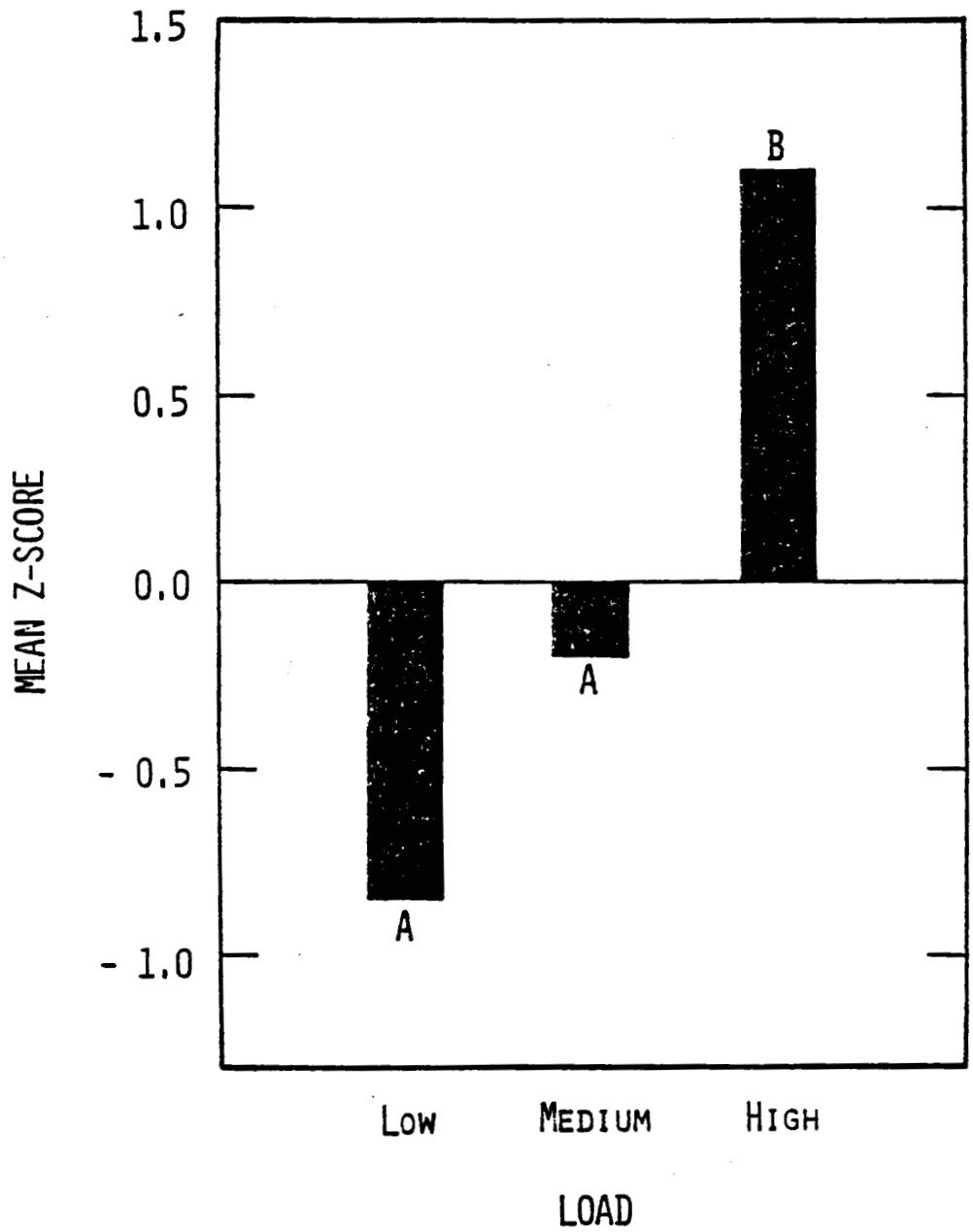

Figure 6. Effect of load on mean standardized scores for the errors of commission technique. (Means with different letters are significantly different, $p < 0.05$).

