Handbook of Human Performance Measures and Crew Requirements for Flightdeck Research

Albert J. Rehmann

December 1995
DOT/FAA/CT-TN95/49

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

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

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.
DOT/FAA/CT-TN95/49

2. Government Accession No.  

3. Recipient's Catalog No.  

4. Title and Subtitle  
HANDBOOK OF HUMAN PERFORMANCE MEASURES AND CREW REQUIREMENTS FOR FLIGHTDECK RESEARCH

5. Report Date  
December 1995

6. Performing Organization Code  
ACT-350

7. Author(s)  
Albert J. Rehmann; CSERIAC

DOT/FAA/CT-TN95/49

9. Performing Organization Name and Address  
Crew System Ergonomics Information Analysis Center (CSERIAC)  
2255 H Street, Building 248  
Wright-Patterson Air Force Base, OH 45433-7022

10. Work Unit No. (TRAIS)  
-  

11. Contract or Grant No.  
DLA900-88-D-0393

12. Sponsoring Agency Name and Address  
U.S. Department of Transportation  
Federal Aviation Administration Technical Center  
Atlantic City International Airport, NJ 08405

13. Type of Report and Period Covered  
Technical Report

-  

15. Supplementary Notes  
-  

16. Abstract  
The Federal Aviation Administration (FAA) Technical Center envisions that their studies will require standard measure of pilot/crew performance. Therefore, the FAA commissioned the Crew System Ergonomics Information Analysis Center (CSERIAC) to (1) identify state-of-the-art pilot/crew performance measures in selected areas of interest, (2) provide guidance material to allow the FAA Technical Center to determine appropriate measures for a given study classification, and (3) provide guidelines on pilot subject characteristics used in their studies. Adhering to accepted standards will allow performance data to be translated between FAA studies and generalized across other government and industry partners.

Three areas of human performance that have achieved the most attention in the literature are: workload, situational awareness, and vigilance. An extensive literature search was conducted on each of these areas and leading experts in the human performance research industry were consulted. The tools and techniques used to measure each of these three areas are investigated. Guidelines are provided to assist the human factors practitioner in choosing the most appropriate performance measure for a given study classification (e.g., part-task, full mission, end-to-end). A set of criteria and guidelines on pilot subject characteristics, such as, number of subjects, experience level required, and the use of different airline flightcrews, is also provided.

17. Key Words  
Reconfigurable Cockpit System (RCS)  
Cockpit Simulation Network (CSN)  
Human Factors

18. Distribution Statement  
This document is available to the public through the National Technical Information Service, Springfield, VA 22161

19. Security Classif.(of this report)  
Unclassified

20. Security Classif.(of this page)  
Unclassified

21. No. of Pages  
94

22. Price  
-  

Form DOT F 1700.7 (8-72)  
Reproduction of completed page authorized
FOREWORD

This report documents work performed by Crew Systems Ergonomics Information Analysis Center (CSERIAC) on subtask 1 out of 4 of the task entitled "Simulation Fidelity Requirements." The task was a provision of an interagency agreement between the Federal Aviation Administration (FAA) Technical Center, Department of Transportation (DOT) and the Defense Technical Information Center (DTIC). It was conducted under DOD Contract Number DLA900-88-D-0393, and the CSERIAC Task Number was 93956-24. The CSERIAC Program Manager was Mr. Don Dreesbach. The CSERIAC Task Leader was Mr. Michael C. Reynolds. The FAA Technical Program Manager (TPM) was Mr. Albert J. Rehmann, and the FAA Project Engineer was Mr. Pocholo Bravo.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXECUTIVE SUMMARY</td>
<td>ix</td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>1.1 Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Statement of Problem/Scope</td>
<td>1</td>
</tr>
<tr>
<td>1.3 Objective</td>
<td>1</td>
</tr>
<tr>
<td>1.4 Purpose</td>
<td>1</td>
</tr>
<tr>
<td>1.5 Method/Procedure</td>
<td>2</td>
</tr>
<tr>
<td>1.6 Organization of Report</td>
<td>3</td>
</tr>
<tr>
<td>2. HUMAN PERFORMANCE MEASURES</td>
<td>3</td>
</tr>
<tr>
<td>2.1 Measurement Criteria</td>
<td>3</td>
</tr>
<tr>
<td>2.2 Workload</td>
<td>6</td>
</tr>
<tr>
<td>2.2.1 Subjective Measures</td>
<td>6</td>
</tr>
<tr>
<td>2.2.2 Performance Measures</td>
<td>8</td>
</tr>
<tr>
<td>2.2.3 Physiological Measures</td>
<td>12</td>
</tr>
<tr>
<td>2.2.4 Guidelines in Workload Assessment</td>
<td>13</td>
</tr>
<tr>
<td>2.3 Situational Awareness</td>
<td>15</td>
</tr>
<tr>
<td>2.3.1 Explicit Measures</td>
<td>16</td>
</tr>
<tr>
<td>2.3.2 Implicit Measures</td>
<td>17</td>
</tr>
<tr>
<td>2.3.3 Subjective Rating Measures</td>
<td>18</td>
</tr>
<tr>
<td>2.4 Vigilance</td>
<td>19</td>
</tr>
<tr>
<td>3. STUDY CLASSIFICATIONS</td>
<td>22</td>
</tr>
<tr>
<td>3.1 Definitions of Study Classifications</td>
<td>22</td>
</tr>
<tr>
<td>3.1.1 Part-Task Environment</td>
<td>23</td>
</tr>
<tr>
<td>3.1.2 Full Mission Simulations</td>
<td>23</td>
</tr>
<tr>
<td>3.1.3 End-to-End Simulations</td>
<td>24</td>
</tr>
<tr>
<td>3.2 Measurement by Study Classification</td>
<td>24</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS (CONTINUED)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>FLIGHTCREW DATA REQUIREMENTS</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>SUMMARY</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>REFERENCES</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>ABBREVIATIONS</td>
<td>37</td>
</tr>
</tbody>
</table>

## APPENDICES

A - DESCRIPTION OF MEASUREMENT TOOLS AND TECHNIQUES
LIST OF ILLUSTRATIONS

Figure

1 Increased Cost, Time and Risk as a Level of Experiment Fidelity 22

LIST OF TABLES

Table

1 Summary of Workload Assessment Technique Capabilities 7
2 Six Common Subjective Rating Scales Judged on Several Criteria 9
3 Example Primary Task Measures 11
4 Secondary Task Methodologies 12
5 Physiological Measures of Workload 12
6 Common Physiological Measures Judged on Several Criteria 13
7 Guidelines for the Use of Subjective Workload Measures 14
8 Guidelines for the Use of Primary Task Workload Measures 14
9 Guidelines for the Use of Secondary Task Workload Measures 15
10 Guidelines for the Use of Physiological Workload Measures 15
11 Situational Awareness Measurement Methods 16
12 Outcomes of a Signal Detection Experiment 17
13 Matrix of Human Performance Measures by Study Classification 25
14 Some Questions to Consider in Selecting an Appropriate Workload Measure 26
EXECUTIVE SUMMARY

A concern in modern aircraft is that flightcrews are inundated with an enormous amount of automation. This has changed the role of the flightcrew and has demanded increased monitoring behaviors than ever before. Because flightcrew behavior is less observable, the challenge in the human factors research industry is to identify pilot performance through new evaluation tools and techniques.

The Federal Aviation Administration (FAA) Technical Center envisions that their studies will require standard measures of pilot/crew performance. Therefore, the FAA commissioned the Crew System Ergonomics Information Analysis Center (CSERIAC) to (1) identify state-of-the-art pilot/crew performance measures in selected areas of interest, (2) provide guidance material to allow the FAA Technical Center to determine appropriate measures for a given study classification, and (3) provide guidelines on pilot subject characteristics used in their studies. Adhering to accepted standards will allow performance data to be translated between FAA studies and generalized across other government and industry partners. This document describes work performed by CSERIAC on subtask 1 out of 4 of the task entitled “Simulation Fidelity Requirements.”

Three areas of human performance that have achieved the most attention in the literature are: workload, situational awareness, and vigilance. An extensive literature search was conducted on each of these areas and leading experts in the human performance research industry were consulted. The assortment of information was reviewed, compiled, and integrated into a convenient handbook applicable to human factors personnel within the FAA.

The FAA has currently in place a variety of testbeds, including the Reconfigurable Cockpit System (RCS) and the Cockpit Simulation Network (CSN). The handbook defines various systems engineering study classifications (e.g., part-task, full-mission, end-to-end) and provides guidelines in the selection of appropriate tools and techniques within each study classification. The use of expert system, knowledge-based tools for matching performance measures to various study classes is also addressed.

A set of criteria and guidelines on pilot subject characteristics is also provided. Qualities of a pilot subject are often discarded in many human performance research studies, and as a result data obtained from these studies may not be representative, nor reflect performance in the real world. Guidelines on number of subjects, experience level required, the use of different airline flightcrews, etc., are provided.

In conclusion, CSERIAC has acquired a network of experts in the area of workload, situational awareness, and vigilance, including related human performance disciplines, such as adaptive
automation, boredom, fatigue, and team decision making. Future efforts may require additional contacts with these experts. As a result of this effort, a database of pilot subject selection criteria has been established which will contain pilot subject characteristics from past and future FAA evaluations. Additionally, any interesting facts or observations will also be recorded.
1. INTRODUCTION.

1.1 BACKGROUND.

The role of the flightcrew in today's modern aircraft has changed considerably since the introduction of the electronic cockpit. Pilots are acting as supervisors and managers of systems, rather than performing traditional manual roles. These changes have placed additional cognitive processing demands on the crew. For the human factors practitioner, evaluation of pilot performance is inherently more difficult to measure and less observable.

1.2 STATEMENT OF PROBLEM/SCOPE.

The emphasis on this paper is to provide guidance material on a few of the most common measurable constructs in the research industry; that is, mental workload, situational awareness, and vigilance. Together, these three areas have received the most attention in the research industry. Workload, for example, has a wide variety of performance metrics that are available to the human factors practitioner. However, because of many definitions and beliefs, it is not always understood what workload measures are more appropriate to use than others. The decisions to be made are numerous and depend on a variety of factors. Some factors are not as easily understood. Therefore, references to expert tools specifically designed to assist the practitioner will be identified.

Many tools are available to measure mental workload, situational awareness, etc., however, only those which are used most common in the research industry, specifically, in the area of flight simulation will be identified. Furthermore, the intent of this report is only to provide a review of the various tools and not to identify how to use them. Appendix A lists the most common tools, along, with references identifying their use.

1.3 OBJECTIVE.

The objective of this research was threefold (1) to identify state-of-the art pilot performance measures, (2) to suggest how these measurement methods and metrics should be employed for use in the Federal Aviation Administration (FAA) Technical Center's pilot/crew performance studies, and (3) to determine a set of criteria or guidelines to guide the human factors practitioner in selecting pilot subjects for evaluation. The ultimate goal is to provide a handbook which can be used by human factors personnel within the FAA.

1.4 PURPOSE.

The purpose of this handbook is to provide guidelines on various human performance measures and flightcrew selection criteria for future human factors Data Link research conducted at the FAA Technical Center. The FAA Technical Center has a unique
assortment of evaluation equipment for conducting pilot performance studies, including the Reconfigurable Cockpit System (RCS). The RCS can be rapidly changed to reflect various cockpit configurations and can be remotely tied into, along with other simulators across the world, the FAA Technical Center's Cockpit Simulation Network (CSN).

This report is one of four reports that have been written by Crew System Ergonomics Information Analysis Center (CSERIAC) under the auspices of the "Simulation Fidelity Requirements" task for the FAA Technical Center. The other reports (Mitman et al., 1994a, 1994b; Reynolds, 1994) contain issues, such as simulator sophistication required, number of simulators required, and generalization of performance data that should also be consulted.

1.5 METHOD/PRECEDURE.

Two approaches to identifying state-of-the-art performance measures were followed. The first approach was to identify general measures of human performance through searches of local libraries and the Defense Technical Information Center (DTIC), National Technical Information Service (NTIS), Aerospace and Compendex on-line databases. As expected, this yielded many references, therefore, further refined searches, limited to the last 10 years and containing pilot/crew performance measures were performed. Results of these searches were filtered and separated into the three specific topic areas of interest: mental workload, situational awareness, and vigilance. Special attention was made to those references which provided recent reviews as opposed to describing specific evaluations which employed or described only a few measures.

The second approach involved soliciting the human factors research industry and universities for subject matter experts (SME) in the various fields of interest. Local SMEs were contacted at Wright-Patterson Air Force Base (WPAFB) within Armstrong Laboratory's Human Engineering Lab (AL/CFH), Wright State University, University of Dayton, and University of Cincinnati. Contacts were made and information was obtained through both phone conversation and personal visits.

On a national level, contacts were made with the National Technology Transfer Center (NTTC). The NTTC, which is sponsored by National Aeronautics and Space Administration (NASA) in cooperation with other federal agencies, serves as a national clearinghouse/gateway for federal technology transfer. An NTTC agent works cooperatively to help refine an individual's request and identifies key contacts in the specified areas of interest. Based on the NTTC support, additional contacts were made with personnel from NASA-AMES, Office of Research and Technology Assessment (ORTA, Brooks AFB), Crew Station Technology Lab (Brooks AFB), and Army Aeroflight Directorate. The personal contacts were made to inquire about recent developments in the area of pilot performance measures not yet published nor in the
public mainstream. As a result of this effort, a network of SMEs was established in all three of the main topic areas, including notable researchers in the area of complacency, boredom, team-decision making and fatigue. A database containing, various SMEs and related government and industry research labs was compiled as a result of this effort. For future consideration, the database will be drawn upon to gather additional information on the various research topics.

1.6 ORGANIZATION OF REPORT.

The report is organized into three major sections. The first section (section 2. Human Performance Measures) addresses the various state-of-the-art performance measures for the three areas previously mentioned.

The second section (section 3. Study Classifications) describes common testbeds or study classifications envisioned to be used by the FAA Technical Center. This section further identifies what performance measures are appropriate for a given study classification. Although other sections may contain additional information on performance measure classification, this section is intended to be the sole section on this subject area.

The third section (section 4. Flightcrew Data Requirements) identifies pilot subject selection criteria that should be considered when conducting human factors research in the different study classifications.

In addition, a reasonably comprehensive reference/bibliography list (over 190 total citations) is provided at the end of the report and in appendix A. Appendix A describes various workload and situational awareness measurement techniques.

2. HUMAN PERFORMANCE MEASURES.

The assessment techniques/tools used in test and evaluation for measuring workload, situational awareness, and vigilance are described in the following section. Again, this is not a complete list, only those to be considered the most common or have been proven to be valid and reliable within a flight simulation environment are addressed. Publications documenting recent reviews of assessment techniques have been highlighted within each area. Emphasis was on providing guidelines, through depiction of various tables, on what methods to use or not to use. Many factors should be considered, not the least of which is cost and the amount of time, resources, and effort in using the tools.

2.1 MEASUREMENT CRITERIA.

A common consideration before conducting a study is to consider several criteria with regards to the measure in question. For example, some measures may be more reliable than others; that is,
the measured output is consistent over repeated test conditions. The following list delineates nine of the most notable criteria. No single measure will have all these attributes, therefore, the practitioner should consider using multiple measures. Most of these criteria were extracted from the American National Standards Institute (ANSI)/American Institute of Aeronautics and Astronautics (AIAA) "Guide to Human Performance Measurements."

Appropriate Level of Detail - Measures should reflect the performance of interest with sufficient detail to permit a meaningful analysis. For example, if one is evaluating alternative control and display relationships, the performance of each step and control activation in a procedural sequence could be important to understand the best configuration or potential for errors. On the other hand, collecting such detailed information might not be appropriate when comparing the effectiveness of two competing systems that had dissimilar procedures. In this case, one should focus on measures of effectiveness (e.g., how well did the operator/maintainer and machine perform the intended purpose of the system) (ANSI/AIAA, 1992).

Reliability - Reliability is the repeatability of a measure. If one measures the same behavior in exactly the same way under identical circumstances, the same value of the metric should result. In human performance measurement, however, individual differences among human operators, decision makers, and maintainers occur; even the same person may respond to successive trials differently because of learning or other effects. To adjust for this, the concept of reliability is extended from a value of a metric to a distribution of a metric; thus, if one obtains the same distribution with repeated measures, the metric is said to be reliable (ANSI/AIAA, 1992).

