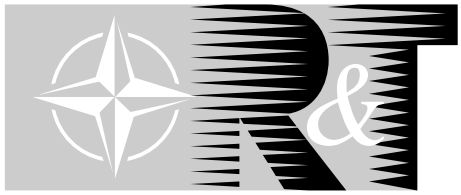
RTO-TR-021
AC/323(HFM-018)TP/19

RTO-TR-021

NORTH ATLANTIC TREATY ORGANIZATION



RESEARCH AND TECHNOLOGY ORGANIZATION

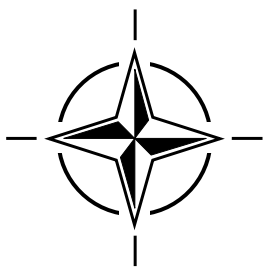
BP 25, 7 RUE ANCELLE, F-92201 NEUILLY-SUR-SEINE CEDEX, FRANCE

RTO TECHNICAL REPORT 21

NATO Guidelines on Human Engineering Testing and Evaluation

(Directives OTAN en matière d'essais et d'évaluations ergonomiques)

Final Report of the RTO Human Factors and Medicine Panel (HFM) Research and Study Group 24 on Human Engineering Testing and Evaluation.



Published May 2001

Distribution and Availability on Back Cover

This page has been deliberately left blank

Page intentionnellement blanche

RTO-TR-021
AC/323(HFM-018)TP/19

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION

BP 25, 7 RUE ANCELLE, F-92201 NEUILLY-SUR-SEINE CEDEX, FRANCE

RTO TECHNICAL REPORT 21

NATO Guidelines on Human Engineering Testing and Evaluation

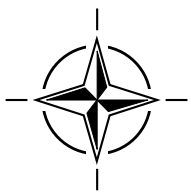
(Directives OTAN en matière d'essais et d'évaluations ergonomiques)

Final Report of the RTO Human Factors and Medicine Panel (HFM) Research and Study Group 24 on Human Engineering Testing and Evaluation.

Authors

J. C. GEDDIE (US) (RSG Chairman)

L. C. BOER (NE), R. J. EDWARDS (UK), T. P. ENDERWICK (US), N. GRAFF (FR), C. PFENDLER (GE), J.-Y. RUISSEAU (FR), P. A. VAN LOON (NE)



The Research and Technology Organization (RTO) of NATO

RTO is the single focus in NATO for Defence Research and Technology activities. Its mission is to conduct and promote cooperative research and information exchange. The objective is to support the development and effective use of national defence research and technology and to meet the military needs of the Alliance, to maintain a technological lead, and to provide advice to NATO and national decision makers. The RTO performs its mission with the support of an extensive network of national experts. It also ensures effective coordination with other NATO bodies involved in R&T activities.

RTO reports both to the Military Committee of NATO and to the Conference of National Armament Directors. It comprises a Research and Technology Board (RTB) as the highest level of national representation and the Research and Technology Agency (RTA), a dedicated staff with its headquarters in Neuilly, near Paris, France. In order to facilitate contacts with the military users and other NATO activities, a small part of the RTA staff is located in NATO Headquarters in Brussels. The Brussels staff also coordinates RTO's cooperation with nations in Middle and Eastern Europe, to which RTO attaches particular importance especially as working together in the field of research is one of the more promising areas of initial cooperation.

The total spectrum of R&T activities is covered by the following 7 bodies:

- AVT Applied Vehicle Technology Panel
- HFM Human Factors and Medicine Panel
- IST Information Systems Technology Panel
- NMSG NATO Modelling and Simulation Group
- SAS Studies, Analysis and Simulation Panel
- SCI Systems Concepts and Integration Panel
- SET Sensors and Electronics Technology Panel

These bodies are made up of national representatives as well as generally recognised 'world class' scientists. They also provide a communication link to military users and other NATO bodies. RTO's scientific and technological work is carried out by Technical Teams, created for specific activities and with a specific duration. Such Technical Teams can organise workshops, symposia, field trials, lecture series and training courses. An important function of these Technical Teams is to ensure the continuity of the expert networks.

RTO builds upon earlier cooperation in defence research and technology as set-up under the Advisory Group for Aerospace Research and Development (AGARD) and the Defence Research Group (DRG). AGARD and the DRG share common roots in that they were both established at the initiative of Dr Theodore von Kármán, a leading aerospace scientist, who early on recognised the importance of scientific support for the Allied Armed Forces. RTO is capitalising on these common roots in order to provide the Alliance and the NATO nations with a strong scientific and technological basis that will guarantee a solid base for the future.

Published May 2001

ISBN 92-837-1068-1

Printed by St. Joseph Ottawa/Hull
(A St. Joseph Corporation Company)
45 Sacré-Cœur Blvd., Hull (Québec), Canada J8X 1C6

NATO Guidelines on Human Engineering Testing and Evaluation

(RTO TR-021 / HFM-018)

Executive Summary

The purpose of this report is to document the efforts of RSG-24, which was initiated by DRG Panel 8 in 1992, and was sponsored after the merger of DRG and AGARD by the Human Factors and Medicine (HFM) Panel of NATO's Research and Technology Organization (RTO). The report presents the RSG's recommended guidelines for accomplishing human engineering test and evaluation. The goal was standardization of test content, procedures and conditions/sequence of test events. The intent was not to impose standardization of system engineering design. The guidelines are expected to facilitate the sharing of data and evaluations, which will cut test costs by reducing duplication and the quantity of test data, required to support decisions.

Human engineering test and evaluation can occur at any point during the acquisition process, but is most often done when a system is undergoing developmental and operational testing. Developmental testing usually occurs earlier and is primarily oriented toward validating the system design, but may include an early assessment of the system's performance in an operational environment. It is often characterized by testing partial systems under benign test conditions, using test subjects who are considerably more knowledgeable of the system than the eventual users of the system will be. Operational testing generally concentrates on testing the whole system, under realistic conditions, using as test subjects, either real users or personnel who closely represent the user in terms of selection, training, and experience. The primary focus of operational testing is to assess the system's ability to perform its mission under realistic conditions.

Human engineering test and evaluation includes data in the following categories:

(a) Data collected to describe relevant characteristics of test participants. Commonality of data in this area is necessary so the test specialist can more accurately use previous test results to make performance predictions for a different system, set of test conditions etc.

(b) Measurement of operator workload. It is important to understand how much of an operator's resources are required to produce satisfactory system performance.

(c) Measurement of human performance. A system's hardware may meet all human engineering design criteria and be liked by operators and maintainers, yet fall short of performance expectations unless one verifies that operators and maintainers satisfactorily perform critical tasks, both under normal, benign conditions and under degraded conditions.

(d) Assessment of user acceptance of the system. Even a system that has been well human engineered for performance and safety may have characteristics that are negatively appreciated by users and that thereby cause system performance to be limited.

(e) Measurements of hardware characteristics. The environment and physical attributes of a system can have positive or negative influences on human performance. Therefore, it is important to measure physical characteristics of the system such as size, weight, light levels, noise levels, crew workspace layout, ingress and egress provisions, temperature, vibration, the brightness, legibility and labeling of displays, and the placement, configuration and force requirements of controls.

Directives OTAN en matière d'essais et d'évaluations ergonomiques

(RTO TR-021 / HFM-018)

Synthèse

Ce rapport a pour objectif de rendre compte des travaux du groupe RSG-24, créé par la commission 8 du GRD en 1992, et repris, suite au fusionnement du GRD et de l'AGARD, par la commission sur les facteurs humains et la médecine (HFM) de l'Organisation pour la recherche et la technologie de l'OTAN (RTO). Le rapport présente les directives proposées par le RSG pour la réalisation d'essais et d'évaluations ergonomiques. Le groupe s'était donné comme objectif la normalisation du contenu, des procédures, des conditions et du séquencement des essais dans ce domaine. Le groupe n'avait pas l'intention d'imposer la normalisation de la conception de l'ingénierie des systèmes. Ces directives devaient faciliter le partage de données et les évaluations, ceci ayant pour effet de faire baisser les coûts des essais en diminuant la redondance et le volume des données d'essais nécessaires pour soutenir la prise de décisions.