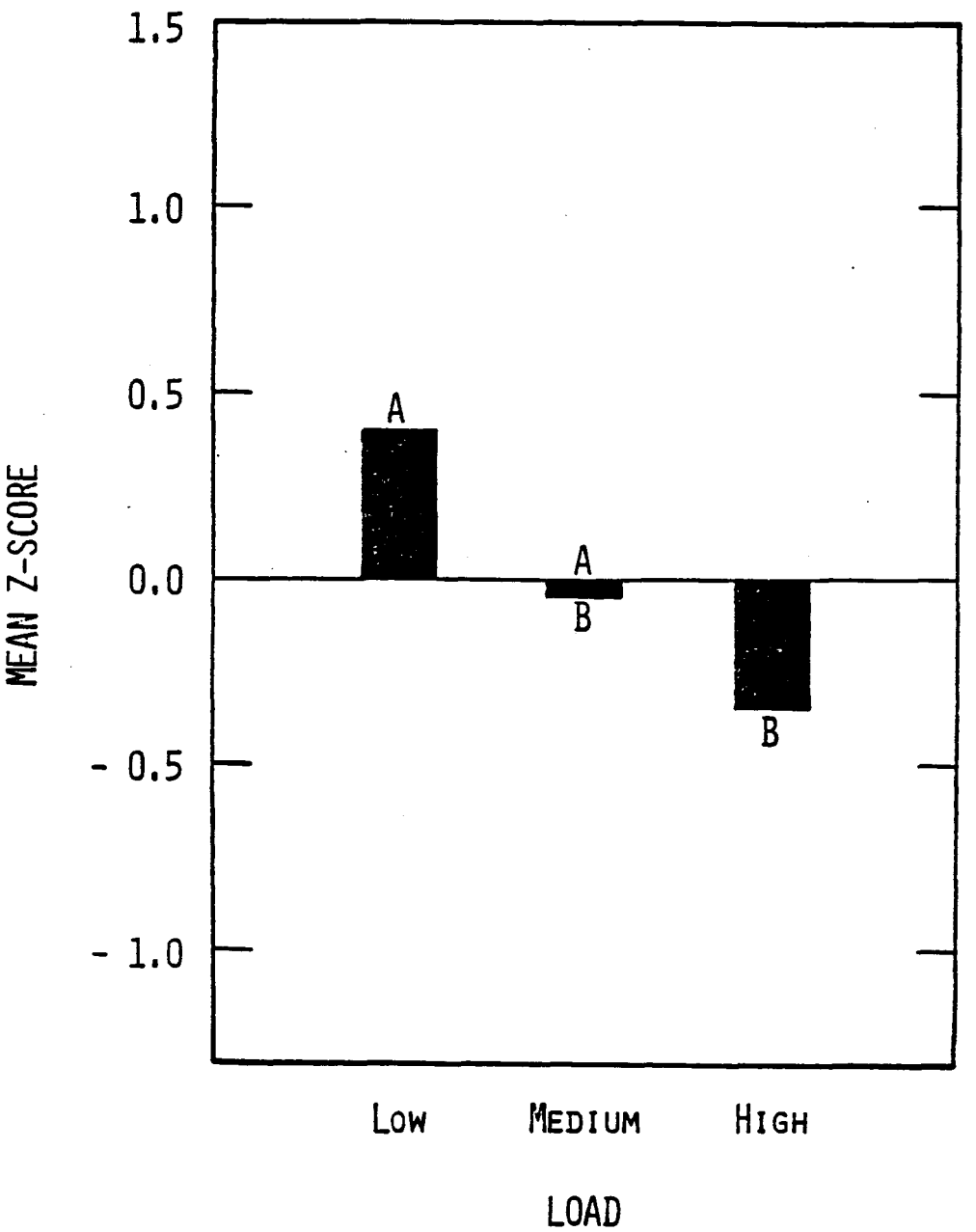

Figure 7. Effect of load on mean standardized scores for the communications response time technique. (Means with different letters are significantly different, $p < 0.05$).

VI. PUBLICATIONS RESULTING FROM THE PROJECT

A. Psychomotor

Connor, S. A. A comparison of pilot workload assessment techniques using a psychomotor task in a moving base aircraft simulator. M. S. Thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, October, 1981.