Validity - Does the measure mean what it is supposed to mean; is it appropriate to use for the intended purpose? There are at least five types of validity: face, concurrent, content, construct, and predictive. Each of these is described separately (ANSI/AIAA, 1992).

Face Validity - Face validity is the most common; here, a subject matter expert usually confirms that the particular metric represents performance that is important for accomplishment of the task.

Concurrent Validity - Concurrent validity is the correlation of a measure with other measures. If two measures correlate highly with each other, they may be measuring the same thing. The higher the correlation, the greater degree of similarity.

Content Validity - Content validity addresses comprehensiveness - proper sampling of the performance in a battery of test items and measures. Have you sampled all of the important areas of
performance or knowledge? Do you have test items or measures that are unimportant, or perhaps irrelevant, to the task?

Construct Validity - Construct validity is concerned with the correlation of a measure (or group of measures) with a construct, theory, or model. One may hypothesize that responses (measures) to a written test battery (the measurement instrument) will be different for various professional groups, such as engineers, physicians, and pilots. By offering the test battery to the various groups, one can classify the responses by group; if the responses are different, the validity of the construct would be demonstrated. Similarly, in the performance domain, one may hypothesize and validate the construct that expert operators perform in a different way than novices.

Predictive Validity - Predictive validity is perhaps the most important characteristic of behavioral measures, yet it often is the most neglected and difficult to obtain. Here, one would like to know that measures being taken in a laboratory, on a mockup, in a simulator, or during training are representative and predictive of the performance of the human being (and that system) in the real world on the job.

Sensitivity - Does the measure react sufficiently well to changes in the independent variable. It is quite possible that the measure chosen may be valid and reliable, but will not show a large enough effect to be measured easily (Wilson, Corlett, 1990).

Transferability - Transferability refers to the capability of a technique to be used in various applications. Some techniques vary from application to application. Consequently, a measure (e.g., flight control inputs) that is applicable to one type (e.g., aircraft system) of evaluation might not be readily transferred to another type of evaluation (e.g., process control system) (Wilson, Corlett, 1990).

Diagnosticity - Diagnosticity can be thought of as the characteristic of a measure to provide information that will tend to isolate the cause of good or bad performance. Measures that have diagnosticity add value to the measure set by providing information that might not be obtained in any other way (ANSI/AIAA, 1992).

Intrusiveness - A measure requiring a technique that in the process of data collection attracts the attention of the subject may clearly affect the subject’s task performance. If it does so, the measure is intrusive. Almost all measures are intrusive to some extent, but their contaminating effect on task performance will vary. Less obtrusive methods of data collection are to be preferred to more intrusive ones (ANSI/AIAA, 1992).

Implementation Requirements - When designing measures or selecting methods of measuring, one must consider the
implementation requirements of the measure set. The general issues to be considered include ease of data collection, robustness of the measurement instruments, and overall data quality control. These issues apply to design and maintenance of laboratory measurement systems and instruments as well as to simulator studies and field exercises (ANSI/AIAA, 1992).

Flexibility - Measurement instruments and automated performance measurement systems should be designed in a manner that will enhance the ability to make changes in measures as situations demand. For automated, computer-driven performance measurement systems, this means placing measurement specifications in tables (or other such mechanisms) so that changes can be made without having to recode or recompile the computer program (ANSI/AIAA, 1992).

2.2 WORKLOAD.

A widely accepted taxonomy of workload assessment techniques is that they generally fall into four major categories: subjective, physiological, performance based, and analytical. The first three categories comprise what is known as empirical techniques, or methods which are used during "operator-in-the-loop" evaluations. The last category, analytical, are predictive techniques that are normally employed in early stages of system design that do not require a pilot subject. From the standpoint that FAA evaluations will be "operator-in-the-loop" simulations, only empirical methods will be described. Table 1, adopted from Eggemeier (1987), provides a general description of how each of the empirical techniques (subjective, physiological, performance [primary and secondary]) relate to five of the nine measurement criteria established in the previous section.

The next three sections identify the various empirical workload assessment techniques. A last section summarizes some advantages and disadvantages of the various techniques. For further information, some excellent recent reviews (Christ et al., 1993; Veltman and Gaillard, 1993; Wierwille and Eggemeier, 1993; Grenell et al., 1991; Eggemeier et al., 1990; Corwin et al., 1989; Lysaght et al., 1989) on workload assessment techniques in general should be consulted.

2.2.1 Subjective Measures.

Perhaps the most popular form of workload assessment employed in the field is in the use of subjective rating scales. The review of the literature uncovered a wide variety of techniques, each having their advantages and disadvantages. The common thread among all rating scales is that they are fairly easy to implement, low cost, and are relatively free of intrusion. Appendix A describes 19 such scales. A general description, strengths and limitations, graphs, etc., are provided for each scale. Although more scales exist, only those chosen are
<table>
<thead>
<tr>
<th></th>
<th>SENSITIVITY</th>
<th>DIAGNOSTICITY</th>
<th>INTRUSIVENESS</th>
<th>IMPLEMENTATION REQUIREMENTS</th>
<th>OPERATOR ACCEPTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SUBJECTIVE TECHNIQUES</strong></td>
<td>Capable of discriminating levels of capacity expenditure in nonoverload situations. Can be used to assess the relative potential for overload among design options.</td>
<td>Not considered diagnostic. Available evidence indicates that rating scales represent a global measure of load. Lack of diagnosticity suggests use as a general screening device to determine if overload exists anywhere within task performance.</td>
<td>Intrusion does not appear to represent a significant problem. Most applications require rating scale completion subsequent to task performance and, therefore, present no intrusion problem.</td>
<td>Instrumentation required is usually minimal, permitting use in a number of environments. Traditional applications require mockups, simulators, or operational equipment. Imposes limits on use during early system development. Recent projective use provides potential for application during early stages. Some familiarization with procedures can be required.</td>
<td>No systematic data. Informal evidence suggests that several rating scales enjoy a high degree of operator acceptance.</td>
</tr>
<tr>
<td><strong>PRIMARY TASK MEASURE</strong></td>
<td>Discriminate overload from nonoverload situations. Used to determine if operator performance will be acceptable with a particular design option.</td>
<td>Not considered diagnostic. Represents a global measure of workload that is sensitive to overloads anywhere within the operator's processing system.</td>
<td>Nonintrusive since no additional operator performance or support required.</td>
<td>Instrumentation for data collection can restrict use in operational environments. Use requires mockups, simulators, or operational equipment. Imposes limits on use during early system development. No operator training required.</td>
<td>No systematic data. No reason to expect negative operator opinion.</td>
</tr>
<tr>
<td><strong>SECONDARY TASK METHODS</strong></td>
<td>Capable of discriminating levels of capacity expenditure in nonoverload situations. Used to assess reserve capacity afforded by a primary task. Can be used to assess the potential for overload among design options.</td>
<td>Capable of discriminating some differences in resource expenditure (e.g., central processing versus motor). Diagnosticity suggests complementary use with more generally sensitive measures, with the latter initially identifying overloads and secondary tasks being used subsequently to pinpoint the locus of overload.</td>
<td>Primary task intrusion has represented a problem in many applications, particularly in the laboratory. Data are not extensive in operational environments. Several techniques (e.g., embedded secondary task, adaptive procedures) have been designed to control intrusion. Potential for intrusion could limit use in operational environments.</td>
<td>Instrumentation for data collection can restrict use in operational environments, but some tasks have been instrumented for in-flight use. Use requires mockups, simulators, or operational equipment. Imposes limits on use during early system development. Some operator training usually required to stabilize secondary task performance.</td>
<td>No systematic data. Requirement to perform secondary task could distract operator. Technique such as embedded secondary task should minimize any acceptance problems.</td>
</tr>
<tr>
<td><strong>PHYSIOLOGICAL TECHNIQUES</strong></td>
<td>Capable of discriminating levels of capacity expenditure in nonoverload situations. Can be used to assess the relative potential for overload among design options.</td>
<td>Some techniques (e.g., event-related brain potential) appear diagnostic of some resources, while other measures (e.g., pupil diameter) appear more generally sensitive. Choice of technique dependent on purpose of measurement (screening for any overload versus identifying locus of overload.</td>
<td>Intrusion does not appear to represent a major problem, although there are data to indicate that some interference can occur.</td>
<td>Instrumentation for data collection can restrict use in operational environments. Use requires mockups, simulators, or operational equipment. Imposes limits on use during early system development. No operator training required.</td>
<td>No systematic data. Instrumentation and recording equipment could represent potential problem, but no significant problems reported in literature.</td>
</tr>
</tbody>
</table>
considered to be the most applicable to the flight simulation environment.

One factor in the selection of a subjective workload rating scale is its diagnosticity. Based on the multiple-resource theory (Wickens, 1980), workload can be thought of as a multidimensional construct. Diagnostic measures such as the Subjective Workload Assessment Technique (SWAT) and the NASA Task Load Index (NASA-TLX), for example, are multidimensional rating scales. They provide information on the various components or sources of workload, as well as an estimate of global workload. Sandry-Garza et al., (1987) concluded that SWAT and TLX are excellent workload tools for commercial transport aircraft applications. Corwin et al., 1989) concluded that in addition to SWAT and TLX, the Bedford Workload and the Pilot Subjective Evaluation (PSE) scales are also applicable for in-flight simulation evaluations.

Two scales that have been recently published are the Dutch Effort Scale (Veltman and Gaillard, 1993) and the Subjective Workload Profile Scale (Tsang and Velazquez, 1993); both are multidimensional scales. The Dutch Effort Scale has been validated in a fixed based simulator environment and has proven to be more sensitive to differences in task loading as compared to the NASA-TLX scale. The Subjective Workload Profile is currently undergoing validation exercises; it has yet to be employed within a flight simulation environment.

Table 2 contains a description of the sensitivity, reliability, diagnosticity, etc., for six of the most common subjective rating scales. Each of these scales are further described in detail in appendix A.

2.2.2 Performance Measures.

Performance based measures utilize some aspect of the operator's capability to perform tasks or system functions in order to provide an assessment of workload. Performance measures can be further broken down into two subcategories: primary task measures and secondary task measures. Each of these areas will be individually described in separate sections.

2.2.2.1 Primary Task Measures.

Primary task measures assess some aspect of the operator's capability to perform the task or system function of interest (Eggemeier and Wilson, 1991). As flight crew task demands increase, their ability to perform at an optimal level decreases. Primary task measures exhibit their greatest sensitivity to variations in workload when the total task demand in a situation exceeds the pilot's capability to process information (Eggemeier et al., 1990). The reason is due to the pilot's ability to overcome moderate levels of workload while still being able to perform at an optimal level on the primary task measure.
Secondary tasks, described in the next section, are typically used in conjunction with primary task measures in order to determine lower to moderate workload shifts in task performance.

**TABLE 2. SIX COMMON SUBJECTIVE RATING SCALES JUDGED ON SEVERAL CRITERIA**

<table>
<thead>
<tr>
<th>Technique</th>
<th>Sensitivity</th>
<th>Reliability</th>
<th>Diagnosticity</th>
<th>Cost/Effort Requirements</th>
<th>Task Time</th>
<th>Ease of Scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytical Hierarchy Process</td>
<td>High</td>
<td>High</td>
<td>Moderate</td>
<td>Low Cost</td>
<td>Low Effort</td>
<td>Requires rating pairs of tasks</td>
</tr>
<tr>
<td>Bedford</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Low Cost</td>
<td>Low Effort</td>
<td>Requires two decisions</td>
</tr>
<tr>
<td>Cooper-Harper</td>
<td>High for psychomotor</td>
<td>High</td>
<td>Low</td>
<td>Low Cost</td>
<td>Low Effort</td>
<td>Requires three decisions</td>
</tr>
<tr>
<td>Modified Cooper-Harper</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Low Cost</td>
<td>Low Effort</td>
<td>Requires three decisions</td>
</tr>
<tr>
<td>NASA-TLX</td>
<td>High</td>
<td>High</td>
<td>Moderate/High</td>
<td>Low Cost</td>
<td>Low Effort</td>
<td>Requires six ratings</td>
</tr>
<tr>
<td>SWAT</td>
<td>High</td>
<td>High</td>
<td>Moderate/High</td>
<td>Low Cost</td>
<td>Low Effort</td>
<td>Requires prior card sort and three ratings</td>
</tr>
</tbody>
</table>

(Source: Lysaght et al., 1989; ANSI/AIAA, 1992)

Typical primary task measures involve the recording of flightcrew control input activity, whether it be from the wheel, column, or pedal. Flight simulation evaluations, for example, that assess the differences between two landing type displays, would be interested in lateral, localizer, and glide slope deviations of final approach. In general, control input activity has demonstrated evidence of validity, reliability, and applicability as primary task measures for evaluating pilot workload (Corwin et al., 1989).

A problem with primary task measures, as opposed to the more general class of secondary class measures, is that a measure must be developed on an individual basis for each application. Care must be taken in selecting an appropriate measure. For example, control input activity might be an acceptable measure when comparing the benefits of Data Link communications to that of voice if only one crew member is aboard. But the same measure in a two-person crew environment would not be applicable. Typically, in the air transport environment, the nonflying pilot handles all radio communications, therefore, control input activity may not be diagnostic of his/her workload.

Traditional primary task measures are speed and accuracy type measures. Speed (or time) would measure the reaction time, for example, to perceive an event, initiate a movement or correction,
and/or perhaps detect a trend of multiple related events occurring in the cockpit. Accuracy measures in a Data Link simulated environment, for example, would be used to measure the accuracy or detection that a signal (Data Link aural/visual alert) was recognized, the appropriate response (WILCO, UNABLE, etc.) was made and that the Data Linked information was accurately conveyed to all crew members. The reciprocal measure of accuracy--errors--would also be an acceptable primary task measure. One could measure the number of errors operating a Data Link display (e.g., incorrect switch hits) in a terminal environment as opposed to an en route environment.

LeMay and Comstock (1990) proposed an overall indicator of performance, or Figure Of Merit (FOM) to establish the effect of workload on efficiency to identify overload conditions. They tested the FOM procedure on simulated landing tasks in which standard communications (voice) with air traffic control (ATC) was compared with a Data Link system. Combined scores for continuous, e.g., lateral position and altitude data, and discrete task performance, e.g., time spent on autopilot manipulation, were added together to obtain an overall FOM. The results indicate that Data Link communications significantly increased variability in overall task performance. The authors stated that the normative FOM technique may be useful to discover problems associated with new technology introduction; they recommended more simulation studies to validate the sensitivity of the procedure.

Control input activity, speed (time) and accuracy (errors) are the most often uses of primary task measures for assessing workload variations. The references provided earlier and those listed in this section provide additional information on primary task measures.

Other primary task measures listed in table 3 have also been used. The sensitivity of these measures is dependent on the level of workload experienced by the operator. If the pilot’s load is too low, then the measures are not sensitive to workload variations in this region. On the other hand, if the pilot’s load is excessive, then these measures by themselves would be highly sensitive (Lysaght et al., 1989). Assuming that the simulation testbed is capable of recording system performance measures, the measures depicted in table 3 can be obtained at a relatively low cost with moderate effort.

2.2.2.2 Secondary Task Measures.

The literature yielded many different kinds of secondary task methodologies (ANSI/AIAA, 1992; Eggemeier and Wilson, 1991; Lysaght et al., 1989). A secondary task is measured in conjunction with a primary task measure. The relative workload associated with the primary task is reflected in the level of performance on the secondary task. For example, if the flightcrew’s workload is fully loaded on the primary task,
performance on a secondary task may be unacceptable. Secondary task paradigms are also used to provide diagnostic information regarding the type of resources (motor, perceptual, etc.) available (spare capacity) or expended by the operator. Because of this, secondary task methodologies are considered sensitive to detecting operator workload, especially during expected low or moderate workload conditions.