Les essais et les évaluations ergonomiques peuvent se faire à n'importe quelle étape du processus d'acquisition, mais ils sont souvent organisés en même temps que les essais de mise au point et de fonctionnement du système. Les essais de mise au point sont généralement organisés plus en amont et sont plutôt orientés vers la validation de la conception du système; mais ils peuvent comporter une évaluation préliminaire des performances du système dans un environnement opérationnel. Ils sont souvent caractérisés par des essais partiels de systèmes dans des conditions peu contraignantes, faisant appel à des expérimentateurs ayant beaucoup plus de connaissances du système que les utilisateurs finaux. Les essais de fonctionnement sont généralement axés sur le système complet, dans des conditions réelles, avec comme acteurs soit des utilisateurs réels, soit des personnes très représentatives de l'utilisateur du point de vue de la sélection, de la formation et de l'expérience. L'objectif principal des essais de fonctionnement est d'évaluer la capacité du système à remplir sa mission dans des conditions représentatives de la réalité.

Les essais et les évaluations ergonomiques font appel à des données appartenant aux catégories suivantes:

(a) Données représentatives des caractéristiques typiques des participants aux essais. Dans ce domaine, il est important d'assurer l'identité des données, pour que le spécialiste puisse se servir des résultats d'essais précédents pour prévoir les performances d'un système différent ou d'un ensemble de conditions d'essais différentes, etc..

(b) Calcul de la charge de travail de l'opérateur. Il est important de bien apprécier le pourcentage des moyens de l'opérateur nécessaire pour assurer des performances système satisfaisantes.

(c) Mesure des performances humaines. Le matériel composant un système peut répondre à tous les critères de conception ergonomiques et plaire aux opérateurs et aux techniciens de maintenance, mais décevoir du point de vue des performances si les opérateurs et les techniciens de maintenance ne réalisent pas certaines tâches critiques de façon satisfaisante, dans des conditions normales, moyennement contraignantes ou dégradées.

(d) Evaluation de l'acceptation du système par l'utilisateur. Même les systèmes ergonomiquement bien conçus du point de vue des performances et de la sécurité peuvent présenter des caractéristiques qui seront perçues négativement par les utilisateurs. Ces caractéristiques risquent alors de limiter les performances du système.

(e) Mesure des caractéristiques matérielles. L'environnement et les attributs physiques d'un système peuvent avoir des conséquences positives ou négatives sur les performances humaines. Par conséquent, il est important de mesurer les caractéristiques physiques d'un système telles que les dimensions, la masse, les niveaux d'éclairage, les niveaux sonores, la disposition de l'espace de travail de l'équipage, les possibilités d'entrée et de sortie, la température, les vibrations, la luminosité, la lisibilité et l'étiquetage des écrans, ainsi que la localisation, la configuration et la force demandée pour manipuler les commandes.

Click inside the blue boxes to view the corresponding section

Contents

Human Factors and Medicine Panel Officers

Chairman:

Dr M.C. WALKER
Director, Centre for Human Sciences
F138 Bldg - Room 204
DERA
Farnborough, Hants GU14 0LX
United Kingdom

Deputy Chairman:

Col. W.D. TIELEMANS
RNLAF/SGO
P O Box 20703
Binckhorstlaan, 135
2500 ES The Hague
The Netherlands

RSG 24 Members and **(Authors/Co-Authors)**

Chairman

Dr J.C. GEDDIE
US Army Research Laboratory
HQ TEXCOM
ATTN: AMSRL-HR-MV
Fort Hood, Texas 76544-5073, U.S.A.
e-mail : geddiejames@otc.army.mil
Tel : +1 254 288 95 72
Fax : +1 254 288 16 91

Members

Dr L.C. BOER
Institute for Perception TNO
Postbus 23 e-mail : boer@tm.tno.nl
NL-3769 ZG Soesterberg, The Netherlands
Tel : +31 346 356 307
Fax : +31 346 353 977
e-mail : boer@tm.tno.nl

Mr T. P. ENDERWICK
SPAWARSYSCEN-SD
53345 Engineer Street
San Diego, Ca 92151-7260, U.S.A.
Tel: +1 619 553 8007
Fax : +1 619 553 9391
e-mail : enderwic@spawar.navy.mil

Mr C. PFENDLER
Forschungsinstitut fur Funk and Mathematik
Abt. Ergonomie und Fuehrungssysteme
Neuenahrer Str. 20
53343 Wachtberg-Werthhoven, Germany
Tel : +49 228 943 54 16
Fax : +49 228 943 5508
e-mail : pfendler@fgan.de

Dr R. J. EDWARDS
DERA Centre for Human Sciences
Head Workspace Ergonomics and
Anthropometry
Building F138 - Room 105
Farnborough, Hants GU14 6TD, United Kingdom
Tel : +44 1252 394 420
Fax : +44 1252 393 097
e-mail : rjedwards@dra.hmg.gb

LtCol N. GRAFF
Section Technique de l'Armée de Terre
Groupement NBC/FH
Route de la Minière, Camp de Satory
78013 Versailles, France
Tel : +33 1 39 67 33 20
Fax : +33 1 39 67 32 13

Mr J. I. RUISSEAU
DGA Division Facteurs Humains
Etablissement Technique d'Angers
BP 36
Route de Laval
49460 Montreuil-Juigne, France
Tel : +33 2 41 93 69 40/68 40
Fax : +33 2 41 936704
e-mail : nrenault@cedocar.fr

Ir. P. A. VAN LOON
DMKL/AB
P O Box 90822
2509 LV The Hague, The Netherlands
Tel : +31 70 316 8372/6655
Fax : +31 70 316 48 99

Major Linda BOSSI (member only)
Head Operational Human Engineering
Human Engineering Sector
DCIEM
1133 Sheppard Ave. W.
North York, Ontario M3M 3B9, Canada
Tel : +1 416 635 2197
Fax : +1 416 635 21 32
e-mail : lindabossi@dciem.dnd.ca

PANEL EXECUTIVE

Dr C. Wientjes
BP 25, 7 rue Ancelle
92201 Neuilly-sur-Seine Cedex, France
Tel: +33 (0)1 55 61 22 60/62
Fax: +33 (0)1 55 61 22 99/98
E-Mail: wientjesc@rta.nato.int

This page has been deliberately left blank

Page intentionnellement blanche

Chapter 1

Introduction

1.1 Background

Scientific inquiry into human performance in military weapon systems first occurred during WWII when Paul Fitts, an experimental psychologist, was asked to find out why the attrition/mortality rate was so high among pilot trainees and new pilots in the Army Air Corps. After systematic study, he identified shortcomings in pilot selection, pilot training, and in the design and arrangement of controls and displays in cockpits. He then recommended corrective actions for each of the problems. When the corrective fixes were made, pilot attrition decreased noticeably and the military services slowly adopted the approach for other types of weapons.

During the ensuing two decades, the concern for the human operators' role in weapon system performance was expanded and systematized into an approach known as the Personnel Subsystem. In the early 70's the approach was revitalized and renamed Human Factors Engineering. The latest emphasis on the human impacts on weapon systems originated in the US Army during the mid-1980s and is known as Manpower, Personnel, and Training Integration (MANPRINT). The objective of this comprehensive management and technical effort is to assure total system effectiveness by continuous integration into the materiel development and acquisition of all relevant information concerning it's human users.

MANPRINT addresses human performance via seven domains: Human Factors Engineering, Safety, Health Hazards, Manpower, Personnel, Training, and Soldier Survivability. The first three of these are concerned primarily with the design of hardware and are addressed by the Human Factors Engineering discipline during the systems engineering design process. These three domains were also the concern of RSG 24, as its purpose was to develop testing guidelines to ensure that human users can safely perform their tasks in a manner that supports system performance and meets operational requirements.