Connor, S. A. and Wierwille, W. W. Comparative evaluation of twenty pilot workload assessment measures using a psychomotor task in a moving base aircraft simulator. Report submitted to NASA-Ames, to appear as a grant report; January, 1982.

Wierwille, W. W., Determinationof sensitive measures of pilot workload as a function of the type of piloting task. Proceedings of the AIAA Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics, January, 1982.

Wierwille, W. W. and Connor, S. A. The sensitivity of twenty measures of pilot workload in a simulated ILS task. Proceedings of the Eighteenth Annual Conference on Manual Control, Dayton, Ohio, June, 1982.

Connor, S. A. and Wierwille, W. W. Evaluation of twenty workload assessment measures using a psychomotor task in a moving-base aircraft simulator. Human Factors, 25, February, 1983 (to appear).

B. Mediational

Rahimi, M., Evaluation of workload estimation techniques in simulated piloting tasks emphasizing mediational activity. Ph.D. Dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, July, 1982.

Rahimi, M. and Wierwille, W. W., Evaluation of the sensitivity and intrusion of workload estimation techniques in piloting tasks emphasizing mediational activity. Proceedings of the 1982 IEEE International Conference on Cybernetics and Society, Seattle, Washington, October, 1982, 593-597.

C. Perceptual

Casali, J. G., A sensitivity/intrusion comparison of mental workload estimation techniques using a simulated flight task emphasizing perceptual piloting behaviors. Ph.D. Dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, July, 1982.

Casali, J. G. and Wierwille, W. W., A sensitivity/intrusion comparison of mental workload estimation techniques using a flight task emphasizing perceptual piloting activities. Proceedings of the 1982 IEEE International Conference on Cybernetics and Society, Seattle, Washinton, October, 1982, 598-602.

D. Communications

Casali, J. G. and Wierwille, W. W. Communications-imposed pilot workload: a comparison of sixteen estimation techniques. To be presented at the Second Symposium on Aviation Psychology, Ohio State University, Columbus, April, 1983.

Casali, J. G. and Wierwille, W. W. A comparative evaluation of rating scale, secondary task, physiological, and primary task workload estimation techniques in a simulated flight task emphasizing communications load. (Submitted to Human Factors, February, 1983).

E. Other

Wierwille, W. W. Instantaneous mental workload: concept and potential methods for measurement. Proceedings 1981 International Conference on Cybernetics and Society, Atlanta, Georgia: IEEE Systems, Man and Cybernetics Society, October, 1981, 604-608.

Wierwille, W. W. and Casali, J. G. Mental workload-an I.E. problem? Ergonomics Newsletter of the I.I.E., February, 1983.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR-166496		

4. Title and Subtitle	5. Report Date
COMPARATIVE EVALUATION OF WORKLOAD ESTIMATION TECHNIQUES IN PILOTING TASKS	Feb. 1980 - Feb. 1983
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Walter W. Wierwille	
9. Performing Organization Name and Address	10. Work Unit No.
Virginia Polytechnic Institute Dept. of IEOR Blacksburg, Virginia 24061	T5743
	11. Contract or Grant No.
	NAG2-17
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
National Aeronautics and Space Administration Washington, D.C. 20546	Contractor Report
	14. Sponsoring Agency Code
	505-35-21

15. Supplementary Notes

Point of Contact: Sandra Hart, MS: 239-3, Ames Research Center,
Moffett Field, CA 94035
415-965-6072 or FTS 448-6072

16. Abstract

In January, 1980, NASA-Ames Research Center awarded a research grant to Virginia Polytechnic Institute and State University (Virginia Tech). The objective of this research was to examine the sensitivity and intrusion of a wide variety of workload assessment techniques in simulated piloting tasks. The study employed four different piloting tasks emphasizing psychomotor, perceptual, mediational, and communications aspects of piloting behaviors. An instrumented moving base general aviation aircraft simulator was used for the study. This document provides a summary of the research.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Workload, Performance, Measurement, Human Factors, Pilot Workloads and Performance	Unclassified - Unlimited Star Category - 03

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	92	

*For sale by the National Technical Information Service, Springfield, Virginia 22161