**TABLE 3. EXAMPLE PRIMARY TASK MEASURES**

<table>
<thead>
<tr>
<th>Example Measures</th>
<th>Sensitivity</th>
<th>Cost/Effort Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airspeed Deviation</td>
<td>Low/High</td>
<td>Low Cost Moderate Effort</td>
</tr>
<tr>
<td>Altitude Deviation</td>
<td>Low/High</td>
<td>Low Cost Moderate Effort</td>
</tr>
<tr>
<td>Bank Angle Deviation</td>
<td>Low/High</td>
<td>Low Cost Moderate Effort</td>
</tr>
<tr>
<td>Control Reversals</td>
<td>Low/High</td>
<td>Low Cost Moderate Effort</td>
</tr>
<tr>
<td>Lateral Deviation</td>
<td>Low/High</td>
<td>Low Cost Moderate Effort</td>
</tr>
<tr>
<td>Pitch Rate</td>
<td>Low/High</td>
<td>Low Cost Moderate Effort</td>
</tr>
<tr>
<td>Roll Rate</td>
<td>Low/High</td>
<td>Low Cost Moderate Effort</td>
</tr>
<tr>
<td>Yaw Rate</td>
<td>Low/High</td>
<td>Low Cost Moderate Effort</td>
</tr>
</tbody>
</table>

(Source: Lysaght et al., 1989)

A problem with secondary tasks is that they are often criticized for their intrusive nature. For example, the Interval Production Task (IPT), requires the operator to generate a series of regular time intervals by executing a motor response (e.g., fingertapping). The irregularities in the tapping rate show workload levels, measured by the performance on the primary task, are increasing. The artificiality of some of the secondary tasks, such as the "fingertapping" task, does not bode well in a flight simulation environment. In operational settings, they may interfere with primary tasks to the point of compromising flight safety. Therefore, "embedded" secondary task methodology was developed to overcome these shortcomings (Shingledecker and Crabtree, 1982). For example, existing system hardware on the flightdeck, such as the radio control panel, or cockpit alerting system can be used to generate tasks normal to everyday flight operations.

Many different tools exist, but those identified in table 4 (including the "embedded" secondary task) have appeared most often in the literature and have been proven successful in flight simulation environments. The references provided earlier provide additional information on their use.
TABLE 4. SECONDARY TASK METHODOLOGIES

<table>
<thead>
<tr>
<th>Secondary Task Measures</th>
<th>Sensitivity</th>
<th>Cost/Effort Requirements</th>
<th>Diagnosticity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choice-Reaction Time</td>
<td>Moderate</td>
<td>Moderate Cost Low Effort</td>
<td>Moderate</td>
</tr>
<tr>
<td>Embedded Secondary Task</td>
<td>High</td>
<td>Low Cost Low Effort</td>
<td>Moderate/High</td>
</tr>
<tr>
<td>Mental Mathematics</td>
<td>Moderate</td>
<td>Moderate Cost Low Effort</td>
<td>Moderate</td>
</tr>
<tr>
<td>Sternberg Memory Task</td>
<td>Moderate</td>
<td>Moderate Cost Low Effort</td>
<td>Moderate</td>
</tr>
<tr>
<td>Time Estimation</td>
<td>Moderate</td>
<td>Moderate Cost Low Effort</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

(Source: Lysaght et al., 1989)

2.2.3 Physiological Measures.

Some recent reviews of physiological measures (Lysaght et al., 1989; Kramer, 1991; Gevins and Leong, 1992) uncovered a variety of different tools (table 5) that can be used to assess variations in workload. The various tools can be classified into three major categories, or physiological subsystems (1) eye related measures; (2) brain related measures; and (3) heart related measures. Other measures, such as skin and muscle activity have also been used. Table 6 provides the sensitivity, diagnosticity, etc., of some of the most commonly used physiological measures.

TABLE 5. PHYSIOLOGICAL MEASURES OF WORKLOAD

<table>
<thead>
<tr>
<th>Eye Related Measures</th>
<th>Heart Related Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blink Duration</td>
<td>Electrocardiogram (EKG, ECG)</td>
</tr>
<tr>
<td>Blink Latency</td>
<td>Heart Rate</td>
</tr>
<tr>
<td>Blink Rate</td>
<td>Heart Rate Variability (HRV)</td>
</tr>
<tr>
<td>Endogenous Eyeblinks</td>
<td></td>
</tr>
<tr>
<td>Eye Movement Analysis</td>
<td></td>
</tr>
<tr>
<td>Pupil Diameter</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Brain Related Measure</th>
<th>Other Common Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electroencephalographic Activity (EEG)</td>
<td>Blood Pressure</td>
</tr>
<tr>
<td>Event Related Potentials (ERP), or</td>
<td>Blood Volume</td>
</tr>
<tr>
<td>Evoked Cortical Potentials (CEP)</td>
<td>Body Fluid Analysis</td>
</tr>
<tr>
<td>Magnetoencephalographic Activity (MEG)</td>
<td>Critical Flicker Frequency (CFF)</td>
</tr>
<tr>
<td>Positron Emission Tomography (PET)</td>
<td>Electrodermal Activity (EDA)</td>
</tr>
<tr>
<td>Regional Cerebral Blood Flow (rCBF)</td>
<td>Electromyographic Activity (EMG)</td>
</tr>
<tr>
<td></td>
<td>Galvanic Skin Response</td>
</tr>
<tr>
<td></td>
<td>Muscle Potential</td>
</tr>
<tr>
<td></td>
<td>Respiration</td>
</tr>
<tr>
<td></td>
<td>Skin Potential</td>
</tr>
<tr>
<td></td>
<td>Speech Quality</td>
</tr>
</tbody>
</table>
A major drawback in the use of physiological measures is that they are fairly expensive to implement and require, in some cases, an expert physiologist for implementation and analysis of the data. However, recent advances in physiological measurement technology have afforded the common practitioner a means in which to use these various tools. Nonetheless, the various publications cited earlier provide an explanation of each of the tools.

**TABLE 6. COMMON PHYSIOLOGICAL MEASURES JUDGED ON SEVERAL CRITERIA**

<table>
<thead>
<tr>
<th>Physiological Techniques</th>
<th>Sensitivity</th>
<th>Cost/Effort Requirements</th>
<th>Diagnosticity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blink Rate</td>
<td>Low</td>
<td>Moderate Cost</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moderate Effort</td>
<td></td>
</tr>
<tr>
<td>Body Fluid Analysis</td>
<td>Low</td>
<td>Low Cost</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low Effort</td>
<td></td>
</tr>
<tr>
<td>Evoked Potentials</td>
<td>Moderate</td>
<td>High Cost</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High Effort</td>
<td></td>
</tr>
<tr>
<td>Eye Movement Analysis/</td>
<td>High</td>
<td>High Cost</td>
<td>High</td>
</tr>
<tr>
<td>Scanning Behavior</td>
<td></td>
<td>High Effort</td>
<td></td>
</tr>
<tr>
<td>Heart Rate</td>
<td>Moderate</td>
<td>Moderate Cost</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moderate Effort</td>
<td></td>
</tr>
<tr>
<td>Heart Rate Variability</td>
<td>Moderate</td>
<td>Moderate Cost</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moderate Effort</td>
<td></td>
</tr>
<tr>
<td>Pupil Diameter/Measures</td>
<td>Moderate</td>
<td>High Cost</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moderate Effort</td>
<td></td>
</tr>
</tbody>
</table>

(Source: Lysaght et al., 1989)

Generally, success has been greatest with the use of eye and heart measures in the flight simulation environment than with brain electrical measures (Corwin et al., 1988; Vikmanis, 1989). Brain electrical measures offer a detailed analysis of operator workload and work best in well controlled laboratory settings.

2.2.4 GUIDELINES IN WORKLOAD ASSESSMENT.

This section summarizes in a convenient format guidelines which should be followed when utilizing the various workload tools. The following tables (tables 7 through 10) are separated by workload category and are comprised of information obtained from Corwin et al. (1989) and Grenell et al. (1991).
### TABLE 7. GUIDELINES FOR THE USE OF SUBJECTIVE WORKLOAD MEASURES

<table>
<thead>
<tr>
<th>SUBJECTIVE MEASURES</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Pilots used for the workload assessment should be diversions, etc.) during the evaluation flights.</td>
<td>High face validity because they tap the subjective experiences of the operator; low cost and ease of implementation, lack of implementation, lack of additional equipment and/or extensive computer programming requirements, lack of intrusion on the primary task, high level of operator acceptance.</td>
<td>High level of intersubject variability. May dissociate (report contrasting results) with performance measures of workload. Therefore, suggest that neither subjective nor performance measures be used as the sole basis for assessing operator workload.</td>
</tr>
<tr>
<td>• When using a subjective measure in-flight, the measure should not be intrusive to the flight related tasks the crew member is trying to accomplish.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• If paper and pencil ratings techniques are to be used inflight, one crew member at a time should record their workload ratings so that the other crew member may attend to flightdeck duties.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• The to-be-rated flight segment (beginning and end points) should be clearly identified to the flight crew for the purpose of obtaining the data for evaluation.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• When used, postflight subjective ratings should be collected from the pilots as soon after the task is operationally feasible.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• To enhance postflight workload evaluation, videotape should be used to aid the crew in recalling their subjective evaluations of crew workload.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 8. GUIDELINES FOR THE USE OF PRIMARY TASK WORKLOAD MEASURES

<table>
<thead>
<tr>
<th>PRIMARY TASK MEASURES</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Control Input Activity should be evaluated only during manual flight path control.</td>
<td>Provides a direction indication of total system or subsystem performance while accounting for the operator in the loop.</td>
<td>Very task and situation dependent and therefore are not readily transferred across different tasks or scenarios. As a result, primary task measures must be carefully selected for each application. Primary task measures do not necessarily provide an indication of an operator's spare or &quot;residual&quot; capacity. For example, while two individuals may exhibit equivalent performance, one may be incapable of meeting additional task demands, while the other may possess the spare resources necessary to meet increased task demands or to perform additional tasks.</td>
</tr>
<tr>
<td>• When possible, state variables (e.g., pitch angle, roll angle, altitude) should be recorded continuously in simulation tests.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• When possible, wheel (ailerons) and stick (elevator) inputs should be employed to represent aircraft control workload throughout the entire flight of an aircraft under manual flightpath control.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Pedal (rudder) activity is normally only representative of aircraft control in the takeoff and approach/landing phase of the flight and should be collected during these flight phases.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• The same flight scenario should be used when comparing workload associated with two different cockpit configurations.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• A flight should be divided into segments for data collection so descriptive statistics can be computed on the continuous measures within each segment.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TABLE 9. GUIDELINES FOR THE USE OF SECONDARY TASK WORKLOAD MEASURES

SECONDARY TASK MEASURES

- When used, secondary tasks should be embedded in the flight task so as to be as non-intrusive as possible.
- Embedded secondary tasks should not appear artificial to the operator so as to maintain operator acceptance and face validity.
- Secondary tasks are most effectively implemented in a simulation environment, where air safety is not a concern and control of the environment is possible.
- Secondary task techniques should be avoided when intrusion will serve as a source of interference for the primary workload measures.

Advantages: Unlike primary task measures, secondary measures offer the advantage of providing an indication of the operator’s residual resources. Since the secondary tasks themselves are not directly linked to the primary task, they are generally transferable among different task scenarios.

Disadvantages: May artificially impact or intrude on primary task performance (and in some cases operator safety), and they are often times met will low acceptance by the operator.

TABLE 10. GUIDELINES FOR THE USE OF PHYSIOLOGICAL WORKLOAD MEASURES

PHYSIOLOGICAL MEASURES

- Data collected with physiological measures can be contaminated by physical movement. Sources of artifact should be controlled when evaluating the implementation of a workload measure.
- The data should be representative of the entire flight segment being evaluated, so some sort of averaging should be used with the flight segment.
- Care should be taken so that the flight crew is protected from hazards, such as electrical shock.
- Care should be taken to assure that the physiological assessment method appears non-career threatening to the crew members it evaluates (e.g., data collected using physiological measures should contain no diagnostic medical information).

Advantages: Do not require overt responses; well suited for tasks which are primarily cognitive in nature; record continuously throughout the task; inherently multi dimensional.

Disadvantages: High cost, expertise required for data collection and interpretation; difficulty in excluding artifacts.

2.3 SITUATIONAL AWARENESS.

The concept of situational awareness, or what operators know about their immediate situation, has only been recently identified as a topic of theoretical interest and development within human factors (Blanchard, 1993). Therefore, the number of techniques or tools is small as compared to, say, workload measures. Although there are not many tools, the framework or taxonomy of techniques is similar to that of workload. Fracker (1991) proposed such a framework which identifies three approaches to situational awareness assessment: Explicit Measures, Implicit Measures, and Subjective Measures. Table 11 summarizes the tools/techniques that fall within each of these
categories. Additional information on these measures are provided in the subsequent three sections.

TABLE 11. SITUATIONAL AWARENESS MEASUREMENT METHODS

<table>
<thead>
<tr>
<th>Explicit Measures</th>
<th>Implicit Measures</th>
<th>Subjective Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAGAT</td>
<td>Signal Detection Theory Measures</td>
<td>SART</td>
</tr>
<tr>
<td>Verbal Protocol Analysis</td>
<td>Secondary Task Measures</td>
<td>SA-SWORD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MST</td>
</tr>
</tbody>
</table>

An explicit measure, the Situational Awareness Global Assessment Technique (SAGAT) and a subjective measure, the Situational Awareness Rating Technique (SART), are the most common and have been cited quite extensively in the human performance research literature. Appendix A provides additional information on these two techniques.

2.3.1 Explicit Measures.

Explicit measures require subjects to self-report material in memory regarding experiences observed during a task. They differ from subjective ratings, in that, pilot subjects are not assigning a numerical value to the content or quality of their awareness. There are two types of explicit measures: retrospective recall and concurrent memory probes. Retrospective measures are used after a task or trial run has been completed; pilot subjects are asked to recall specific events, or the number of times a specific event (e.g., threats, system malfunctions) occurred. For example, Kibbe (1988) had laboratory subjects perform a radar warning receiver (RWR) monitoring task. During the task, five different types of threats appeared on the RWR several times. Following the task subjects were asked to recall and position threat events along a timeline representing their flight path in addition to estimating the number of times each type of threat had occurred. Kibbe found that the more severe the threat, the more accurate its recall.

However, a problem with retrospective measures is that sometimes pilots are "forgetting" and therefore are unable to recall correct information. With the concurrent, or memory probe method, the subject pilots are asked to recall specific events near the time they occurred rather than afterwards. The SAGAT measure (Bolstad, 1991; Endsley, 1990; Endsley, 1988) mentioned earlier is an excellent example of a concurrent memory probe technique; the simulation is frozen at specified times and pilots are then asked to recall information.

Sarter and Woods (1994) describe the use of a probe/concurrent technique in assessing the quality of 20 airline pilots' mental models of the operation of the Flight Management System (FMS). Probe questions were presented during the performance of a simulated flight in a Boeing 737-300 part-task trainer. Results
indicated that most of the difficulties in pilot interaction with automation is related to the lack of mode awareness.

Sullivan and Blackman (1991) describe the use of verbal protocol analysis for the assessment of situational awareness. With verbal protocol analysis techniques, pilots are asked to verbalize their thought content while executing a mission. More experienced pilots are thought to have more information stored in long-term memory stores, and therefore would require less verbalization of thought content (see section 4. Flightcrew Data Requirements, on definition of an experienced pilot).

2.3.2 Implicit Measures.

The goal of implicit measures is to derive measures of situational awareness directly from task performance rather than rely on pilot self-reporting techniques as is done with explicit type measurement techniques. The most straightforward approach in obtaining an implicit metric of situation awareness is to use signal detection theory measurement methods and techniques. For example, a subject pilot might be asked to report whether they detect a Data Link aural alert (a hit) within a high workload terminal environment, or not (miss). To account for all possible alternatives, an indication or measure of false alarms (responding as if they heard an alert when it did not actually go off) and correct negatives (not responding when an alert did not go off) are also recorded. This method could be used, for example, to identify which of two possible Data Link alerting schemes would provide better performance.

Table 12 is a schematic representation of the four possible outcomes that can occur; the outcomes are representative of those collected/measured in a classic signal detection experiment (Coren et al., 1984). This type of method has been used quite extensively in the military, specifically in the detection of enemy threats and weapons management.

<table>
<thead>
<tr>
<th>Signal</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>Hit</td>
<td>Miss</td>
</tr>
<tr>
<td>Absent</td>
<td>False Alarm</td>
<td>Correct Negative</td>
</tr>
</tbody>
</table>

Table 12. OUTCOMES OF A SIGNAL DETECTION EXPERIMENT

Target events (Data Link alerts, enemy threats, etc.,) can be tailored to match the objectives of any study, provided they conform to the following three objectives (1) target events as well as the responses must be unambiguously defined so that the
presence and absence of both are clear and countable, (2) the sets of events and responses must be finite, and (3) only one response is unique to any one target event.

Hahn and Hansman (1992), for example, designed an implicit measure of situational awareness as the ability of pilot subjects to detect erroneous clearances. The study was designed to determine the relationship of situational awareness to automated FMS programming of Data Linked clearances and the readback of ATC clearances. The results indicated that the error detection performance (measure of situational awareness) and pilot preference results indicate that automated programming of the FMS may be superior to manual programming.