Human engineering testing and evaluation are conducted to ensure that the intended users of a system can operate and maintain it. This type of testing determines if the system's equipment meets human engineering, safety and other criteria relevant to human use, while meeting all mission performance requirements. Analyses to be carried out must verify that effective, efficient and safe operation of the total human-machine system is supported by the overall architecture, workspace design and workload imposed by the system's design.

Human engineering testing and evaluation is an important component of any test effort as it is generally the only activity that looks at the influence of human performance on system performance. The need for ensuring accurate human performance is obvious when operators are in the control loop of a system as they directly affect performance of the system. However, even in the most automated system, humans are often assigned critical functions such as enabling, programming, initializing, calibrating, verifying, validating, designating, and authorizing. Indeed, accurate human performance becomes more critical with increased automation because of the absence of a human operator. For example, once a cruise missile is in flight, the operator is out of the missile's control loop and cannot compensate for any errors that may have escaped detection when the crew loaded the missile onto a launch platform, programmed it with a flight plan and launched it.

There are presently no widely accepted techniques and methods for accomplishing human engineering tests. The approach taken to accomplishing human engineering testing generally depends on the organization's tradition and upon the individual specialist's education, experience, and intuition. In the absence of a common approach and data set, human engineering test results are difficult to compare for different test conditions, systems, armed services, or nations. As a result, duplicative testing becomes necessary, and important lessons learned and generalities of findings are lost.

Recognizing these deficiencies, Panel 8 of the Defence Research Group (DRG) approved the formation of an Exploratory Group (EG.K) on "Human Engineering Testing and Evaluation" at its fall 1990 meeting. The primary goal of EG.K was to assess the level of interest among NATO nations in development of guidelines for conducting human engineering testing and evaluation and to lay the framework for developing such guidelines.

The EG.K met on two occasions to produce Terms of Reference (TOR) and Programme of Work (POW) documents which defined what the effort would consist of and how it would be carried out. In June of 1992, Panel 8 approved the TOR and POW and forwarded them to the DRG with the recommendation that a Research Study Group (RSG) be formed to execute the work proposed in these documents. In September of 1992, the DRG accepted the recommendation and approved the establishment of RSG-24 on Human Engineeering Test and Evaluation. Meetings of RSG-24 began in the spring of 1993 and continued semiannually until the fall of 1998.

1.2 Purpose

The purpose of this report is to document the efforts of RSG-24 and to present its recommended guidelines for accomplishing human engineering test and evaluation. The goal was standardization of test content, procedures and conditions /sequence of test events. The intent was not to impose standardization of system engineering design. The guidelines are expected to facilitate the sharing of data and evaluations which will cut test costs by reducing duplication and the quantity of test data required to support decisions.

The intended audience for these guidelines includes all persons who conduct human engineering testing on military systems. The professional tester benefits from using the guidelines because he will be able to compare his test results to those gathered on other systems, under other conditions, or by other nations. The novice tester gains the additional benefit of having a readily available reference to proven techniques.

1.3 Scope

The RSG decided to develop testing guidelines that would describe test and evaluation methods already in existence and in use by test agencies. The idea was to conduct a comprehensive survey of techniques in use by test organizations throughout the armed services of the participating NATO nations and to select "best practices" as the recommended guidelines. Stipulating that techniques must be easily used in a field test environment further narrowed the possible domain of methods. Choosing among these existing methods, the guidelines were to be robust enough to be used by any nation or service, regardless of the type of system, stage of system development, or test conditions.

Human engineering test and evaluation can occur at any point during the acquisition process, but is most often done when a system is undergoing developmental and operational testing. Developmental testing usually occurs earlier and is primarily oriented toward validating the system design, but may include an early assessment of the system's performance in an operational environment. It is often characterized by testing partial systems under benign test conditions, using test subjects who are considerably more knowledeable of the system than the eventual users of the system will be. Operational testing generally concentrates on testing the whole system, under realistic conditions, using as test subjects, either real users or personnel who closely represent the user in terms of selection , training, and experience. The primary focus of Operational testing is to assess the system's ability to perform its mission under realistic conditions.

Human engineering test and evaluation includes data in the following categories:

(a) Data collected to describe relevant characteristics of test participants. Commonality of data in this area is necessary so the test specialist can more accurately use previous test results to make performance predictions for a different system, set of test conditions etc. It also helps the test specialist to ensure that tests are conducted using participants who are representative of the intended user.

(b) Measurement of operator workload. It is important to understand how much of an operator's resources are required to produce satisfactory system performance. Two systems with the same level of overall performance may impose quite different levels of workload on operators. Collection of these data also aids decisions and tradeoffs needed to ensure that workload levels are even across time and optimized for short term task performance as well as for sustained operation.

(c) Measurement of human performance. A system's hardware may meet all human engineering design criteria and be liked by operators and maintainers and still fall short of performance expectations unless we verify that operators and maintainers perform critical tasks to criterion, both under normal, benign conditions and under degraded conditions.

(d) Assessment of user acceptance of the system. Even a system that has been well human engineered for performance and safety may have characteristics that users find aversive and cause system performance to be limited by that aspect of the human-machine interface.

(e) Measurements of hardware characteristics. The environment and physical attributes of a system can have positive or negative influences on human performance. Therefore, it is important to measure physical characteristics of the system such as size, weight, light levels, noise levels, crew workspace layout, ingress and egress provisions, temperature, vibration, the brightness, legibility and labeling of displays, and the placement, configuration and force requirements of controls. These measurements are then analysed and compared to design requirements or best engineering practices to determine if they are within acceptable ranges.

It is useful to distinguish between testing and evaluation since they are separate and distinct activities and are often done by different persons and agencies.

Testing is the first part of the sequence and is concerned with test design, execution, and reporting. Its purpose is to gather and present specific data that are needed to answer questions regarding a system's capability to satisfy mission and user requirements. These data may include quantitative measurements of physical properties such as force, time, distance, etc. as well as qualitative estimates of attributes such as comfort, preference, and ease of use. Testing is normally done by test agencies, but is occasionally done by laboratories and other research and development organizations. Persons who design the tests and analyze test data normally have advanced degrees related to human factors, while data collectors may have only on- the-job training.

Evaluation, in this context, refers to decisions made based on knowledge provided by the test data. Its purpose is to determine the significance of the test data in relation to a set of criteria and/or issues established by appropriate authority that defines the required characteristics or capabilities of the system under test. Evaluation of test data can be a direct comparison of the data to a criterion or an issue, with a conclusion about whether it met, failed to meet, or exceeded the criterion or answered the issue, and a statement concerning the significance of this result on mission performance of the item under test. Evaluation is often done by the agency procuring the system by decision makers and other persons who have no formal training in human engineering. Compared to testing, evaluation is usually done with much less scientific rigor and involves consideration of other variables such as economic, political, or schedule constraints.

1.4 Method

The RSG adopted the following process to develop testing guidelines:

(a) A lead nation was appointed for each of the human engineering test and evaluation categories or topics of interest described above (see Table 1). The lead nation's delegation was responsible for moderating discussion of their topic and for writing initial and final drafts of the report 's chapter dealing with that topic.

(b) Using common data collection instruments developed by the RSG, each member of the RSG collected inputs on each of the topics from his nation's armed services, coordinating them so as to arrive at the RSG meeting with his nation's agreed upon position.

(c) The RSG worked as a team to develop consensus on any issues and to organize the inputs into agreed upon test guidelines.

(d) The RSG reviewed initial drafts of each chapter and provided revisions to the lead nations for incorporation in the final draft of the chapters.

1.5 Organization of the Report

The remainder of the report is arranged by topic as depicted in Table 1, with the addition of a final chapter that presents the overall conclusions and recommendations.

TABLE 1

Chapter	Lead Nation
2. Description of Test Participants	France
3. Measurement of Operator Workload	Germany
4. Measurement of Human Task Performance	United States
5. Assessment of User Acceptance	United States
6. Measurement of Hardware Characteristics	France

Chapter 2

Description of Test Participants

Because human performance influences total human-machine system performance, methods to account for the variance contributed by the human component of the system are needed. Collection of data on test participant characteristics that are believed to be relevant to their performance of tasks required in operation and maintenance of the system offers a means of doing that. It also provides a means for human engineering test results to be meaningfully shared across nations.