The secondary task approach to deriving workload measures can also be used to derive situational awareness measures. Pilot performance tasks that contribute to overall pilot situational awareness can be indirectly measured through the performance observed on secondary task measures. The same tools and techniques identified earlier for secondary task workload methods can be adopted and used to identify situational awareness performance. As in workload measures, the secondary task should be constructed so as not to interfere with the pilots primary task. This will ensure that any effects observed will be solely due to the performance of the primary task. When using implicit measures, it is good practice to utilize SMEs in selecting pilot performance tasks which are indicative of high situational awareness.

2.3.3 Subjective Rating Measures.

As in workload, the most popular of situational awareness type ratings is the subjective rating scale. Subjective scales can either be direct or comparative. In direct ratings, pilots assign a numerical value to their situational awareness in any one flight (or flight segment). A “Likert” scale is used whereby pilot subjects are asked to assign values on a discrete, integer-based scale having an odd number of discrete options and consisting of a range generally from 1 to 5, from strongly disagree to strongly agree, respectively (Stramler, 1993). Comparative ratings are designed to evaluate a pilot’s situational awareness on one flight compared to another; a pilot would assign a value to the ratio of one to the other.

The most common used subjective scales are direct rating scales. The SART scale (appendix A), for example, has three different forms: 3-dimensional, 10-dimensional and overall. The scales were developed based on factors believed to be components of high situational awareness. A concern with direct rating scales is that because of differences in individual’s perceived rating of situational awareness, they generally cannot be compared across raters.
The comparative scales offer a way to control for subject variability by permitting ratio scores as defined earlier. Fracker and Davis (1990) developed a comparative scale based on the Subjective Workload Dominance Technique (SWORD, appendix A) called SA-SWORD.

A recent review by Metalis (1993) provides a review of research activities that employ various subjective techniques; one of these tools is the Mission Simulation Technique (MST). The MST provides a statistically weighted measure of situational awareness based on a battery of measures, which include explicit, or probe, type measures and subjective techniques. Flying performance, captured through variations in airspeed, is also factored in. The battery of measures offers a global assessment of situational awareness and is useful in evaluating, for example, a new cockpit design concept (e.g., a new avionics display) against a standard.

In conclusion, the development of situational awareness tools and techniques is only in its infancy and as such there are only a few to choose from. However, utilizing the basic framework as proposed by Fracker (1991), specialized probe questions and events/targets to measure can be developed specific to individual research objectives; that is, the researcher is not confined to just known measurement tools such as SART, SAGAT, SA-SWORD, etc. Vidulich et al. (1994) provides an annotated bibliography of over 200 papers that can be useful in investigating concepts and development of specialized situational awareness metrics and tools.

2.4 VIGILANCE.

The ability of flightcrews to maintain a constant focus of attention is a growing problem in today's advanced aircraft because of the increased levels of automation which shift flightcrews into a supervisory role from the active, manual role of the past. An assessment of recent reviews of the vigilance literature (see et al., 1994; Warm, 1993) and personal contacts with leading experts has resulted in a state-of-the-art database of information.

It was learned, however, that most vigilance research has been conducted in a strictly controlled laboratory environment in contrast to the operational research environment common to flight management (FM) research. There appears to be no documented vigilance assessment tools for the flightdeck research environment beyond direct observation methods, reaction time to respond to alarms/alerts, quality of decision making, etc. Two reasons for this deficiency have been cited in the literature. First, most vigilance tasks and associated assessment techniques that are implemented in the laboratory setting require stimulus event rates (number of alarms/signals per hour, etc.,) that would exceed those in a typical flight simulation environment. Stramler (1993) defines a vigilance type task to be:
A state in which an individual sustains a high level of attention in an attempt to detect a signal, a change in a signal, or a particular activity.

Secondly, in order for researchers to discover subtle changes in pilot behavioral effects, they need to minimize the number of concurrent operator tasks. In a multitask flight simulation environment, it would be difficult to pinpoint causal effects due to uncontrolled or extraneous variables. As a result, the experiments are typically not representative of real-life tasks, let alone aviation-type tasks.

Those experiments which are representative are usually conducted during actual inflight operational settings. For example, Cabon et al., (1991), conducted a recent study in collaboration with Airbus Industrie to identify factors which can modify a flightcrews vigilance and performance during long-haul flights. They utilized a battery of physiological data (EKG, heart rate, and motor activity) and direct observation of aircrew tasks. Results revealed that generally low vigilance phases are identified around 30 minutes after takeoff, at the beginning of the cruise phase of flight, and during periods of low communication. A total of 45 of 50 planned flights were studied. Flights were achieved on different kinds of planes, including, the Airbus A320 and Boeing 747-400 and Douglas DC10.

Because of resources and the amount of cost involved, conducting inflight experiments is not always available to the everyday practitioner. Based on this assumption, one could also argue that this may also contribute to the low number of vigilance experiments conducted in operational settings. Given the lack of available tools in the aviation simulation setting, the rest of this section will emphasize some significant advancements in the area of vigilance research that deserve mentioning.

It has been stated (Dember and Warm 1979; Parasuraman, 1984) performance decrements observed during vigilance tasks are characteristic of operator underload conditions, or periods of low workload. However, a number of more recent published works (Warm et al., 1991; Becker et al., 1991 and 1992; Scerbo et al., 1992) have reexamined the effect that vigilance type tasks have on the human operator. Warm et al., (1991) provides the following discussion on the relation of vigilance and workload:

Vigilance tasks do not represent underload situations. Instead, the cost of mental operations in vigilance is substantial, with mental demand and frustration the primary contributors. Recognition of the workload characteristics of vigilance tasks may be helpful in understanding both vigilance performance and the stress associated with the need to sustain attention for long periods of time.
In the Warm, Becker and Scerbo studies, workload ratings were obtained in nonaviation type laboratory tasks with the NASA-TLX subjective workload rating scale (appendix A). The results of their studies have implications in the aviation environment. For example, reducing workload through automation, and therefore increasing the pilot's monitoring load may be counterproductive; it could be increasing workload—not reducing it.

Vigilance research has also shown that motivation (Dember et al., 1992) plays a major role in the ability of operators to perform well in vigilance type tasks (monitoring of automation, etc.,). A motivational variable, choice, allows the operator to select among different opportunities for action. In a practical sense, a flightcrew that had a choice in what operations were manual versus automated would have a countermeasure to monitoring inefficiency (Parasuraman et al., 1994). In a dynamic automation environment, flightcrews would be more closely coupled to the system and would detect malfunctions/failures at a higher rate than with a constant, or static automation system. The ability of the flightcrew to tailor automation to specific needs is known as "adaptive automation." Morrison and Gluckman (1994) and Scerbo (1994) discuss recent concepts of adaptive automation that should be considered. Both articles provide discussion on a general framework for implementing adaptive automation concepts and address research based principles and guidelines for its use.

In addition to Scerbo (1994), two other recent published works on adaptive automation, such as Molloy and Parasuraman's (1992) and Singh's et al., (1993), provide references to tools which may be applicable to vigilance research. In all three articles, monitoring of failures was recorded in a crude flight-simulation task, a revised version of the MultiAttribute Task (MAT) Battery, developed by Comstock and Arnegard (1992). The MAT is a multitask flight simulation package comprising component tasks of tracking, system monitoring, fuel management, communications, and scheduling, each of which can be performed manually or under automation control. The MAT technique can be used to evaluate the effects of various automation strategies on pilot's decision making and system monitoring capabilities. In addition to evaluating monitoring performance, Singh et al., (1993), measured operator individual differences with the Complacency Potential Rating Scale (CPRS) developed in an earlier study (Singh et al., 1992). The CPRS subjective rating scale measures operator complacency on four complacency-related dimensions: trust, confidence, reliance, and safety.

In summary, vigilance related research is typically reserved for the part-task laboratory environment. Related research disciplines, such as workload and adaptive automation, have yielded tools (e.g., MAT, CPRS, and NASA-TLX) that might prove valuable as research tools in the evaluations of Data Link related automation concepts.
3. STUDY CLASSIFICATIONS.

The second main objective of this research was to define how the measurement techniques/tools identified for workload, situational awareness and vigilance could best be allocated to the FAA Technical Center's pilot/crew performance studies. This section will be divided into two parts. The first part provides definitions of the various study classifications, and the second part provides information on how the various tools should be allocated to each study environment.

3.1 DEFINITIONS OF STUDY CLASSIFICATIONS.

There are three main types of study classifications envisioned to be used by the FAA during the conduct of their pilot/crew performance studies: part-task, full-mission, and end-to-end. Together, these classes represent the spectrum of behavioral test environments (excluding the operational flight test environment) contained in a systems engineering process to human performance assessment (Blanchard and Fabrycky, 1990).

The major underlying difference between the types of studies is the level of fidelity; as you move from part-task to end-to-end type evaluations the level of fidelity of the experiment, the cost and amount of time for preparation, and associated risk are increasing. Figure 1 provides a graphic of this trend.

![Graph showing increased cost, time, and risk with increased fidelity of experiment]

FIGURE 1. INCREASED COST, TIME AND RISK AS A LEVEL OF EXPERIMENT FIDELITY

The FAA has already in place various research testbeds, such as the RCS and the high fidelity CSN. The CSN can be comprised of a number of external simulators, including the RCS, that can be remotely tied into the FAA's unique ATC simulation facilities. These test vehicles/platforms provide the capability to evaluate
a variety of cockpit control/display system interface, operational, procedural and system level issues. The sections which follow define the characteristics of each study class.

3.1.1 Part-Task Environment.

Part-task studies are evaluations used to extract basic human behavior performance. Typically, mockups or scaled representations of the physical characteristics of equipment and systems are used. A mockup is a tool to assist development by enhancing conceptualization of the human machine interface and to evaluate functional, operational, and procedural issues. Furthermore, only a single subsystem, or component of the flightdeck is evaluated. For example, the mockup could be as simple as a two-dimensional cardboard representation of a control panel layout, or as sophisticated as a computer representation of the subsystem (ANSI/AIAA, 1992). The most elaborate mockup would be, for example, the system itself. Using the FMS Control Display Unit (CDU), for example, one could evaluate basic interface design issues; how the menus could be formatted, how they could be accessed, determine the types of errors made, and problems encountered by the operator. A series of part-task studies may be required before optimal format design is achieved or before the next level of evaluation is desired.

The sophistication level of the mockup required depends on the particular human behavior of interest. For example, simple interface characteristics such as color or font, may be evaluated on simple paper mockups. Conversely, more complex issues such as, location or reachability, may require a device such as the RCS. In conclusion, the more elaborate the mockup or representation is, the higher the face validity of the experiment.

3.1.2 Full Mission Simulations.

The full-mission simulation study is a level of fidelity higher than part-task studies, in that the focus is not on just one subsystem or component on the flightdeck; it is a multitask environment. In full-mission studies, the practitioner can assess combined operational, procedural, and integration issues that are not of interest, collectively, in part-task evaluations. The interaction of systems, the interaction and communication of flightcrews is more realistic. Studies can be tailored towards specific flight regimes, such as landing, takeoff or cruise, and/or the transitions between these environments. ATC functions do not need to be elaborate; the focus is on the flightdeck and the behaviors and performance of the flightcrew. Depending on what is being evaluated, ATC functions may not be necessary at all. When conducting Data Link studies some form of ATC representation needs to be included. However, a full-mission study does not require a high fidelity simulation testbed such as the FAA's ATC simulation network. High fidelity ATC simulation is reserved for end-to-end simulations which is discussed in the next section.
3.1.3 End-to-End Simulations.

Short of an actual flight test, an end-to-end simulation attempts to identify operational readiness of a proposed system. An end-to-end evaluation involves realistic and elaborate simulations of both ground and airborne operations and is used to investigate pilot and controller interactions in a fully scripted operational scenario. The fidelity of the simulation allows investigation of operational, procedural, and integration issues from a system point of view, not necessarily just from the cockpit perspective.

3.2 MEASUREMENT BY STUDY CLASSIFICATION.

The task of selecting an appropriate and practical measure for a particular study classification can be a difficult task. First and foremost, proper allocation requires knowledge about the specific objectives of the test, that is, what specific human behavior is of interest. For example, if situational awareness measures are desired, proper allocation requires knowing the limitations and constraints of the test environment; if a simulator does not have "freeze" capability, then the class of explicit/probe situational awareness techniques (e.g., SAGAT) would not be possible.

Secondly, based on the discussion on vigilance, it was found that vigilance measures are geared towards part-task laboratory environments, with exception to the operational in-flight test environment. However, it was concluded that tools from related disciplines (e.g., adaptive automation, complacency, workload) may be useful.

Third, there is also the subject of cost. Physiological measures of workload, for example, require specialized equipment that may not be available. Implementing the measures and analyzing the data may require an expert; obtaining such an expert may not be cost effective.

A fourth factor to consider, for example, is operational realism. End-to-end simulations, for example, are the highest fidelity study types; a measure that is intrusive or requires artificial measurement techniques such as a secondary task (e.g., fingertapping) would interfere with the pilot's flying task. Embedded secondary measures that use existing piloting type tasks would be the proper choice.

The aforementioned examples are only a few of the questions that might be considered in choosing an appropriate measure. To help simplify the complexities involved, two approaches are recommended and outlined. First, the information summarized in Table 13 provides in convenient matrix form guidelines to follow in choosing a broad class of measures for the study classifications identified. Although no specific tools are mentioned in Table 13, the guidelines should help in identifying if a particular class of workload, situational awareness and
vigilance measures is appropriate. The guidelines are derived from knowledge obtained through review of the human performance literature and contacts with SMEs in the field. Note the caveats provided in table 13.

TABLE 13. MATRIX OF HUMAN PERFORMANCE MEASURES BY STUDY CLASSIFICATION

<table>
<thead>
<tr>
<th>Human Performance Measure</th>
<th>Part-Task Environment</th>
<th>Full Mission Simulation</th>
<th>End-to-End Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workload</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subjective</td>
<td>X¹</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Primary Task</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Secondary Task</td>
<td>X</td>
<td>X</td>
<td>X²</td>
</tr>
<tr>
<td>Physiological</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Situational Awareness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explicit</td>
<td>X³</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Implicit</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Subjective</td>
<td>X¹</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Vigilance</td>
<td>X</td>
<td>X⁴</td>
<td>X³</td>
</tr>
</tbody>
</table>

¹ Limited on Generalizability of Results
² Embedded Secondary Task Only
³ 3-dimensional mockup (e.g., RCS) or part-task trainer required.
⁴ Extensive Resource Commitment Required. Simulation Time is Excessive.

A second approach would be to utilize available expert systems. Recent advances in expert systems allow a user to interact with microcomputer based applications in selecting appropriate assessment techniques. Two tools, the Operator Workload Knowledge-based Expert System Tool (OWLKNEST, Hill and Harris, 1990) and the Workload Consultant for Field Evaluations (WC FIELDE, Casper et al., 1987) have been widely used in research and can be used. WC FIELDE is currently a tool available from CSERIAC’S product and services line. Although, OWLKNEST and WC FIELDE are designed to assist in determining an appropriate workload technique, the questions which are asked may determine, for example, the use of a post-test subjective situational awareness tool as opposed to a concurrent (during simulation) probe type technique. No expert tools for situational awareness or vigilance type measures were discovered in the literature.

Table 14 shows some example questions (11 total) which were extracted from 23 questions that formed the basis of the OWLKNEST expert system tool development program (Lysaght et al., 1989). Those selected were determined to be relevant to “operator-in-the-loop” evaluations. Note also in table 14 that a description is also provided of the reason why the questions are asked. Similar questions are contained in the WC FIELDE expert system tool.
### TABLE 14. SOME QUESTIONS TO CONSIDER IN SELECTING AN APPROPRIATE WORKLOAD MEASURE

<table>
<thead>
<tr>
<th>Question that may be asked</th>
<th>Reason for the question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What is the time frame in which the workload analysis must be complete?</td>
<td>Determine the impact of the analysis time frame on techniques selected, e.g., if time is short use subjective techniques.</td>
</tr>
<tr>
<td>2. What computer software facilities are available?</td>
<td>If no software exists, then use pencil and paper subjective techniques.</td>
</tr>
<tr>
<td>3. What sort of laboratory facilities are available for empirical work?</td>
<td>Some empirical techniques require specialized facilities or equipment. Primary and secondary techniques may require equipment to present tasks and record responses. Subjective techniques may use computers or paper and pencil. Physiological techniques may require equipment, such as sensors, to record physical responses.</td>
</tr>
<tr>
<td>4. What staff support is available either in house or through another organization?</td>
<td>It is necessary to have the expertise (or the expert) available on various topics.</td>
</tr>
<tr>
<td>5. How much staff or manpower is available to do the workload analysis?</td>
<td>Certain empirical techniques are very labor intensive. Certain techniques are more flexible than others in terms of manpower requirements.</td>
</tr>
<tr>
<td>6. Why is workload assessment being done?</td>
<td>The reason (e.g., comparison of two or more candidate systems, examination of individual differences) assessment is being done will influence the types of techniques used</td>
</tr>
<tr>
<td>7. Is this a derivative system or a brand new one?</td>
<td>If it is a derivative system then the system can probably be tested in a generic simulator using the old simulation model with mock-ups of the new operator controls and procedures.</td>
</tr>
<tr>
<td>8. What are the primary measures of human performance in the system?</td>
<td>Primary measures are time, accuracy (or error), both time and accuracy, fine structure of behavior.</td>
</tr>
<tr>
<td>9. What operator performance characteristics are relevant to the Particular man-machine system?</td>
<td>Categories of behavior expected to be influenced by the man-machine system: Perceptual, Mediational, Communication, mediational or communication processes</td>
</tr>
<tr>
<td>10. Can the operator be interrupted during a mission or are there blocks of time during the mission in which the operator can fill out forms.?</td>
<td>Subjective measures require some time for filling out the rating forms. If the operator cannot be interrupted, then it is better to video tape the session and have the ratings completed later.</td>
</tr>
<tr>
<td>11. Does the operator have spare time to do other things at various points in the mission?</td>
<td>Secondary tasks may be used if there is some spare time.</td>
</tr>
</tbody>
</table>

(Source: Lysaght et al., 1989)

In conclusion, the two approaches identified can help in determining the most appropriate technique. With regards to table 13 it appears that (except for a few exceptions) any workload, situational awareness, and vigilance assessment technique can be employed in any study class identified. However, practitioners must be careful in interpreting data from part-task simulations; the more artificial the test environment...
is the less amount of confidence in generalizing the collected data to the real-world (Mitman et al., 1994b). Conversely, sacrificing measurement control (part-task) in lieu of operational realism (end-to-end) can invite additional problems, namely, problems associated with the control of unwanted, unrealized variables that can affect your data. To this end, the solution would be to utilize all types of study classifications in the research, design, and evaluation of aviation system concepts.

4. FLIGHTCREW DATA REQUIREMENTS.

The third main objective of this research was to determine a set of criteria or guidelines in selecting pilot subjects for evaluation. A number of factors, such as experience level, differences in type rating, number of flightcrews, etc., were extracted from the literature and are reported in the following. Bulleted items refer to recommended guidelines which are followed by supporting comments from the research literature.

Guidelines

- Studies which are conducted in a simulator, such as a full-mission or end-to-end evaluation, should employ if possible current type rated pilots in the aircraft being evaluated.

However, because of cost and time constraints, it may not be possible to obtain current type rated pilots in that aircraft. In this case, personnel that are essentially equivalent in terms of training and skill level should be used (ANSI/AIAA, 1992). If this should fail, then consider restructuring the tasks to be measured, such that special experience or particular pilot attributes are not required.

- If the focus is on a particular type subsystem or display, then experience with that system would prove beneficial.

However, some concerns have been expressed with regards to use of only highly trained or experienced subjects, in the evaluations of new and/or prototype display systems. These users may experience little or no problems with the system and therefore results would not be indicative or representative of how the system really is. Jorna, 1992, reports that using newly trained subjects or less-experienced pilots in evaluating prototype systems may be better because they are relatively free of experience biases.

Fracker, 1991, points out that "if experienced or only partially trained operators are included in a study, the correlation between measured 'situational awareness' and the criterion may appear low for reasons that have nothing to do with the
Situational Awareness (SA) metric itself.” Therefore, the best solution is to use a broad level of experience levels, both highly trained (e.g., computer smart) and naive users. However, at the end of the study, a differential analysis of the two groups might prove useful.

Based on the review of the literature, it was never explicitly stated what amount of flight hours denotes an “experienced” pilot versus a “nonexperienced” pilot. For practical considerations, and for use as a general guideline, CSERIAC recommends at least 1000 hours flight time. This is only a general guideline and as such other factors may need to be considered, for example, specific equipment time, captain versus first-officer time, individual aircraft time, etc.

- Subjects that are type rated in the same aircraft, but represent different airlines, is acceptable.

The above guideline depends in part on the degree of difference between two company’s aircraft configurations and the specific component(s) being studied. If more than one component is being measured, than the differences between two aircraft configurations is more important. For example, one airline might exercise automation levels at a higher rate than others. In any case, some investigation into airline procedures prior to the test would prove useful. Prior to a study conducted by Battiste and Bortolussi (1988), difference training was conducted to acquaint 19 airline captains with a Delta configured Boeing 727. The number of hours required for difference training varied from 2 to 4 hours and continued until all subjects were at a common level of awareness. If the required amount of training exceeds time and budget constraints, then perhaps pilots from the same company should be used.

- More than one crew is necessary in any evaluation.

It is not acceptable to have only one crew as test subjects. This is because a single crew may, for unknown reasons, have certain peculiarities that judge the design atypical (ANSI/AIAA, 1992). An absolute minimum of two crews is necessary, and more than two are desirable, especially when evaluating designs in an operational setting.

General Comments

A review of the literature did not uncover any guidelines on the issues of age and gender. In both cases, consideration should be given to the amount of experience only and to what type of aircraft, subsystems, etc., they have experience with. However, a point worth mentioning is that in limited observation and discussion older pilots tend to not accept new technology as do younger pilots. Based on this notion, the selection of pilot
subjects should represent pilots of all ages and different training emphasized.

Depending on pilot experience levels and associations with different airlines, a sufficient amount of training may be conducted prior to a study or evaluation. The training or shakedown phase serves two purposes (1) it ensures that all study participants are trained in a similar way and up to a certain level, and (2) it provides immediate feedback to the practitioner whether a certain subject would qualify as a test participant in subsequent data collection exercises. In human factors performance research pilots must be trained to a point where there is little continued improvement, or until an asymptotic level of performance is achieved (ANSI/AIAA, 1992).

To conclude, a recent historical review of the human factors literature (Moroney and Reising, 1992) was conducted to assess various characteristics of subjects employed in human factors experiments. The results revealed that approximately 41 percent of the articles provided inadequate detail on how experimental subjects were selected. It would appear that this is an inadequacy in the human factors research industry.

As a way to build upon the set of flightcrew data requirements and guidelines, CSERIAC will establish a database of pilot subject characteristics. The database will contain subject characteristics of pilot/crew personnel used in past and future data link system studies. Correlations between subject characteristics and performance, workload, situational awareness, etc., will be drawn where possible and justifiable. Additionally, any interesting facts or observations will also be recorded.

5. SUMMARY.

To summarize, an extensive literature search utilizing CSERIAC in-house and local capabilities was conducted to determine state-of-the-art human performance measures in the area of workload, situational awareness, and vigilance. Having obtained various tools, guidelines were then created to determine (1) how these tools should be allocated to study classifications envisioned to be used by the FAA, and (2) what pilot subject characteristics should be considered prior to conducting an evaluation.

As a result of this effort, over 190 references and over 10 personal contacts with experts in the field were made. At least two documents deserve special mentioning as they were cited frequently in the handbook. They were (1) Operator Workload: Comprehensive Review and Evaluation of Operator Workload Methodologies (Lysaght, et al., 1989), and (2) Guide to Human Performance Measurements (ANSI/AIAA, 1992).

As expected, it was found that more tools/techniques exist for workload measures as opposed to the other human performance measures reviewed. However, it was revealed that a few workload
techniques (SWORL, NASA-TLX embedded secondary task, etc.) are adaptable to both situational awareness and vigilance studies. Furthermore, related disciplines, such as adaptive automation and complacency, uncovered tools (e.g., MAT, CPRS) that can also be adapted to FAA human factors Data Link research. Continued awareness with regards to these disciplines will be emphasized and any new developments will be noted.

Future recommendations are to continue updating this handbook as new tools and techniques become available, especially in the area of situational awareness and vigilance. Both areas are topics which are recently receiving the most attention.

To conclude, the database of pilot subject selection criteria will be maintained as an active database; results, ideas, etc., from future FAA Data Link human factors research will be recorded. The network of human performance experts in the field will be strengthened through continued conversations, visits, and associations at various meetings and symposia.
6. REFERENCES.


Reynolds, M. (1994). Certification Of The Data Link Installation Within the AVIA Boeing 727-100 Flight Simulator in Accordance with FAA Notice 8110.50, Guidelines For Airworthiness Approval Of Airborne Data Link Systems And Applications. CSERIAC-FAA-TR-94-
06. Dayton, OH: Crew System Ergonomics Information Analysis Center.


7. **LIST OF ABBREVIATIONS.**

AFB  
Air Force Base

AHP  
Analytical Hierarchy Process

AIAA  
American Institute of Aeronautics and Astronautics

ANSI  
American National Standards Institute

ATC  
Air Traffic Control

CDU  
Control Display Unit

CEP  
Evoked Cortical Potentials

CFF  
Critical Flicker Frequency

CPRS  
Complacency Potential Rating Scale

CSERIAC  
Crew System Ergonomics Informatics Analysis Center

CSN  
Cockpit Simulator Network

DOD  
Department of Defense

DOT  
Department of Transportation

DTIC  
Defense Technical Information Center

ECG  
Electrocardiogram

EDA  
Electrodermal Activity

EEG  
Electroencephalographic Activity

EKG  
Electrocardiogram

EMG  
Electromyographic Activity

ERP  
Event Related Potentials

FAA  
Federal Aviation Administration

FMS  
Flight Management System

FOM  
Figure-Of-Merit

HRV  
Heart Rate Variability

ILS  
Instrument Landing System

IPT  
Interval Production Task

MAT  
Multi-Attribute Task Battery

MEG  
Magnetoencephalographic Activity

NASA  
National Aeronautics and Space Administration

NASA-TLX  
NASA Task Load Index

NTIS  
National Technical Information Service

NTTC  
National Technology Transfer Center

ORTA  
Office of Research and Technology Assessment

OW  
Overall Workload

OWLKNEST  
Operator Workload Knowledge-Based Expert System Tool

PET  
Positron Emission Tomography

POSWAT  
Pilot Objective/Subjective Workload Assessment Technique

PSE  
Pilot Subjective Evaluation

RCBF  
Regional Cerebral Blood Flow

RCS  
Reconfigurable Cockpit System

RWR  
Radar Warning Receiver

SA  
Situational Awareness

SAGAT  
Situational Awareness Global Assessment Technique

SART  
Situational Awareness Rating Technique

SME  
Subject Matter Expert

SWAT  
Subjective Workload Assessment Technique

SWORD  
Subjective Workload Dominance Technique

TGF  
Target Generation Facility

TPM  
Technical Program Manager

VTOL  
Vertical Takeoff and Landing
<table>
<thead>
<tr>
<th>WC FELDE</th>
<th>Workload Consultant for Field Evaluations</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCVTE</td>
<td>Workload/Compensation/Interference/Technical Effectiveness</td>
</tr>
<tr>
<td>WPAFB</td>
<td>Wright-Patterson Air Force Base</td>
</tr>
</tbody>
</table>
APPENDIX A

DESCRIPTION OF MEASUREMENT TOOLS AND TECHNIQUES
APPENDIX A. DESCRIPTION OF MEASUREMENT TOOLS AND TECHNIQUES

TABLE OF CONTENTS

SUBJECTIVE WORKLOAD ASSESSMENT TECHNIQUES

<table>
<thead>
<tr>
<th>Technique</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytical Hierarchy Process</td>
<td>A-3</td>
</tr>
<tr>
<td>Bedford Workload Scale</td>
<td>A-5</td>
</tr>
<tr>
<td>Cooper-Harper Rating Scale</td>
<td>A-7</td>
</tr>
<tr>
<td>Crew Status Survey</td>
<td>A-8</td>
</tr>
<tr>
<td>Dynamic Workload Scale</td>
<td>A-10</td>
</tr>
<tr>
<td>Equal-Appearing Intervals</td>
<td>A-11</td>
</tr>
<tr>
<td>Flight Workload Questionnaire</td>
<td>A-12</td>
</tr>
<tr>
<td>Hart and Hauser Rating Scale</td>
<td>A-13</td>
</tr>
<tr>
<td>Honeywell Cooper-Harper Rating Scale</td>
<td>A-14</td>
</tr>
<tr>
<td>Magnitude Estimation</td>
<td>A-15</td>
</tr>
<tr>
<td>Modified Cooper-Harper Rating Scale</td>
<td>A-16</td>
</tr>
<tr>
<td>NASA Bipolar Rating Scale</td>
<td>A-18</td>
</tr>
<tr>
<td>NASA Task Load Index</td>
<td>A-20</td>
</tr>
<tr>
<td>Overall Workload Scale</td>
<td>A-24</td>
</tr>
<tr>
<td>Pilot Objective/Subjective Workload Assessment Technique</td>
<td>A-25</td>
</tr>
<tr>
<td>Pilot Subjective Evaluation</td>
<td>A-26</td>
</tr>
<tr>
<td>Subjective Workload Assessment Technique</td>
<td>A-27</td>
</tr>
<tr>
<td>Subjective Workload Dominance Technique</td>
<td>A-30</td>
</tr>
<tr>
<td>Workload/Compensation/Interference/Technical Effectiveness</td>
<td>A-31</td>
</tr>
</tbody>
</table>

SITUATIONAL AWARENESS ASSESSMENT TECHNIQUES

<table>
<thead>
<tr>
<th>Technique</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crew Situational Awareness</td>
<td>A-32</td>
</tr>
<tr>
<td>Situational Awareness Global Assessment Technique</td>
<td>A-33</td>
</tr>
<tr>
<td>Situational Awareness Rating Technique</td>
<td>A-34</td>
</tr>
</tbody>
</table>

Bibliography ......................................................... A-36

Caveat: The narrative descriptions provided for both the workload and situational tools are as provided in the Guide to Human Performance Measurements, American National Standards Institute (ANSI)/American Institute of Aeronautics and Astronautics (AIAA), 1992. Only those tools considered to be valuable to the FAA planned data link human performance research studies are contained in this appendix.
LIST OF FIGURES

Figure A1: AHP Rating Scale......................................................... A-3
Figure A2: Bedford Workload Scale................................................... A-5
Figure A3: Cooper-Harper Rating Scale............................................ A-7
Figure A4: Crew Status Survey.......................................................... A-9
Figure A5: Dynamic Workload Scale.................................................. A-10
Figure A6: Hart and Hauser Rating Scale.......................................... A-13
Figure A7: Honeywell Cooper-Harper Rating Scale............................ A-14
Figure A8: Modified Cooper-Harper Rating Scale.............................. A-16
Figure A9: NASA Bipolar Rating Scale............................................. A-19
Figure A10: NASA TLX Rating Sheet ............................................... A-21
Figure A11: Pilot Subjective Evaluation Scale.................................... A-26
Figure A12: The WCI/TE Scale Matrix............................................. A-31

LIST OF TABLES

Table A1: Definitions of AHP Scale Descriptors.................................. A-3
Table A2: NASA Bipolar Rating-Scale Definitions.............................. A-18
Table A3: Rating-Scale Definitions.................................................... A-20
Table A4: SWAT Scales........................................................................ A-27
Table A5: Definitions of SART Rating Scales..................................... A-34
Analytical Hierarchy Process

General description - The analytical hierarchy process (AHP) uses the method of paired comparisons to measure workload. Specifically, subjects rate which of a pair of conditions has the higher workload. All combinations of conditions must be compared. Therefore, if there are \( n \) conditions, the number of comparisons is \( 0.5n(n - 1) \).

Strengths and limitations - Lidderdale (1987) found high consensus in the ratings of both pilots and navigators for a low-level tactical mission. Complex mathematical procedures must be employed (Lidderdale, 1987; Lidderdale and King, 1985; Saaty, 1980). Budescu, Zwick, and Rapoport (1986) provide critical value tables for detecting inconsistent judgments and subjects. Vidulich and Tsang (1987) concluded that AHP ratings were more valid and reliable than either an overall workload rating or NASA-TLX. Vidulich and Bortolussi (1988a) reported that AHP ratings were more sensitive to attention than secondary reaction times. Vidulich and Tsang (1988) reported high test/retest reliability.