These data enable us to ensure that tests are conducted using participants whose performance is representative of that of the intended user, or at least identify any relevant differences between the characteristics of the sample of test participants and the target population of system users.

The data to accomplish this fall into three basic categories: biographic data, anthropometric data, and vocational data.

Biographic Data

This category of data includes administrative and identification data, information on military and civilian training the person brings with him to the test and system-specific training, and results of evaluations at the end of system-specific training.

Anthropometric Data

Body measurement

Minima and maxima for such parameters as height and weight should be specified, giving due consideration to the range of these dimensions expected in typical users when the system is fielded. One example application might be to establish percentile cutoff scores first for height, then determine upper and lower percentile cutting scores for weight at each height.

No personnel with a restricted-duty profile should be permitted to participate unless a task analysis reveals that the restriction on his or her activities has no impact on the task required in the test.

Sensory Acuity

All participants should have had a recent (within the last 12 months) test of vision and audition. Minimum standards should be stated for each of these sensory modalities depending upon an analysis of the requirements of the tasks to be performed. A good starting point for standards might be those required of new inductees to the military. Although there is room here for the application of judgment in individual tests, standards should nonetheless be established and stated, together with their rationale.

Vocational Data

The required military occupational specialty (MOS), complete with skill-level suffix, should be specified. This specification should include a determination of whether the specified MOS must be the participant's primary MOS or whether a person with this specialty in a secondary MOS is acceptable. If alternate specialties include the required training and are acceptable substitutes, these should also be listed. In addition, the training-requirements specifications should state whether the MOS must be a school-trained qualification or whether an on-the-job-training (OJT) qualified person meets the requirements.

Detailed description of his or her characteristics should be obtained by collecting, for example, the following information on each test participant

1. Scores from aptitude tests administered when the participant first entered the military,

2. Scores from the participants' most recent MOS tests,

3. A list (by descriptive title) of additional system-related training or courses completed,

4. Results of tests administered during system-specific training, including:

    a. Minimum performance required, and

    b. Level of performance attained by the participant.

As an aid to standardize collection of these data, the RSG proposes use of the Test Participant Data Form provided below. The form has three parts:

- the first one, to be completed by the test participant, and checked out by the test conductor (paragraph A of the form),

- the other ones, to be fulfilled by the test conductor (Parts B and C of the form).

Part A can usually be provided by each participant from memory. Parts B and C should be available from personnel files. Other pertinent administrative data, if any, can be collected or confirmed via a simple data collection form administered on the first day of the test.

Parts A and B should be the same for most tests. Part C item selection depends upon the system being tested. Examples presented are not exhaustive.

TEST PARTICIPANT
DATA FORM

A. To be completed by test participant.

1. Last Name: _____ First name: _____

2. Date (MM DD YY): _____

3. Gender: M F (circle one)

4. Draftee: D Volunteer: V (circle one)

5. Active service: A Reserve: R (circle one)

6. Grade/rank: _____

7. Military specialty code: _____

8. Id#: _____

9. Crew position (in test): _____

10. Months of experience (in tested crew position): _____

11. Height (centimeters): _____

12. Weight (kilograms): _____

13. Date of birth:
(DD MM YY) _____

14. Date entered service:
(DD MM YY) _____

15. Civilian education:

(a) 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
(circle number of years)

(b) Major area (if applicable): _____

(c) Highest level completed: _____

16. Preferred hand (circle one): Right Left

B. To be completed by test conductor for each participant:

17. Physical profile: _____

18. Armed services aptitudes battery standard scores: _____

19. Latest military specialty test score: _____

20. End-of-training proficiency results:

20.1. Written or oral exam scores: _____

20.2. Score required to pass: _____

20.3. Task performance:

	Required	Obtained
Task a _____	_____	_____
Task b _____	_____	_____
Task c _____	_____	_____
Task d _____	_____	_____
Task e _____	_____	_____

21. Other military training:

School	Course	Completed Successfully? (y/n)
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

22. Other civilian training

School	Course	Completed Successfully? (y/n)
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

C. Other measures as appropriate (including methods used)

For example;
23. Visual acuity, depth perception, preferred eye (right - left)...
24. Auditory acuity ...
25. Anthropometric data ...
26. Color perception ...
27. Operation hours/type of system ...
28. Etc...

For more information, see appendix.

Appendix

Part A:

- items 4 and 5 do not conflict: for example, a test participant can be either a draftee belonging to a reserve corps or a volunteer in active service.

- item 15 (civilian education): related to primary and secondary school, college and graduate school, that means the total number of years completed.

Part B:

- item 17: according to national standards, such as SIGYCOP, ABOZHIS, PULHES, PULHEEMS. Explanation on each physical profile rating must be provided, including database norms used for physical profile evaluation.

- item 18: according to national standards, such as ASVAB, BARB, EVP. Explanation of each aptitude battery profile must be provided, including database norms used for battery tests.

- item 19: test scores must be given related to a reference point (e.g.:threshold, mean. . .)

- item 20.3: provide a description of each task, if applicable.

Part C:

Part C depends upon which systems which are tested. Examples given in this section are not exhaustive.

- item 25: Anthropometric data provided in this section must be referred to the global population, or to some specific sample, as required.

Glossary:

If specific terms are to be used by the test conductor or when wording is unusual, explanations must be given in a glossary.

This page has been deliberately left blank

Page intentionnellement blanche

Chapter 3

Measurement of Operator Workload

1 Summary

Workload can be defined as the portion of human resources an operator expends when performing a specified task. As these resources are limited it must be prevented that the human operator is overloaded by a task conducted within a system which can result in severe performance decrements. Furthermore, two systems with the same level of overall performance may impose quite different levels of workload on operators. That is why it is important to measure not only performance but also operator workload required to produce satisfactory system performance. Collection of these data also aids decision and trade-off needed to insure that workload levels are acceptably distributed across time and optimised for short term tasks as well as for sustained operations.

Workload is assessed by subjective methods, secondary tasks, physiological techniques and measures of primary task performance. Furthermore, workload is evaluated by analytical techniques, which can be used already in the design phase of a system. Analytical techniques are not discussed here, as the goal of the research study group was to mention only empirical methods. Measures of primary tasks will be discussed in the chapter on performance measurement. Physiological and secondary task methods are not as important as rating scales for test and evaluation and that is also why they are described only in short overviews. The reasons will be explained more in detail in the corresponding chapters (e.g., controversial results, limited applicability, high implementation demands, etc.). It is recommended to apply physiological and secondary task methods only as additional tools for workload measurements (e.g. for continuous workload measurement).

For human engineering testing and evaluation subjective methods, especially rating scales, which assess operator or observer judgements of a task have many advantages and relatively few disadvantages in measuring operator workload compared to the other approaches. From the many rating scales used in test and evaluation four methods - NASA-TLX (NASA Task Load Index), SWAT (Subjective Workload Assessment Technique), MCH (Modified Cooper Harper Scale), and ZEIS (Sequential Judgement Scale) - are recommended for workload measurement and will be described in detail in the following chapters. They have been selected as they are all used in more than one nation and they can fulfil satisfactorily a number of standards against which all workload measurement methods must be compared (see the following chapters). In the following recommendations are given when to use these four methods.

NASA-TLX can be applied with a very broad range of tasks. It should be used when there is a special interest in diagnosticity, i.e., in detection of the sources of workload, which is possible by the six subscales of the method. NASA-TLX has good face validity and validity. In comparative evaluations with other rating scales NASA-TLX proved to be the most sensitive technique also with lower workload levels. When used during task performance rating of the six subscales might cause interference with the main task, especially in high workload situations.