Data requirements - Four steps are required to use the AHP. First, a set of instructions must be written. A verbal review of the instructions should be conducted after the subjects have read the instructions to ensure their understanding of the task. Second, a set of evaluation sheets must be designed to collect the subjects' data. An example is presented in Figure A1. Each sheet has the two conditions to be compared in separate columns, one on the right side of the page, the other

<table>
<thead>
<tr>
<th>WORKLOAD JUDGMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSOLUTE</td>
</tr>
<tr>
<td>ILS APPROACH WITH HUD</td>
</tr>
<tr>
<td>ILS APPROACH WITHOUT HUD</td>
</tr>
</tbody>
</table>

Figure A1: AHP Rating Scale

on the left. A 17-point rating scale is placed between the two sets of conditions. The scale uses five descriptors in a pre-defined order and allows a single point between each for mixed ratings (see Figure A1). Vidulich (1988) defined the scale descriptors (see Table A1). Third, the date

Table A1: Definitions of AHP Scale Descriptors

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQUAL</td>
<td>The two task combinations are absolutely equal in the amount of workload generated by the simultaneous tasks.</td>
</tr>
<tr>
<td>WEAK</td>
<td>Experience and judgment slightly suggest that one of the combinations of tasks has more workload than the other.</td>
</tr>
<tr>
<td>STRONG</td>
<td>Experience and judgment strongly suggest that one of the combinations has higher workload.</td>
</tr>
<tr>
<td>VERY STRONG</td>
<td>One task combination is strongly dominant in the amount of workload, and this dominance is clearly demonstrated in practice.</td>
</tr>
<tr>
<td>ABSOLUTE</td>
<td>The evidence supporting the workload dominance of one task combination is the highest possible order of affirmation (adapted from Vidulich, 1988, p. 5).</td>
</tr>
</tbody>
</table>
must be scored. The scores range from +8 (absolute dominance of the left-side condition over the right-side condition) to -8 (absolute dominance of the right-side condition over the left-side condition). Finally, the scores are input, in matrix form, into a computer program. The output of this program is a scale weight for each condition and three measures of goodness of fit.

*Thresholds* - Not stated.
Bedford Workload Scale

General description - Roscoe (1984) described a modification of the Cooper-Harper scale created by trial and error with the help of test pilots at the Royal Aircraft Establishment at Bedford, England. The Bedford Workload Scale retained the binary decision tree and the four- and ten-rank ordinal structures of the Cooper-Harper scale (see Figure A2). The three-rank ordinal structure asked pilots to assess whether: (1) it was possible to complete the task, (2) the workload...
was tolerable, and (3) the workload was satisfactory without reduction. The rating-scale end points were; workload insignificant to task abandoned. In addition to the structure, the Cooper-Harper (1969) definition of pilot workload was used: “... the integrated mental and physical effort required to satisfy the perceived demands of a specified flight task.” The concept of spare capacity was used to help define levels of workload.

Strengths and limitations - The Bedford Workload Scale was reported to be welcomed by pilots. Roscoe (1984) reported that pilots found the scale “easy to use without the need to always refer to the decision tree.” He also noted that it was necessary to accept ratings of 3.5 from the pilots. These statements suggest that the pilots emphasized the ten- rather than the four-rank, ordinal structure of the Bedford Workload Scale. Roscoe (1984) found that pilot workload ratings and heart rates varied in similar manners during close-coupled inflight maneuvers in a Bae 125 twin-jet aircraft. He felt that the heart-rate information complemented and increased the value of subjective workload ratings. He also noted the lack of absolute workload information provided by the Bedford Workload Scale and by heart-rate data. Wainwright (1987) used the scale during certification of the Bae 146 aircraft. Tsang and Johnson (1987) concluded that the Bedford Scale provided a good measure of spare capacity. Roscoe (1987) reported that the scale was well accepted by aircrews. Corwin, Sandry-Garza, Biferno, Boucek, Logan, Jonsson, and Metalis (1989) concluded that the Bedford scale is a reliable and valid measure of workload based on flight simulator data. Vidulich and Bortolussi (1988b) reported significant differences in Bedford ratings across four flight segments. However, the workload during hover was rated less than that during hover with a simultaneous communication task. Further, the scale was not sensitive to differences in control configurations nor combat countermeasure conditions. Finally Lidderdale (1987) reported that post-flight ratings were very difficult for aircrews to make.

Data requirements - Roscoe (1984) suggested the use of short, well-defined flight tasks to enhance the reliability of subjective workload ratings.

Thresholds - Minimum value is 1, maximum is 10.
Cooper-Harper Rating Scale

General description - The Cooper-Harper Rating Scale is a decision tree that uses adequacy for the task, aircraft characteristics, and demands on the pilot to rate handling qualities of an aircraft (see Figure A3).

Strengths and limitations - The Cooper-Harper Rating Scale is the current standard for evaluating aircraft handling qualities. It reflects differences in both performance and workload and is behaviorally anchored. It requires minimum training, and a briefing guide has been developed (see Cooper and Harper, 1969, pp. 34-39). Cooper-Harper ratings have been sensitive to variations in controls, displays, and aircraft stability (Crabtree, 1975; Krebs and Wingert, 1976; Labacqz and Aiken, 1975; and Schultz, Newell, and Whitbeck, 1970). Conner and Wierwille (1983) reported significant increases in Cooper-Harper ratings as the levels of wind gust increased and/or as the aircraft pitch stability decreased. Harper and Cooper (1984) describe a series of evaluations of the rating scale.

Data requirements - The scale provides ordinal data that must be analyzed accordingly. The Cooper-Harper scale should be used for workload assessment only if handling difficulty is the major determinant of workload. The task must be fully defined for a common reference.

Thresholds - Ratings vary from 1 (excellent, highly desirable) to 10 (major deficiencies). Noninteger ratings are not allowed.

HANDLING QUALITIES RATING SCALE

![Decision Tree Diagram]

Figure A3: Cooper-Harper Rating Scale
Crew Status Survey

General description - The original Crew Status Survey was developed by Pearson and Byars (1956) and contained 20 statements describing fatigue states. The staff of the Air Force School of Aerospace Medicine Crew Performance Branch, principally Storm, Perelli, and Gray, updated the original survey. They selected the statements anchoring the points on the fatigue scale of the survey through iterative presentations of drafts of the survey to aircrew members. The structure of the fatigue scale was somewhat cumbersome, since the dimensions of workload, temporal demand, system demand, system management, danger, and acceptability were combined on one scale. However, the fatigue scale was simple enough to be well received by operational crews. The fatigue scale of the survey was shortened to seven statements and subsequently tested for sensitivity to fatigue as well as for test/retest reliability (Miller and Narvaez, 1986). Finally, a seven-point workload scale was added. The current Crew Status Survey (see Figure A4) provides measures of both self-reported fatigue and workload as well as space for general comments. Ames and George (1992) are modifying the workload scale to enhance reliability. Their scale descriptors are:

(1) Nothing to do; No system demands.

(2) Light activity; Minimum demands.

(3) Moderate activity; Easily managed; Considerable spare time.

(4) Busy; challenging but manageable; Adequate time available.

(5) Very busy; Demanding to manage; Barely enough time.

(6) Extremely busy; Very difficult; Nonessential tasks postponed.

(7) Overloaded; System unmanageable; Important tasks undone; unsafe. (p4).

Strengths and limitations - These scales have been found to be sensitive to changes in task demand and fatigue but are independent of each other (Courtright, Frankenfeld, and Rokicki, 1986). Storm and Parke (1987) used the Crew Status Survey to assess the effects of temazepam on FB-111A crewmembers. The effect of the drug was not significant. The effect of performing the mission was, however. Specifically, the fatigue ratings were higher at the end than at the beginning, of a mission. Gawron, et al. (1988) analyzed Crew Status Survey ratings made at four times during each flight. They found a significant segment effect on fatigue and workload. Fatigue ratings increased over the course of the flight (preflight = 1.14, predrop = 1.47, postdrop = 1.43, and postflight = 1.56). Workload ratings were highest around a simulated air drop (preflight = 1.05, predrop = 2.86, postdrop = 2.52, and postflight = 1.11). George, Nordeen, and Thurmond (1991) collected workload ratings from Combat Talon II aircrew members during arctic deployment. None of the median ratings were greater than four. However, level 5 ratings occurred for navigators during airdrops and self-contained approach run-ins. These authors also used the Crew Status Survey workload scale during terrain-following training flights on Combat Talon II. Pilots and copilots gave a median rating of 7. The ratings were used to identify major crewstation deficiencies. However, George and Hollis (1991) reported confusion between
adjacent categories at the high workload end of the Crew Status Survey. They also found adequate ordinal properties for the scale but very large variance in most order-of-merit tables.

<table>
<thead>
<tr>
<th>NAME</th>
<th>DATE AND TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBJECTIVE FATIGUE</td>
<td></td>
</tr>
<tr>
<td>(Circle the number of the statement which describes how you feel RIGHT NOW.)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Fully Alert; Wide Awake; Extremely Peppy</td>
</tr>
<tr>
<td>2</td>
<td>Very Lively; Responsive; But Not At Peak</td>
</tr>
<tr>
<td>3</td>
<td>Okay; Somewhat Fresh</td>
</tr>
<tr>
<td>4</td>
<td>A Little Tired; Less Than Fresh</td>
</tr>
<tr>
<td>5</td>
<td>Moderately Tired; Let Down</td>
</tr>
<tr>
<td>6</td>
<td>Extremely Tired; Very Difficult to Concentrate</td>
</tr>
<tr>
<td>7</td>
<td>Completely Exhausted; Unable to Function Effectively; Ready to Drop</td>
</tr>
<tr>
<td>COMMENTS</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NAME</th>
<th>DATE AND TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>WORKLOAD ESTIMATE</td>
<td></td>
</tr>
<tr>
<td>(Circle the number of the statement which best describes the MAXIMUM workload you experienced during the past work period. Put an X over the number of the statement which best describes the AVERAGE workload you experienced during the past work period.)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Nothing to do; No System Demands</td>
</tr>
<tr>
<td>2</td>
<td>Little to do; Minimum System Demands</td>
</tr>
<tr>
<td>3</td>
<td>Active Involvement Required, But Easy to Keep Up</td>
</tr>
<tr>
<td>4</td>
<td>Challenging, But Manageable</td>
</tr>
<tr>
<td>5</td>
<td>Extremely Busy; Barely Able to Keep Up</td>
</tr>
<tr>
<td>6</td>
<td>Too Much to do; Overloaded; Postponing Some Tasks</td>
</tr>
<tr>
<td>7</td>
<td>Unmanageable; Potentially Dangerous; Unacceptable</td>
</tr>
<tr>
<td>COMMENTS</td>
<td></td>
</tr>
</tbody>
</table>

**Figure A4: Crew Status Survey**

*Data requirements* - Although the Crew Status Survey is printed on card stock, subjects find it difficult to fill in the rating scale during high workload periods. Further, sorting (for example, by the time completed) the completed card-stock ratings after the flight is also difficult and not error free. A larger character-size version of the survey has been included on flight cards at the Air Force Flight Test Center. Verbal ratings prompted by the experimenter work well if: (1) subjects can quickly scan a card-stock copy of the rating, scale to verify the meaning of a rating and (2) subjects are not performing a conflicting verbal task. Each scale can be used independently.

*Thresholds* - 1 to 7 for subjective fatigue; 1 to 7 for workload (see Figure A4).
Dynamic Workload Scale

General description - The Dynamic Workload Scale is a seven-point workload scale (see Figure A5) developed as a tool for aircraft certification. It has been used extensively by Airbus Industries.

<table>
<thead>
<tr>
<th>Workload Assessment</th>
<th>Criteria</th>
<th>Effort or Stress</th>
<th>Appreciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light</td>
<td>Reserve Capacity: Ample</td>
<td></td>
<td>Very Acceptable</td>
</tr>
<tr>
<td>Moderate</td>
<td>Adequate</td>
<td>Some</td>
<td>Well Acceptable</td>
</tr>
<tr>
<td>Fair</td>
<td>Sufficient</td>
<td>Recurring</td>
<td>Acceptable</td>
</tr>
<tr>
<td>High</td>
<td>Reduced</td>
<td>Repetitive</td>
<td>High but Acceptable</td>
</tr>
<tr>
<td>Heavy</td>
<td>Little</td>
<td>Frequent</td>
<td>Just Acceptable</td>
</tr>
<tr>
<td>Extreme</td>
<td>None</td>
<td>Continuous</td>
<td>Not Acceptable CONTINUOUSLY</td>
</tr>
<tr>
<td>Supreme</td>
<td>Impairment</td>
<td>Impairment</td>
<td>Not Acceptable Instantaneously</td>
</tr>
</tbody>
</table>

Figure A5: Dynamic Workload Scale

Strengths and limitations - Speyer, Fort, Fouillot, and Bloomberg (1987) reported high concordance between pilot and observer ratings as well as sensitivity to workload increases.

Data requirements - Dynamic Workload Scale ratings must be given by both a pilot and an observer-pilot. The pilot is cued to make a rating; the observer gives a rating whenever workload changes or five minutes have passed.

Thresholds - Two is minimum workload; eight, maximum workload.
Equal-Appearing Intervals

*General description* - Subjects rate the workload in one of several categories using the assumption that each category is equidistant from adjacent categories.

*Strengths and limitations* - Hicks and Wierwille (1979) reported sensitivity to task difficulty in a driving simulator. Masline (1986) reported comparable results with the magnitude estimates and SWAT ratings but greater ease of administration. Masline, however, warned of rater bias.

*Data requirements* - Equal intervals must be clearly defined.

*Thresholds* - Not stated.
Flight Workload Questionnaire

General description - The Flight Workload Questionnaire is a four-item, behaviorally anchored rating scale. The items and the end points of the rating scales are: workload category (low to very high), fraction of time busy (seldom have much to do to fully occupied at all times), how hard had to think (minimal thinking to a great deal of thinking), and how felt (relaxing to very stressful).

Strengths and limitations - The questionnaire is sensitive to differences in experience and ability. For example, Stein (1984) found significant differences in the flight workload ratings between experienced and novice pilots. Specifically, experienced pilots rated their workload during an air transport flight lower than novice pilots did. However, Stein also found great redundancy in the value of the ratings given for the four questionnaire items. This suggests that the questionnaire may evoke a response bias. The questionnaire provides a measure of overall workload but cannot differentiate between flight segments and/or events.

Data requirements - Not stated.

Thresholds - Not stated.
Hart and Hauser Rating Scale

*General description* - Hart and Hauser (1987) used a six-item rating scale (see Figure A6) to measure workload during a nine-hour flight. The items and their scales were: stress (completely relaxed to extremely tense), mental/sensory effort (very low to very high), fatigue (wide awake to worn out), time pressure (none to very rushed), overall workload (very low to very high), and performance (completely unsatisfactory to completely satisfactory). Subjects were instructed to mark the scale position that represented their experience.

![Stress Scale](image)

**Figure A6: Hart and Hauser Rating Scale**

*Strengths and limitations* - The scale was developed for use inflight. In the initial study, Hart and Hauser (1987) asked subjects to complete the questionnaire at the end of each of seven flight segments. They reported significant segment effects in the seven-hour flight. Specifically, stress, mental/sensory effort, and time pressure were lowest during a data recording segment. There was a sharp increase in rated fatigue after the start of the data-recording segment. Overall workload was rated as higher by the aircraft commander than by the copilot. Finally, performance received the same ratings throughout the flight.

*Data requirements* - The scale is simple to use but requires a stiff writing surface and minimal turbulence.

*Thresholds* - Not stated.
Honeywell Cooper-Harper Rating Scale

*General description* - This rating scale (see Figure A7) uses a decision-tree structure for assessing, overall task workload.

*Strengths and limitations* - The Honeywell Cooper-Harper Rating Scale was developed by Wolf (1978) to assess overall task workload. North, Stackhouse, and Graffunder (1979) used the scale to assess workload associated with various Vertical Take-Off and Landing (VTOL) aircraft displays. For the small subset of conditions analyzed, the scale ratings correlated well with performance.

*Data requirements* - Subjects must answer three questions related to task performance. The ratings are ordinal and must be treated as such in subsequent analyses.

*Thresholds* - Minimum is 1, maximum is 9.

![Decision Tree Diagram]

Figure A7: Honeywell Cooper-Harper Rating Scale
Magnitude Estimation

*General description* - Subjects are required to estimate workload numerically in relation to a standard.

*Strengths and limitations* - Borg (1978) successfully used this method for evaluating workload. Helm and Heimstra (1981) reported a high correlation between workload estimates and task difficulty. Masline (1986) reported sensitivity comparable to estimates from the equal-appearing intervals method and SWAT. Gopher and Braune (1984), however, found a low correlation between workload estimates and reaction-time performance. In contrast, Kramer, Sirevaag, and Braune (1987) reported good correspondence to performance in a fixed-based flight simulator. Hart and Staveland (1988) suggest that the presence of a standard enhances interrater reliability. O'Donnell and Eggemeier (1986), however, warned that subjects may be unable to retain an accurate memory of the standard over the course of an experiment.

*Data requirements* - A standard must be well defined.

*Thresholds* - Not stated.
Modified Cooper-Harper Rating Scale

General description - Wierwille and Casali (1983) noted that the Cooper-Harper scale represented a combined handling-qualities/workload rating scale. They found that it was sensitive to psychomotor demands on an operator, especially for aircraft handling qualities. They wanted to develop an equally useful scale for the estimation of workload associated with cognitive functions, such as "perception, monitoring, evaluation, communications, and problem solving." The Cooper-Harper scale terminology was not suited to this purpose. A modified Cooper-Harper rating scale (see Figure A8) was developed to "increase the range of applicability to situations commonly found in modern systems." Modifications included: (1) changing the rating scale end points to very easy and impossible, (2) asking the pilot to rate mental workload level rather than controllability, and (3) emphasizing difficulty rather than deficiencies. In addition, Wierwille and Casali (1983) defined mental effort as "minimal" in rating 1, while mental effort is not defined as minimal until rating 3 in the original Cooper-Harper scale. Further, adequate performance begins at rating 3 in the modified Cooper-Harper but at rating 5 in the original scale.

![Modified Cooper-Harper Rating Scale](image)

Figure A8: Modified Cooper-Harper Rating Scale

Strengths and limitations - Investigations were conducted to assess the modified Cooper-Harper scale. They focused on perception (e.g., aircraft engine instruments out of limits during simulated
flight), cognition (e.g., arithmetic problem solving during simulated flight), and communications (e.g., detection of, comprehension of, and response to own aircraft call sign during simulated flight).

The modified Cooper-Harper is sensitive to various types of workloads. For example, Casali and Wierwille (1983) reported that modified Cooper-Harper ratings increased as the communication load increased. Wierwille, Rahimi, and Casali (1985) reported significant increase in workload as navigation load increased. Casali and Wierwille (1984) reported significant increases in ratings as the number of danger conditions increased. Skipper, Rieger, and Wierwille (1986) reported significant increases in ratings in both high communication and high navigation loads. Wolf (1978) reported the highest workload ratings in the highest workload flight condition (i.e., high wind gust, and poor handling qualities). Bittner, Byers, Hill, Zaklad, and Christ (1989) reported reliable differences between mission segments in a mobile air defense system. Byers, Bittner, Hill, Zaklad, and Christ (1988) reported reliable differences between crew positions in a remotely-piloted vehicle system. These results suggested that the modified Cooper-Harper scale is a valid, statistically reliable indicator of overall mental workload. However, it carries with it the underlying assumptions that high workload is the only determinant of the need for changing the control/display configuration. Wierwille, Casali, Connors, and Rahimi (1985) concluded that the modified Cooper-Harper Rating Scale provided consistent and sensitive ratings of workload across a range of tasks. Wierwille, Skipper, and Rieger (1985) reported the best consistency and sensitivity with the modified Cooper-Harper from five alternatives tests. Warr, Colle, and Reid (1986) reported that the modified Cooper-Harper Ratings were as sensitive to task difficulty as SWAT ratings. Kilmer, Knapp, Burdsal, Borresen, Bateman, and Malzahn (1988), however, reported that the modified Cooper-Harper rating scale was less sensitive than SWAT ratings to changes in tracking, task difficulties. Hill, Iavecchia, Byers, Bittner, Zaklad, and Christ (1992) reported that the modified Cooper-Harper scale was not as sensitive or as operator accepted as the NASA TLX or the overall workload scale.

Data requirements - Wierwille and Casali (1983) recommend the use of the modified Cooper-Harper in experiments where overall mental workload is to be assessed. They emphasize the importance of proper instructions to the subjects. Since the scale was designed for use in experimental situations, it may not be appropriate to situations requiring an absolute diagnosis of a subsystem.

Thresholds - Not stated.
NASA Bipolar Rating Scale

General description - The NASA Bipolar Rating Scale has ten subscales. The titles, endpoints, and descriptions of each scale are presented in Table A2; the scale itself, in Figure A9. If a scale is not relevant to a task, it is given a weight of zero (Hart, Battiste, and Lester, 1984). A weighting procedure is used to enhance intrasubject reliability by 50 percent (Miller and Hart, 1984).

Strengths and limitations - The scale is sensitive to flight difficulty. For example, Bortolussi, Kantowitz, and Hart (1986) reported significant differences in the bipolar ratings between an easy and a difficult flight scenario. Bortolussi, Hart, and Shively (1987) and Kantowitz, Hart, Bortolussi, Shively, and Kantowitz (1984) reported similar results. However, Haworth, Bivens, and Shively (1986) reported that, although the scale discriminated control configurations in a single-pilot configuration, it did not do so in a pilot/copilot configuration. Biferno (1985) reported a correlation between workload and fatigue ratings for a laboratory study. Bortolussi, Kantowitz, and Hart (1986) and Bortolussi, Hart, and Shively (1987) reported that the bipolar scales discriminated two levels of difficulty in a motion-based simulator task. Vidulich and Pandit (1986) reported that the bipolar scales discriminated levels of training in a category search task. Haworth, Bivens, and Shively (1986) reported correlations of 0.79 with Cooper-Harper ratings and 0.67 with SWAT ratings in a helicopter nap-of-the-earth mission. Vidulich and Tsang, (1985a, 1985b, 1985c) reported that the NASA Bipolar Scales were sensitive to task demand, had higher interrater reliability than SWAT, and required less time to complete than SWAT.

Table A2: NASA Bipolar Rating-Scale Definitions

<table>
<thead>
<tr>
<th>TITLE</th>
<th>ENDPOINTS</th>
<th>DESCRIPTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Workload</td>
<td>low/high</td>
<td>The total workload associated with the task considering all sources and components.</td>
</tr>
<tr>
<td>Task Difficulty</td>
<td>low/high</td>
<td>Whether the task was easy, demanding, simple or complex, exacting or forgiving.</td>
</tr>
<tr>
<td>Time Pressure</td>
<td>low/high</td>
<td>The amount of pressure you felt due to the rate at which the task elements occurred. Was the task slow and leisurely or rapid and frantic.</td>
</tr>
<tr>
<td>Performance</td>
<td>good/poor</td>
<td>How successful you think you were in doing what we asked you to do and how satisfied you were with what you accomplished.</td>
</tr>
<tr>
<td>Mental/Sensory Effort</td>
<td>low/high</td>
<td>The amount of mental and/or perceptual activity that was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.).</td>
</tr>
<tr>
<td>Physical Effort</td>
<td>low/high</td>
<td>The amount of physical activity that was required (e.g., pushing, pulling, turning, controlling, activating, etc.).</td>
</tr>
<tr>
<td>Frustration Level</td>
<td>Fulfilled, Exasperated</td>
<td>How insecure, discourage, irritated, and annoyed versus secure, gratified, content and complacent you felt.</td>
</tr>
<tr>
<td>Stress Level</td>
<td>Relaxed, Tense</td>
<td>How anxious, worried, uptight, and harassed or calm, tranquil, placid, and relaxed you felt.</td>
</tr>
<tr>
<td>Fatigue</td>
<td>Exhausted, Alert</td>
<td>How tired, weary, worn out, and exhausted or fresh, vigorous, and energetic you felt.</td>
</tr>
<tr>
<td>Activity Type</td>
<td>Skill based, Rule based, Knowledge based</td>
<td>The degree to which the task required mindless reaction to well-learned routines or required the application of known rules or required problem solving and decision making.</td>
</tr>
</tbody>
</table>
**Figure A9: NASA Bipolar Rating Scale**

*Data requirements* - The number of times a dimension is selected by a subject is used to weight each scale. These weights are then multiplied by the scale score, summed, and divided by the total weight to obtain a workload score. The minimum workload value is zero; the maximum, 100. The scale provides a measure of overall workload but is not sensitive to short-term demands. Further, the activity-type dimension must be carefully explained to pilots before use in flight.

*Thresholds* - Not stated.
NASA Task Load Index

General description - The NASA Task Load Index (TLX) is a multi-dimensional subjective workload rating technique (see Figure A10). In TLX, workload is defined as the "cost incurred by human operators to achieve a specific level of performance." The subjective experience of workload is defined as an integration of weighted subjective responses (emotional, cognitive, and physical) and weighted evaluation of behaviors. The behaviors and subjective responses, in turn, are driven by perceptions of task demand. Task demands can be objectively quantified in terms of magnitude and importance. An experimentally based process of elimination led to the identification of six dimensions for the subjective experience of workload: mental demand, physical demand, temporal demand, perceived performance, effort, and frustration level. The rating-scale definitions are presented in Table A3.

Strengths and limitations - Sixteen investigations were carried out, establishing a database of 3461 entries from 247 subjects. All dimensions were rated on bipolar scales ranging from 1 to 100, anchored at each end with a single adjective. An overall workload rating was determined from a weighted combination of scores on the six dimensions. The weights were determined from a set of relevance ratings provided by the subjects.

Table A3: Rating-Scale Definitions

<table>
<thead>
<tr>
<th>TITLE</th>
<th>ENDPOINTS</th>
<th>DESCRIPTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental Demand</td>
<td>low/high</td>
<td>How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?</td>
</tr>
<tr>
<td>Physical Demand</td>
<td>low/high</td>
<td>How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?</td>
</tr>
<tr>
<td>Temporal Demand</td>
<td>low/high</td>
<td>How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?</td>
</tr>
<tr>
<td>Performance</td>
<td>good/poor</td>
<td>How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?</td>
</tr>
<tr>
<td>Effort</td>
<td>low/high</td>
<td>How hard did you have to work (mentally and physically) to accomplish your level of performance?</td>
</tr>
<tr>
<td>Frustration Level</td>
<td>low/high</td>
<td>How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task? (NASA Task Load Index, p. 13)</td>
</tr>
</tbody>
</table>

Many of the conclusions drawn by Hart and Staveland (1988) were based on correlations. However, the distributions of subjective responses to many tasks were skewed. Hart and Staveland did not indicate whether the distributions were normalized prior to the analyses, suggesting a lack of homoscedasticity (a normal distribution around a regression line) for many of the correlations calculated. Additionally, multicollinearity among ratings was not controlled by the calculation of semi-partial correlation values. These deficiencies call into question conclusions based on the correlations.
Figure A10: NASA TLX Rating Sheet
On the other hand, there was at least one striking similarity between TLX data and the structure of SWAT, another multi-dimensional technique. When ten dimensions were under consideration during the process of elimination, the four dimensions considered most important by the subjects paralleled the three dimensions selected for use in SWAT: mental effort, time pressure, and psychological stress. The four TLX dimensions were, in order of importance, time pressure, frustration, stress, and mental effort. Frustration may be thought of as highly related to both time pressure and psychological stress. The similarity between the two approaches may support the hypothesis that perceptions of workload are, indeed, highly dependent on perceptions of time pressure, mental effort, and psychological stress. An attempt to compare TLX to SWAT by Hart and Staveland (1988) was marred by the failure to use partializing techniques to compute interrelationships of the Subjective Workload Assessment Technique among the dimensions. Vidulich and Tsang (1985) compared the SWAT and TLX. They stated that the collection of ratings is simpler with SWAT. However, the SWAT card sort is more tedious and time consuming.

Hart and Staveland (1988) concluded that the TLX provides a sensitive indicator of overall workload as it differed among tasks of various cognitive and physical demands. They also stated that the weights and magnitudes determined for each TLX dimension provide important diagnostic information about the sources of loading within a task. They reported that the six TLX ratings took less than a minute to acquire and suggested the scale would be useful in operational environments. Battiste and Bortolussi (1988) reported significant workload effects as well as a test-retest correlation of + 0.769. Corwin, Sandry-Garra, Biferno, Boucek, Logan, Jonsson, and Metalis (1989) reported that NASA TLX was a valid and reliable measure of workload. Bittner, Byers, Hill, Zaklad, and Christ (1989), Byers, Bittner, Hill, Zaklad, and Christ (1988), Hill, Byers, Zaklad, and Christ (1989), Hill, Zaklad, Bittner, Byers, and Christ (1988), and Shively, Battiste, Matsumoto, Pepitone, Bortolussi, and Hart (1987), based on in-flight data, stated that TLX ratings significantly discriminated flight segments. Vidulich and Bortolussi (1988b) replicated the significant flight-segment effect but reported no significant differences in TLX ratings between control configurations, nor between combat countermeasure conditions. Tsang and Johnson (1987) reported good correlations between NASA TLX and a uni-dimensional workload scale. In a later study, these authors (Tsang and Johnson, 1989) reported reliable increases in NASA TLX ratings when target acquisition and engine-failure tasks were added to the primary flight task. Battiste and Bortolussi (1988) reported no significant correlation between SWAT and NASA TLX in a simulated B-727 flight. Vidulich and Tsang, (1987) replicated the Tsang, and Johnson finding as well as reported a good correlation between NASA TLX and the Analytical Hierarchy Process. In the same year, Nataupsky and Abbott (1987) successfully applied NASA TLX to a multi-task environment. Finally, Hill, Iavecchia, Byers, Bittner, Zaklad, and Christ (1992) reported that the NASA TLX was sensitive to different levels of workload and high in user acceptance. Their subjects were Army operators. Nygren (1991) reported that NASA TLX is a measure of general workload experienced by aircrews.

Data requirements - Use of the TLX requires two steps. First, subjects rate each task performed on each of the six subscales. Hart suggests that subjects should practice using the rating scales in a training session. Second, subjects must perform 15 pair-wise comparisons of six workload scales. The number of times each scale is rated as contributing more to the workload of a task is
used as the weight for that scale. Separate weights should be derived for diverse tasks; the same
weights can be used for similar tasks. Note that a set of IBM PC compatible programs has been
written to gather ratings and weights and to compute the weighted workload scores. The
programs are available from the Human Factors Division at NASA Ames Research Center,
Moffett Field, CA.

Thresholds - Not stated.

A-23
Overall Workload Scale

*General description* - The Overall Workload (OW) Scale is a bipolar scale requiring subjects to provide a single workload rating.

*Strengths and limitations* - The OW scale is easy to use but is less valid and reliable than NASA TLX or AHP ratings (Vidulich and Tsang, 1987). Hill, Iavecchia, Byers, Bittner, Zaklad, and Christ (1992) reported that OW was consistently more sensitive to workload and had greater operator acceptance than the Modified Cooper-Harper rating scale or the Subjective Workload Assessment Technique (SWAT).

*Data requirements* - Not stated.

*Thresholds* - Not stated.
Pilot Objective/Subjective Workload Assessment Technique

*General description* - The Pilot Objective/Subjective Workload Assessment Technique (POSWAT) is a ten-point subjective scale developed at the Federal Aviation Administration's Technical Center (Stein, 1984). The scale is a modified Cooper-Harper scale but does not include the binary decision tree that is characteristic of the Cooper-Harper scale. It does, however, divide workload into five categories: low, minimal, moderate, considerable, and excessive. Like the Cooper-Harper, the lowest three levels (1 through 3) are grouped into a low category.

*Strengths and limitations* - Stein (1984) reported that POSWAT ratings significantly differentiated experienced and novice pilots and high (initial and final approach) and low (en route) flight segments. There was also a significant learning effect: workload ratings were significantly higher on the first than on the second flight. Although the POSWAT scale was sensitive to manipulations of pilot experience level for flights in a light aircraft and in a simulator (Mallery and Maresh, 1987), the scale was cumbersome. Seven dimensions (workload, communications, control inputs, planning, "deviations," error, and pilot complement) are combined on one scale. Further, the number of ranks on the ordinal scale are confusing since there are both five and ten levels. In the Mallery and Maresh (1987) study, POSWAT ratings were obtained once per minute during simulated and actual flights. This high rate of data acquisition was also used by Rosenberg, Rehmann and Stein (1982). The latter investigators found that pilots reliably reported workload differences in a tracking task on a simple ten-point non-adjetival scale. Therefore, the cumbersome structure of the POSWAT scale may not be necessary.