SWAT can also be applied with most tasks. But in contrast to NASA-TLX the method has been found to be not so sensitive in low workload situations. As SWAT has only three subscales it can also be used during task performance with low risk of interference with the main task but information about the sources of workload is limited. Critical SWAT values predicting operator overload exist. Validity and face validity of SWAT is rated lower than with NASA-TLX. Studies demonstrate that the complex and cumbersome data analysis of SWAT (and of NASA-TLX) can possibly be dropped without any disadvantages.

MCH has been designed for workload assessment of tasks where the operator is less involved in active control of the system and more occupied with activities typical for modern man-machine-systems like perception, monitoring, evaluation, communication, and problem solving. The scale has demonstrated sensitivity in such tasks during simulated flight but was less useful in other environments. It gives an estimate of overall workload but no diagnostic information. In comparative studies sensitivity of MCH has been rated lower than with NASA-TLX and with SWAT.

ZEIS has been designed for evaluation of vehicle handling difficulty but is assumed to be applicable to a much wider variety of tasks as task difficulty is regarded as one of the most relevant dimensions of workload. The scale gives an overall estimate of workload without any diagnostic information. In comparative evaluations validity of ZEIS resulted to be relatively high but validation has been restricted as yet mainly to motor tasks. As they are unidimensional methods ZEIS and MCH might also be used during task performance without interfering with the main task.

The list of these four methods is not exhaustive and other methods can also be conceived to be useful tools for workload measurement. As an example two other promising approaches - SWORD (Subjective Workload Dominance Technique) and the Bedford Scale - are added to the following description of workload measurement scales.

Before this detailed descriptions are given different aspects of workload measurement in a test and evaluation environment in different nations are presented.

2 Workload Measurement in Different Nations

2.1 Testing in the Dutch Armed Forces

NATO's Defence Research Group, Panel 8, Research Study Group 24 collects data on human engineering testing and evaluation. The level of mental workload imposed on the military operator by the man-machine system is important in this context because under- and overload can undermine the potential of the whole system. The present document discusses the current status of workload testing in the Dutch armed forces and the trends for the future.

Current Dutch military procurement procedures do not mention the concept of mental workload. Procurement specifications may contain a clause stating that the to be procured system should satisfy "current ergonomic standards". In subsequent acceptance testing all specifications are evaluated. The emphasis is on technical performance. If at all, workload is evaluated in a crude and intuitive way only and although it is possible that more extensive and systematic workload tests will be undertaken following a suspicion of questionable workloads, no such cases are known.

Extensive and systematic workload tests are occasionally performed by external contractors, mainly TNO in Soesterberg, often on the personal initiative of a military officer who has a background in, or is interested in, behavioural science. The impact of these studies on actual military systems is limited. Examples are cockpit studies in which pilots under-training flew detailed missions while subjective, objective, and physiological data were collected. The aim is fundamental knowledge and the development of an in-flight workload assessment module. Other examples are the air traffic control studies in which operators on duty performed a secondary task when possible. The aim was to evaluate whether the workload was tolerable. Possible consequences were in terms of the numbers of personnel rather than in terms of system design.

For future trends, the growing awareness of the importance of human factors and the potential of the human factors contribution to design deserves mentioning. The work of RSG 14 "on analysis techniques for man-machine systems design" has had a sizeable impact in this respect. In the Royal Netherlands airforce this awareness has led to a limited participation of a human factors specialist in the procurement of an air refueling system. In the Royal Netherlands navy this awareness has led to a participation of human factors specialists in the design of warships and in the design of naval command and control facilities. This participation is expected to become permanent.

2.2 Mental Workload Measurement in a Test and Evaluation Environment (USA)

A great deal of research has been undertaken in the development and application of mental workload assessment techniques over the past several years. The workload assessment literature over this period includes several comprehensive reviews (e.g., Hancock and Meshkati, 1988; Lysaght et al., 1989; Moray, 1988; O'Donnell and Eggemeier, 1986; Wierwille and Williges,1978). However, an informed practitioner, faced with the problem of

performing a workload evaluation in a system under test could have some difficulty in selecting and applying techniques. This is attributable both to the fact that there are many techniques (e.g., Lysaght et al., 1989; Moray, 1988; O'Donnell and Eggemeier, 1986; Wierwille and Williges, 1978) available, and that some conceptual issues involved are very complex.

Assessment procedures might be divided into three categories: subjective procedures, performance-based techniques, and physiological techniques. Subjective procedures are based on operator judgements of the workload associated with performance of a task. Performance-based techniques assess workload by measuring performance of tasks. Physiological techniques use differences in physiological responses to task performance to estimate the workload imposed by performing the task. There are a number of individual assessment techniques associated with each category.

There are several properties which a workload assessment technique should have in order to be appropriate for use in assessing workload in T&E. The most important of these is *sensitivity,* which is the degree to which a given workload technique can distinguish differences in levels of load imposed on an operator. Sensitivity is a major consideration in the choice of a measurement technique because the intent is to assess differences in the workload imposed by the system under test. The property of *global sensitivity is* the capability to reflect variations in different types of resource expenditure or factors that influence workload. Most subjective assessment procedures and certain primary task and physiological measures demonstrate a broad bandwidth of sensitivity and therefore qualify as globally sensitive measures of operator workload.

Another property is *intrusion,* which is an undesirable property in which introduction of the workload measuring technique causes a change in operator-system performance.

For T&E, another important consideration is *implementation requirements* which includes any equipment or instrumentation that is necessary to present information (e.g., the stimuli required for performance of a secondary task) or record data (e.g., operator heart rate). Implementation requirements also include the data collection procedures and any operator training (e.g., practice of a secondary task, familiarisation with rating scales) that is associated with use of a technique.

Generally speaking, non intrusive workload techniques that have global sensitivity are of utmost importance in T&E applications that require a general or overall screening of the workload levels associated with a system or design option. Such techniques can be used with some confidence to determine whether an overall workload problem exists or to differentiate the workload associated with two or more design options.

"Pure" forms of operator behaviour are rare in T&E. Usually operators perform functions that involve numerous forms of behaviour, and the forms vary from task to task. A prime objective of workload measurement during T&E must be to obtain an overall assessment. Techniques used for this purpose should have global sensitivity. Otherwise, a shift in specific forms of operator behaviour may result in failure to detect workload differences.

There are several subjective assessment techniques that possess demonstrated sensitivity to a variety of demand manipulations within complex systems (e.g., aircraft, air defence). Some of the most frequently employed techniques that were designed for use in a variety of systems and that are therefore appropriate to a range of T&E applications include the Modified Cooper Harper (MCH) Scale (Wierwille and Casali, 1983), the Subjective Workload Assessment Technique (SWAT; Reid and Nygren, 1988), and the NASA Task Load Index (TLX; Hart and Staveland, 1988).

The MCH, for example, has been successfully applied to assess workload in many flight simulation experiments that incorporated several types of demand manipulations (e.g., Casali and Wierwille, 1983, 1984; Itoh et al., 1989; Skipper, Rieger, and Wierwille, 1986; Wierwille et al., 1985b). SWAT has also been extensively employed to assess pilot workload. Representative applications (e.g., Battiste and Bortolussi, 1988; Corwin et al., 1989; Hughes et al., 1990; Nataupsky and Abbott,1987; Ward and Hassoun, 1990) have demonstrated the sensitivity of SWAT to different demand manipulations in the flight environment and are representative of results that have been obtained with SWAT in other systems, such as military tanks (Whitaker et al., 1989). Likewise, TLX sensitivity has been demonstrated in many flight experiments that incorporated different demand manipulations

(e.g., Battiste and Bortolussi,1988; Corwin et al.,1989; Nataupsky and Abbott, 1987; Shively et al., 1987; Tsang and Johnson, 1989; Vidulich and Bortolussi, 1988). All three measures and the Overall Workload Scale (Vidulich and Tsang, 1987) have also been applied to derive an index of operator workload which proved sensitive to several variables in a series of evaluations conducted within U.S. Army systems (e.g., remotely piloted vehicle, mobile air defence) under various test conditions (Bittner et al., 1989; Byers et al., 1988; Hill et al., 1988).

The pattern of results that has emerged from these evaluations indicates that the MCH, SWAT, and TLX procedures represent globally sensitive measures of operator workload. Similar patterns have been demonstrated with other techniques, such as the Bedford Scale (e.g., Roscoe, 1987; Roscoe and Ellis, 1990), which is designed to assess pilot workload. Current results therefore indicate that subjective techniques can be generally classified as globally sensitive indexes of workload. Two of the noted techniques (i.e., SWAT and TLX) are multidimensional and therefore can provide some diagnostic information on the sources of workload represented by the subscales (e.g., mental effort load, time load) of the respective procedures. SWAT incorporates three such subscales, and NASA-TLX includes six subscales.

Comparative evaluations involving two or more techniques have been conducted within a variety of scenarios that range from the laboratory to simulation and T&E environments (e.g., Eggemeier and Wilson, 1991; Lysaght et al., 1989). For the most part, representative studies performed within a flight simulation environment (e.g., Battiste and Bortolussi, 1988; Corwin et al., 1989; Nataupsky and Abbott, 1987; Tsang and Johnson, 1989; Vidulich and Bortolussi, 1988) have not shown marked and consistent sensitivity differences among techniques. The previously referenced series of experiments (Bittner et al.,1989; Byers et al., 1988; Hill et al., 1988) conducted within Army systems did, however, demonstrate that the TLX procedure consistently demonstrated higher loadings than did the SWAT and MCH on the overall index of subjective workload that was derived in each case. In addition, Nygren (1991) compared the psychometric properties of the SWAT and TLX procedures and concluded that neither technique was generally preferable to the other. SWAT was viewed as having the greatest potential for identification of factors such as cognitive mechanisms affecting mental workload judgements. Conversely, TLX was seen as particularly appropriate for problems in applied settings and was considered potentially more sensitive than SWAT at low levels of workload. Its applicability in applied settings and potential sensitivity at low levels of workload suggest that TLX be given strong consideration in T&E applications, although the specific objectives of an evaluation should be considered when choosing a technique.

Intrusion is not usually considered an important factor in applications of subjective measures because these techniques are typically administered after task performance. For instance, the previously referenced systematic series of flight simulation experiments (Casali and Wierwille 1983, 1984; Wierwille et al. 1985a) resulted in no differential intrusion with application of the MCH relative to the other performance-based and physiological assessment procedures that were evaluated. Some applications of subjective techniques to complex systems (e.g., Corwin et al., 1989) have utilized postmission subjective ratings that were supported by videotapes of the mission itself. This procedure represents an approach that can be considered when rating collection is precluded by mission constraints or when intrusion that could be associated with concurrent-task performance and rating scale completion is to be eliminated. Without videotapes or other aiding, delay of ratings for substantial periods following task performance is not recommended. Such delays have the potential to be associated with the loss of critical rating information from operator memory. Current evidence (e.g., Lutmer and Eggemeier, 1990; Moroney et al., 1992), however, suggests that delays of up to 15 minutes may not be critical with some rating scales under certain laboratory or simulator conditions.

Additional important considerations in T&E applications include implementation requirements and user acceptance. Subjective assessment procedures typically require little instrumentation but differ in other implementation requirements. Nominal applications of multidimensional procedures such as SWAT and TLX, for example, require that data necessary to combine subscales into one workload estimate be gathered from operators. The scale development procedure for the SWAT scale, for instance, can require approximately 1 hr to complete. More detailed information concerning scale development requirements is available for the SWAT (e.g., Reid et al., 1989) and the TLX (NASA Ames Research Center). User acceptance of subjective procedures is typically high, and comparative information concerning techniques is not extensive. In one of the system test comparisons noted earlier, Byers et al. (1988) reported higher user acceptance ratings of the TLX procedure relative to the SWAT and MCH techniques.

Finally, subjective measures also provide the potential for application in a projective mode, which can permit some predictive assessment of the workload associated with proposed systems or design options that have not yet been implemented. Such projective applications are based on descriptions of a proposed system or design option (e.g., display formats), which are used by subject matter experts in order to project or estimate the workload that would be associated with performance under stipulated conditions or mission scenarios (e.g., Reid et al., 1984; Vidulich et al., 1991). Although reports of projective applications are not extensive, evidence indicates that projective ratings gathered from subject matter experts may correlate highly with subjective ratings obtained after participation in simulated missions (e.g., Eggleston, 1984; Vidulich et al., 1991).

References

Battiste, V. and Bortolussi M. (1988). Transport pilot workload: A comparison of two subjective techniques. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp.15 -154). Santa Monica, CA: Human Factors and Ergonomics Society.

Bittner, A.C., Byers, J.C., Hill, S.G., Zaklad, A.L., and Christ, R.E. (1989). Generic workload ratings of a mobile air defence system (LOS-F-H). In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1476-1480). Santa Monica. CA: Human Factors and Ergonomics Society.

Byers, J.C., Bittner; A.C., Hill, S.G., Zaklad, A.L., and Christ, R.E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 1145-1149). Santa Monica, CA: Human Factors and Ergonomics Society.

Casali, J.G. and Wierwille, W.W. (1983). A comparison of rating scale, secondary task, physiological, and primary task workload estimation techniques in a simulated flight task emphasising communications load. *Human Factors, 25*, 623-641.

Casali, J.G. and Wierwille, W.W. (1984). On the measurement of pilot perceptual workload: A comparison of assessment techniques addressing sensitivity and intrusion issues. *Ergonomics, 27*, 1033-1050.

Corwin, W.H., Sandry-Garza, D.L., Biferno, M.H., Boucek, G.P., Logan, A.L., Jonsson, J.E., and Metalis S.A. (1989). *Assessment of crew workload measurement; methods, techniques, and procedures: Volume 1. Process, methods, and results* (Tech. Report WRDC-TR-897006). Wright-Patterson Air Force Base, OH: Wright Research and Development Center, Air Force Systems Command.

Eggemeier, F.T. and Wilson, G.F. (1991). Subjective and performance-based assessment of workload in multitask environments. In D. L. Damos (Ed.), *Multiple task performance* (pp. 217-278). London: Taylor & Francis.

Eggleston, R.G. (1984). A comparison of projected and measured workload ratings using the Subjective Workload Assessment Technique (SWAT). In *Proceedings of the 1984 IEEE National Aerospace and Electronics Conference* (pp. 827-831). New York: Institute of Electrical and Electronics Engineers.

Hancock, P.A. and Meshkati, N. (Eds.). (1988). *Human mental workload.* Amsterdam: North- Holland.

Hart, S.G. and Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of experimental and theoretical research. In P.A. Hancock and N. Meshkati (Eds.), *Human mental workload* (pp. 139183). Amsterdam: North-Holland.

Hart, S.G. and Wickens, C.D. (1990). Workload assessment and prediction. In H.R. Booher (Ed.), *Manprint: An approach to systems integration* (pp. 257-296). New York: Van Nostrand Reinhold.

Hill, S.G., Byers, J.C., Zaklad, A.L., and Christ, R.E. (1989). Subjective workload assessment during 48 continuous hours of LOS-F-H operations. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1129-1133). Santa Monica, CA: Human Factors and Ergonomics Society.

Hill, S.G., Zaklad, A.L., Bittner, A.C., Byers, J.C., and Christ, R.E. (1988). Workload assessment of a mobile air defence system. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 1068-1072). Santa Monica, CA: Human Factors and Ergonomics Society.

Hughes, E.R., Hassoun, J.A., Ward, G.F., and Rueb, J.D. (1990). *An assessment of selected workload and situation awareness metrics in a part-mission simulation* (Tech. Report ASD, TR-90-5009). Wright-

Patterson Air Force Base, OH: DCS for Integrated Engineering and Technical Management, Aeronautical Systems Division, Air Force Systems Command.

Itoh, Y., Hayashi, Y., Tsukui, L, and Saito, S. (1989). Heart rate variability and subjective mental workload in flight task validity of mental workload measurement using H. R. V. method. In M. J. Smith and G. Salvendy (Eds.), *Work with computers: Organizational, management, stress and health aspects* (pp. 209-216). Amsterdam: Elsevier.