*Data requirements* - Stein (1984) suggested not analyzing POSWAT ratings for short flight segments if the ratings are given at one-minute intervals.

*Thresholds* - Not stated.
Pilot Subjective Evaluation

*General description* - The Pilot Subjective Evaluation (PSE) workload scale (see Figure A11) was developed by Boeing for use in the certification of the Boeing 767 aircraft. The scale is accompanied by a questionnaire. Both the scale and the questionnaire are completed with reference to an existing aircraft selected by the pilot.

*Strengths and limitations* - Fadden (1982) and Ruggerio and Fadden (1987) stated that the ratings of workload greater than the reference aircraft were useful in identifying aircraft design deficiencies.

*Data requirements* - Both the PSE scale and the questionnaire must be completed by each subject.

*Thresholds* - 1, minimum workload; 7, maximum workload.

![Pilot Subjective Evaluation Scale](image)

Figure A11: Pilot Subjective Evaluation Scale
Subjective Workload Assessment Technique

*General description* - The Subjective Workload Assessment Technique (SWAT) combines ratings of three different scales (see Table A4) to produce an interval scale of mental workload. These scales are: (1) time load, which reflects the amount of spare time available in planning, executing, and monitoring a task; (2) mental effort load, which assesses how much conscious mental effort and planning are required to perform a task; and (3) psychological stress load, which measures the amounts of risk, confusion, frustration, and anxiety associated with task performance. A more complete description is given in Reid and Nygren (1988). A description of the initial conjoint measurement model for SWAT is described in Nygren (1982, 1983).

**Table A4: SWAT Scales**

<table>
<thead>
<tr>
<th>Time Load</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Often have spare time. Interruptions or overlap among activities occur infrequently or not at all.</td>
</tr>
<tr>
<td>2.</td>
<td>Occasionally have spare time. Interruptions or overlap among activities occur frequently.</td>
</tr>
<tr>
<td>3.</td>
<td>Almost never have spare time. Interruptions or overlap among activities are frequent or occur all the time.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mental Effort Load</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention.</td>
</tr>
<tr>
<td>2.</td>
<td>Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention required.</td>
</tr>
<tr>
<td>3.</td>
<td>Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Psychological Stress Load</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Little confusion, risk, frustration, or anxiety exists and can be easily accommodated.</td>
</tr>
<tr>
<td>2.</td>
<td>Moderate stress due to confusion, frustration, or anxiety noticeably adds to workload. Significant compensation is required to maintain adequate performance.</td>
</tr>
<tr>
<td>3.</td>
<td>High to very intense stress due to confusion, frustration, or anxiety. High to extreme determination and self-control required (Potter and Bressler, 1989, pp. 12-14).</td>
</tr>
</tbody>
</table>

*Strengths and limitations* - SWAT has been found to be a valid (Albery, Repperger, Reid, Goodyear, and Roe, 1987; Haworth, Bivens, and Shively, 1986; Masline, 1986; Reid, Shingledecker, and Eggemeier, 1981; Reid, Shingledecker, Nygren, and Eggemeier, 1981; Vidulich and Tsang, 1987; Warr, Colle, and Reid, 1986), sensitive (Eggemeier, Crabtree, Zingg, Reid and Shingledecker, 1982), reliable (Corwin, Sandry-Garza, Biferno, Boucek, Logan, Jonsson, and Metalis, 1989; Gidcomb, 1985), and relatively unobtrusive (Crabtree, Bateman, and Acton, 1984; Courtright and Kuperman, 1984; Eggemeier, 1988) measure of workload. Further, SWAT ratings are not affected by delays of up to 30 minutes (Eggemeier, Crabtree, and LaPointe, 1983), nor by intervening tasks of all but difficult tasks (Eggemeier, Melville, and Crabtree, 1984; Lutmer and Eggemeier, 1990). Warr (1986) reported that SWAT ratings were less variable than modified Cooper-Harper ratings.
SWAT has been used in diverse environments, for example, a high-G centrifuge (Albery, Ward, and Gill, 1985), command, control, and communications centers (Crabtree, Bateman, and Acton, 1984), nuclear power plants (Beare and Dorris, 1984), domed flight simulators (Reid, Eggemeier, and Shingledecker, 1982), tank simulators (Whitaker, Peters, and Garinther, 1989); and the benign laboratory setting (Graham and Cook, 1984; Kilmer, Knapp, Burdual, Borrensen, Bateman, and Malzahn (1988)). In the laboratory, SWAT has been used to assess the workload associated with critical tracking and communication tasks (Reid, Shingledecker, and Eggemeier, 1981), memory tasks (Eggemeier, Crabtree, Zingg, Reid, and Shingledecker, 1982; Eggemeier and Stadler, 1984; Potter and Acton, 1985), and monitoring tasks (Noteastine, 1984).

Usage in simulated flight has also been extensive (Haworth, Bivens, and Shively, 1986; Nataupsky and Abbott, 1987; Schick and Hann, 1987; Skelly and Purvis, 1985; Skelly, Reid, and Wilson, 1983; Thiessen, Lay, and Stern, 1986; Ward and Hassoun, 1990). For example, Bateman and Thompson (1986) reported that SWAT ratings increased as task difficulty increased. Their data were collected in an aircraft simulator during a tactical mission. Vickroy (1988), also using an aircraft simulator, reported that SWAT ratings increased as the amount of air turbulence increased.

Usage in flight has been extensive. For example, Pollack (1985) used SWAT to assess differences in workload between flight segments. She reported that C-130 pilots had the highest SWAT scores during the approach segment of the mission. She also reported higher SWAT ratings during the preflight segments of tactical, rather than proficiency, missions. Haskell and Reid (1987) found significant difference in SWAT ratings between flight maneuvers and also between successfully completed maneuvers and those that were not successfully completed. Gawron, et al. (1988) analyzed SWAT ratings made by the pilot and copilot four times during each familiarization and data flight: (1) during the taxi out to the runway, (2) just prior to a simulated drop, (3) just after a simulated drop, and (4) during the taxi back to the hangar. There were significant segments effects. Specifically, SWAT ratings were highest before the drop and lowest for preflight. The ratings during postdrop and postflight were both moderate.

In addition, ratings of the time, effort, and stress scales may be individually examined as workload components (Eggemeier, McGhee and Reed, 1983). Finally, Eggleston (1984) found a significant correlation between projected SWAT ratings made during system concept evaluation and those made during ground-based simulation of the same system. Nygren (1991) stated that SWAT provides a good cognitive model of workload, sensitive to individual differences.

Experience with SWAT has not been all positive, however. For example, Boyd (1983) reported that there were significant positive correlations between the three workload scales in a text-editing task. This suggests that the three dimensions of workload are not independent. This, in turn, poses a problem for use of conjoint measurement techniques. Derrick (1988) and Hart (1986) suggest that three scales may not be adequate for assessing workload. Further, experience at the Air Force Flight Test Center at Edwards Air Force Base with SWAT suggests that task demands during flight tests often preclude the acquisition of multiple ratings. Battiste and Bortolussi (1988) reported a test/retest correlation of +0.751 but also stated that, of the 144 SWAT ratings reported during, a simulated B-727 flight, 59 were zero. Corwin (1989) reported no difference
between inflight and postflight ratings of SWAT in only two of three flight conditions. Kilmer, et al. (1988) reported that SWAT was more sensitive to changes in difficulty of a tracking task than the modified Cooper-Harper Rating Scale was. Gidcomb (1985) reported casual card sorts and urged emphasizing the importance of the card sort to SWAT raters. A computerized version of the traditional card sort is being developed at the Air Force School of Aerospace Medicine. This version eliminates the tedium and dramatically reduces the time to complete the SWAT card sort. Haworth, Bivens, and Shively (1986) reported that, although the SWAT was able to discriminate control configuration conditions in a single-pilot configuration, it could not discriminate these same conditions in a pilot/copilot configuration. Wilson, Hughes, and Hassoun (1990) reported no significant differences in SWAT ratings among display formats, in contrast to pilot comments. Van de Graaff (1987) reported considerable (60 points) intersubject variability in SWAT ratings during an in-flight approach task. Hill, Iavecchia, Byers, Bittner, Zaklad, and Christ (1992) reported that SWAT was not as sensitive to workload or as accepted by Army operators as NASA TLX and the Overall Workload Scale.

Data requirements - SWAT requires two steps to use: scale development and event scoring. Scale development requires subjects to rank, from lowest to highest workload, 27 combinations of three levels of the three workload subscales. The levels of each subscale are presented in Table A4. Programs to calculate the SWAT score for every combination of ratings on the three subscales are available from the Harry G. Armstrong Aerospace Medical Research Laboratory at Wright-Patterson Air Force Base. A user's manual is also available from the same source.

During event scoring, the subject is asked to provide a rating (1, 2, -) for each subscale. The experimenter then maps the set of ratings to the SWAT score (1 to 100) calculated during the scale development step. Reid (1987) suggests that the tasks to be rated be meaningful to the subjects and, further, that the ratings not interfere with performance of the task. Acton and Colle (1984) reported that the order in which the subscale ratings are presented does not affect the SWAT score. However, it is suggested that the order remain constant to minimize confusion.

Thresholds - Minimum value is 0, maximum value is 100. High workload is associated with the maximum value.
Subjective Workload Dominance Technique

*General description* - The Subjective Workload Dominance (SWORD) technique uses judgment matrices to assess workload.

*Strengths and limitations* - SWORD is a sensitive and reliable workload measure (Vidulich, 1989).

*Data requirements* - There are three required steps: (1) a rating scale listing all possible pairwise comparisons of the tasks performed must be completed, (2) a judgment matrix comparing each task to every other task must be filled in with each subject's evaluation of the tasks, and (3) ratings must be calculated using a geometric means approach.

*Thresholds* - Not stated.
Workload/Compensation/Interference/Technical Effectiveness

General description - The Workload/Compensation/Interference/Technical Effectiveness (WCI/TE) rating scale (see Figure A12) requires subjects to rank the sixteen matrix cells and then rate specific tasks. The ratings are converted by conjoint scaling, techniques to values of 0 to 100.

Strengths and limitations - Wierwille and Connor (1983) reported sensitivity of WCI/TE ratings to three levels of task difficulty in a simulated flight task. Wierwille, Casali, Connor, and Rahimi (1985) reported sensitivity to changes in difficulty in psychomotor, perceptual, and mediational tasks. O'Donnell and Eggemeier (1986) suggest that the WCI/TE should not be used as a direct measure of workload.

![Figure A12: The WCI/TE Scale Matrix](image)

Data requirements - Subjects must rank the sixteen matrix cells and then rate specific tasks. Complex mathematical processing is required to convert the ratings to WCI/TE values.

Thresholds - 0 is minimum workload, 100 is maximum workload.
Crew Situational Awareness

*General description* - Mosier and Chidester (1991) developed a method for measuring situational awareness of air transport crews. Expert observers rate crew coordination performance and identify and rate performance errors (type 1, minor errors; type 2, moderately severe errors; and type 3, major, operationally significant errors). The experts then develop information transfer matrices identifying time and source of item requests (prompts) and verbalized responses. Information is then classified into decision or non-decision information.

*Strengths and limitations* - The method was sensitive to type of errors and decision prompts.

*Data requirements* - The method requires open and frequent communication among air crew members. It also requires a team of expert observers to develop the information transfer matrices.

*Thresholds* - Not stated.
Situational Awareness Global Assessment Technique

**General description** - The most well-known objective measure of SA is the Situational Awareness Global Assessment Technique (SAGAT) (Endsley, 1988a). SAGAT was designed around real-time, human-in-the-loop simulation of a military cockpit but could be generalized to other systems. Using SAGAT, the simulation is stopped at random times, and the operators are asked questions to determine their SA at that particular point in time. Subjects' answers are compared with the correct answers that have been simultaneously collected in the computer database. "The comparison of the real and perceived situation provides an objective measure of ... SA" (Endsley, 1988b, p. 101). This same technique could be used with any complex system that is simulated, be it a nuclear power plant control room or the engine room of a ship. In addition, if an operational system is properly instrumented, SAGAT is also applicable in this environment. SAGAT uses a graphical computer program for the rapid presentation of queries and data collection. In addition to possessing a high degree of face validity, the SAGAT technique has been tested in several studies, which demonstrated: (1) empirical validity (Endsley, 1989, 1990a) - the technique of freezing the simulation did not impact subject performance and subjects were able to reliably report SA knowledge for up to six minutes after a freeze without memory decay problems; (2) predictive validity (Endsley, 1990b) - linking SAGAT scores to subject performance; and (3) content validity (Endsley, 1990a) - showing appropriateness of the queries used (for an air-to-air fighter cockpit).

**Strengths and limitations** - SAGAT provides unbiased objective measures of SA across all of the operators' SA requirements that can be computed in terms of errors or percent correct and can be treated. However, Sarter and Woods (1991) suggest that SAGAT does not measure SA but rather measures what pilots can recall. Further, Fraker and Vidulich (1991) identified two major problems with the use of explicit measures of SA, such as SAGAT: (1) decay of information and (2) inaccurate beliefs.

**Data requirements** - The proper queries must be identified prior to the start of the experiment.

**Thresholds** - Tolerance limits for acceptable deviance of perceptions from real values on each parameter should be identified prior to the start of the experiment.
Situational Awareness Rating Technique

General description - An example of a subjective measure of SA is the Situational Awareness Rating Technique (SART) (Taylor, 1990). SART is a questionnaire method that concentrates on measuring the operator's knowledge in three areas: (1) demands on attentional resources, (2) supply of attentional resources, and (3) understanding of the situation (see Table A5). The reason that SART measures three different components (there is also a 10-dimensional version) is that the SART developers feel that, like workload, SA is a complex construct; therefore, to measure SA in all its aspects, separate measurement dimensions are required. Because information processing and decision making are inextricably bound with SA (since SA involves primarily cognitive rather than physical workload), SART has been tested in the context of Rasmussen's Model of skill-, rule-, and knowledge-based behavior. Selcon and Taylor (1989) conducted separated studies looking at the relationship between SART and rule- and knowledge-based decisions, respectively. The results showed that SART ratings appear to provide diagnosticity in that they were significantly related to performance measures of the two types of decision making. Early indications are that SART is tapping the essential qualities of SA, but further validation studies are required before this technique is commonly used.

Table A5: Definitions of SART Rating Scales

<table>
<thead>
<tr>
<th>Demand on Attentional Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instability: Likelihood of situation changing suddenly.</td>
</tr>
<tr>
<td>Complexity: Degree of complication of situation.</td>
</tr>
<tr>
<td>Variability: Number of variables changing in situation.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Supply of Attentional Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arousal: Degree of readiness for activity.</td>
</tr>
<tr>
<td>Concentration: Degree to which thoughts bear on situation.</td>
</tr>
<tr>
<td>Division: Amount of division of attention in situation.</td>
</tr>
<tr>
<td>Spare Capacity: Amount of attention left to spare for new variables.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Understanding of the Situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Quantity: Amount of information received and understood.</td>
</tr>
<tr>
<td>Information Quality: Degree of goodness of information gained.</td>
</tr>
</tbody>
</table>

Strengths and limitations - SART is a subjective measure and, as such, suffers from the inherent reliability problems of all subjective measures. The strengths are that SART is easily administered and was developed in three logical phases: (1) scenario generation, (2) construct elicitation, and (3) construct structure validation (Taylor, 1989). SART has been prescribed for comparative system design evaluation (Taylor and Selcon, 1991). SART is sensitive to differences in performance of aircraft attitude recovery tasks and learning comprehension tasks (Selcon and Taylor, 1991; Taylor and Selcon, 1990). However, Taylor and Selcon (1991) state "There

A-34
remains considerable scope for scales development, through description improvement, interval justification and the use of conjoint scaling techniques to condense multi-dimensional ratings into a single SA score" (p. 11). These authors further state that "The diagnostic utility of the Attentional Supply constructs has yet to be convincingly demonstrated" (p. 12).

Data requirements - Data are on an ordinal scale; interval or ratio properties cannot be implied.

Thresholds - The data are on an ordinal scale and must be treated accordingly when statistical analysis is applied to the data. Non-parametric statistics may be the most appropriate analysis method.
Bibliography


Crabtree, M. S. (1975). Human factors evaluation of several control system configurations, including workload sharing with force wheel steering during approach and flare (AFFDL-TR-75-43). Wright-Patterson Air Force Base, OH: Flight Dynamics Laboratory.


A-39