Lutmer, P.A. and Eggemeier, F.T. (1990, May). *Effect of intervening task performance on subjective ratings of mental workload.* Presented at the Fifth Mid-Central Ergonomics Meeting, Dayton, OH.

Lysaght, R.J., Hill, S.G., Dick, A.O., Plamondon, B.D., Linton, P.M., Wierwille, W.W., Zaklad, A.L., Bittner, A.C., and Wherry, R.J. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies* (Tech. Report 851). Fort Bliss, TX: U.S. Army Research Institute, Field Unit.

Moray, N. (1988). Mental workload since 1979. In D. J. Oborne (Ed.), *International Reviews of Ergonomics* (pp. 123-150). London: Taylor & Francis.

Moroney, W.F., Biers, D.W., Eggemeier, F.T., and Mitchell, J.A. (1992). A comparison of two scoring procedures with the NASA task load index in a simulated flight task. In *Proceedings of the 1992 National Aerospace and Electronics* Conference (pp. 734-740). New York: Institute of Electrical and Electronics Engineers.

NASA Ames Research Center. (n.d.). *NASA Task Load Index (NASA-TLX): A paper and pencil package, Version 1.0.* Moffett Field, CA: Human Performance Research Group, NASA Ames Research Center.

Nataupsky, M. and Abbott, T.S. (1987). Comparison of workload measures on computer-generated primary flight displays. In *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 548-552). Santa Monica, CA: Human Factors and Ergonomics Society.

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors, 33*, 17-33.

O'Donnell, R.D. and Eggemeier, F. T. (1986). Workload assessment methodology. In K.R. Boff, L. Kaufman, and J. Thomas (Eds.), *Handbook of perception and human performance: Volume 11. Cognitive processes and performance* (pp. 421142/49). New York: Wiley.

Reid, G. B. and Nygren, T.E. (1988). The Subjective Workload Assessment Technique: A scaling procedure for measuring mental workload. In P.A. Hancock and N. Meshkati (Eds.), *Human mental workload* (pp. 185-218). Amsterdam: North-Holland.

Reid, G.B., Potter, S.S., and Bressler, J.R. (1989). *Subjective Workload Assessment Technique (SWAT): A user's guide* (Tech. Report AAMRL-TR-89-023). Wright-Patterson Air Force Base, OH: USAF Armstrong Laboratory.

Reid, G.B., Shingledecker, C.A., Hockenberger, R.L., and Quinn, T.J. (1984). A projective application of the subjective workload assessment technique. In *Proceedings of the 1984 IEEE National Aerospace and Electronics Conference* (pp. 824-826). New York: Institute of Electrical and Electronics Engineers.

Roscoe, A.H. (1987). In-flight assessment of workload using pilot ratings and heart rate. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload* (AGARDograph 282). (pp. 78-82). Neuilly sur Seine, France: Advisory Group for Aerospace Research and Development.

Roscoe, A.H. and Ellis, G.A. (1990). *A subjective rating scale for assessing pilot workload in flight: A decade of practical use* (Tech. Report 90019). Farnborough, England: Royal Aerospace Establishment.

Shively, R., Battiste, V., Matsumoto, J., Pepiton, D., Bortolussi, M., and Hart, S.G. (1987). Inflight evaluation of pilot workload measures for rotorcraft research. In *Proceedings of the Fourth Symposium on Aviation Psychology* (pp. 637-643). Columbus, OH: Department of Aviation, Ohio State University.

Skipper, J.H., Rieger, C.A., and Wierwille, W.W. (1986). Evaluation of decision tree rating scales for mental workload estimation. *Ergonomics, 29*, 585-599.

Tsang, P.S. and Johnson, W.W. (1989). Cognitive demands in automation. *Aviation, Space and Environmental Medicine, 60,* 130-135.

Vidulich, M.A. and Bortolussi, M.R. (1988). *Speech recognition in advanced rotorcraft: Using speech controls to reduce manual control overload.* Presented at the American Helicopter Society National Specialists' Meeting-Automation Applications in Rotorcraft, Atlanta, GA.

Vidulich, M.A. and Tsang, P.S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. In *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 1057-1061).Santa Monica, CA: Human Factors and Ergonomics Society.

Vidulich, M.A., Ward, G.F., and Schueren, J. (1991). Using the Subjective Workload Dominance (SWORD) technique for projective workload assessment. *Human Factors, 33,* 677-691.

Ward, G.F., and Hassoun, J.A. (1990). *The effects of headup display (HUD) pitch ladder articulation, pitch number, location and horizon line length on unusual attitude recovery in the F-16* (Tech. Report ASD-TR-90-5008). Wright-Patterson Air Force Base, OH: DCS for Integrated Engineering and Technical Management, Aeronautical Systems Division, Air Force Systems Command.

Whitaker, L., Peters, L., and Garinther, G. (1989). Tank crew performance: Effects of speech intelligibility on target acquisition and subjective workload assessment. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1411-1413). Santa Monica, CA: Human Factors and Ergonomics Society.

Wierwille, W.W. and Casali, J.G. (1983). A validated rating scale for global mental workload measurement applications. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 129-133). Santa Monica, CA: Human Factors and Ergonomics Society.

Wierwille, W.W., Casali, J.G., Connor, S.A., and Rahimi, M. (1985a).Evaluation of the sensitivity and intrusion of mental workload estimation techniques. In W. B. Rouse (Ed.), *Advances in man-machine systems research* (Vol. II, pp. 51-127). Greenwich, CT: JAI Press.

Wierwille, W.W. and Connor, S.A. (1983).Evaluation of twenty workload assessment measures using a psychomotor task in a moving-base aircraft simulator. *Human Factors, 25,* 1-16.

Wierwille, W.W., Rahimi, M., and Casali, J.G. (1985b).Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediation activity. *Human Factors, 27,* 489-502.

Wierwille, W.W. and Williges, R. (1978). *Survey of operator workload assessment techniques* (Final Tech. Report S-78-101). Blacksburg, VA: Systemetrics, Inc.

2.3 Measurement of Operator Workload Under Field Conditions (Germany)

The present paper is written according to the program of work of NATO RSG 24:" Human Engineering Testing and Evaluation" and is based on findings from testing and evaluation documents and on results on military ergonomics research. Special emphasis is given to the measurement aspects of workload assessment methods. In this respect three different kinds of relationships between the workload predictor and workload are distinguished:

- a predictor shows concordance with workload, if the predictor scores significantly vary in the same direction as the workload levels as expected (e.g., significant increase in heart rate with increasing workload)
- a predictor shows discordance with workload, if the predictor scores vary in the opposite direction to the workload levels than expected (e.g., error decrease in a secondary task while workload increases)
- a predictor shows indifference, when no significant relationship between predictor scores and workload levels can be found.

The following findings reflect approximately the present state of the art of workload measurement in German military ergonomics.

Standardized Test Procedures

Up to now a standardized test procedure for workload measurement only exists for test and evaluation of land based vehicles. It is based on the "Allied Vehicle Testing Publications" (AVTP) of NATO (AC/ 225, Panel II/WGE 3) in connection with the STANAGs 4357 and 4358 and is also an agreement between the members of the West European Union (WEU 4 FT 6). The respective test procedures are presented in the AVTPs and also in the WEU Trial Series 09-100. From an ergonomic standpoint it is important that the standardized test procedure for workload assessment as well as all other ergonomic test procedures have to be followed not only in bi- or multilateral tests but are also obligatory for test and evaluation of systems in the Federal Armed Forces Testcenter for Wheeled and Tracked Vehicles since 1992.

Results From Military Ergonomics Research

The findings are based on typical field experiments, in which the stressor variable could not unequivocally be quantified. Furthermore some laboratory experiments are mentioned. The results are described under the following classes of workload measurement methods:

- subjective methods
- performance measurement methods
- physiological methods
- secondary task methods

Analytical techniques are not mentioned because of the specific restrictions of the prediction.

Subjective Methods

Most workload research has been done with subjective methods. In the selected analyses subjectively experienced workload was measured with longterm (32 days) submarine cruises (Rummel, 1992a) and with multiple hours small military unit missions (Wiegand, 1988).

To measure workload and subjective fitness to perform Rummel used the BLV Questionnaire (Beanspruchung und subjektives Leistungsvermögen: Workload and subjective fitness to perform; Bronner and Karger, 1985).

In the BLV questionnaire 24 items are assigned to the following four scales:

- Psychische Anspannung (psychological tension)
- Subjektive Leistungsfähigkeit (subjective fitness to perform)
- Leistungsaversion (performance aversion)
- Ermüdung (fatigue)

Significant effects during the mission were found with the scales "subjective fitness to perform", "performance aversion" and "fatigue" whereas "psychological tension" did not differentiate significantly.

Concordance of scores from subjective methods with workload was also observed nearly always in the analysis of Wiegand, who used a self report questionnaire ("Selbstbeobachtungsfragebogen") from Bartenwerfer (1960) and from Apenburg (1986) ("Eigenzustandsskala") in a pre-post comparison. The questionnaire of Bartenwerfer includes the following eight scales:

- Unternehmungslust (enterprising)
- Konzentration (concentration)
- Zeichen geistiger Erschöpfung (symptoms of mental exhaustion)
- Zeichen körperlicher Erschöpfung (symptoms of physical exhaustion)
- Aktivation (activation)

- Bereitschaft zur geistigen Tätigkeit (readiness for mental activities)
- Bereitschaft zur körperlichen Tätigkeit (readiness for physical activities)
- Ausgeglichenheit (psychological balancedness)

With the exception of the scale "psychological balancedness", which differentiated at $p<0.05$, all other scales differentiated at $p<0.001$.

The questionnaire from Apenburg (1986) comprises 36 items with ratings with six steps which are combined into the following eight subscales:

- Anstrengungsbereitschaft (readiness to make an effort)
- Kontaktbereitschaft (readiness to make contacts)
- Soziale Anerkennung (social recognition)
- Selbstsicherheit (selfassurance)
- Stimmungslage (mood)
- Spannungslage (tension level)
- Erholtheit (recoverdness)
- Schläfrigkeit (drowsiness)

With the questionnaire from Apenburg the subscales "readiness to make an effort", "recoverdness" and "drowsiness" differentiated at $p<0.001$ and the subscales "readiness to make contacts" and "tension level" showed significant results at $p<0.05$, whereas the other subscales did not show any significant effect.

In another experiment a tracking task with simplified car dynamics with three difficulty levels was used and workload was measured with a rating scale with seven verbal descriptors (Pfendler,1982). The rating scale differentiated significantly between all task levels at $p<0.01$.

In a learning experiment with a pattern detection task (Pfendler, 1993) a German version of NASA-TLX was compared with the Sequential Judgement Scale ZEIS from Pitrella and Käppler (1988). From the 12 learning blocks, from which 66 multiple comparisons can be derived, 17 comparisons proved to be significantly different with ZEIS and 16 with NASA-TLX at $p<0.05$. Both workload measurement methods showed consistence in 15 comparisons.

Performance Measurement Methods

Whereas the subjective methods differentiated in general in concordance with workload in the pre-post comparisons under field conditions, this was not the case for most of the performance tests. These predictors responded primarily indifferent or in discordance to workload. As an example the data of the analysis of Rummel (1992b) and Wiegand (1988) are presented again.

Rummel used a test battery with five cognitive tests (Rummel, 1992):

- QRST memory task
- mental arithmetics task
- sentence verification task
- dotting task
- number search task

Only the corrections with the dotting task differentiated in concordance with workload. The other tasks did not show any significant differences or showed performance improvements discordant to workload.

The same tendency was found in the analysis from Wiegand (1988). No differences could be found with the scores in the two hand control tester, in the colour word interference test and in flicker fusion frequency. Discordant to workload responded the scores in complex reaction tests. Concordant results to workload could be demonstrated only with simple reaction time ($p<0.001$) and with continuous time estimation ($p<0.001$), which showed a considerable reduction in variability. Another relevant result was the divergence between objective and subjective data from complex reaction tests. After the loading mission a performance increment in this test was subjectively experienced as performance decrement.

Physiological Methods

Physiological methods have shown remarkable good differentiation between different levels of physical workload from environmental factors under laboratory conditions with quantifiable stressors like noise, vibration and climate factors. That is why physiological methods which measure heart rate, blood pressure, galvanic skin response, muscle potentials and brain activity have been included in the standardized test procedure AVTP 90-100. As some of them are very prone to artefacts these measures are used under field conditions only under special circumstances.

Problems with artefacts exist also when measuring mental workload with physiological measures. But in comparison to physical workload, differentiation is often poor even under more restricted conditions or the results are at least contradictory. In the tracking task with simplified car dynamics from Pfendler (1982), heart rate could differentiate significantly in only one from three comparisons and two significant comparisons could be found with the best of four measures of heart rate irregularity, whereas three other measures did not show any significant result.

A physiological predictor not prone to artefacts is salivary-cortisol-secretion, which responds highly concordant to mental workload also under field conditions (Deinzer et al.,1991; Hellhammer, 1992). For further validation, for potential acceptance of the method as a standard tool in military ergonomics, more experiments will be done with multiple levels of workload.

Secondary Task Methods

There are only a few findings with the secondary task method. As an example, in the studies of Wiegand (1974, 1989) mental workload of soldiers of the Federal Armed Forces, when driving military trucks with and without traffic, was measured. There was also a differentiation between experienced and unexperienced drivers. In the first experiment the subtest "digit span" (reproduction of digits) of the "Wechsler Adult Intelligence Scale" was used as a secondary task. Lists with 20 items of three, four and five digits had to be reproduced backward while driving. Only the lists with three and four digits differentiated significantly between driving with and without traffic ($p<0.05$) with the unexperienced military truck drivers. Due to traffic law demands, a new predictor had to be selected. Based on results from laboratory experiments continuous time estimation technique was used, in which Subjects had to produce time intervals of 20 sec. The predictor did not differentiate in respect to measures of central tendency, but all measures of variability (standard deviation, coefficient of variation, sum of differences) differentiated in accordance to workload between levels of traffic and between levels of driving experience of military truck drivers.

In the tracking task with simplified car dynamics Pfendler (1982) also used a secondary task. Subjects had to monitor a display where a pointer randomly moved between three sections. The "alarm" sections were on the left and right side and the "normal" section was in the centre of the display. Subjects had to press a key, as long as the pointer was moving in one of the alarm sections of the display. Percent of missed signals was calculated and two of three comparisons between the difficulty levels of the tracking task were significant ($p<0.01$).

References

Apenburg, E. (1986). Befindlichkeitsbeschreibung als Methode der Beanspruchungsmessung. *Psychologie und Praxis. Zeitschrift für Arbeits- und Organisationspsychologie, 30* (N. F. 4) 1, 3-14.

Bronner, R. and Karger, J. (1985). Beanspruchungs-Messung in Problemlöseprozessen - Modifikation eines Tests zur Erfassung psychischer Beanspruchung. *Psychologie und Praxis. Zeitschrift für Arbeits- und Organisationspsychologie, D-29, 4*, 173-184.

Bartenwerfer, H. (1960). *Beiträge zum Problem der psychischen Beanspruchung*. Forschungsbericht des Landes Nordrhein-Westfalen, Nr. 808.

Deinzer, R. et al. (1991). Salivary Cortisol in unexperienced parachute jumpers. *Neuroendocrinology Letters, 13*, 222.

Hellhammer, D. (1992). Variabilität der Cortisolreaktion auf psychologischen Stress in Abhängigkeit vom Sozialstatus bei jungen Männern (unpublished).

Pfendler, C. (1982). Bewertung der Brauchbarkeit von Methoden zur Messung der mentalen Beanspruchung bei Kfz-Lenkaufgaben. *Zeitschrift für Arbeitswissenschaft, 36 (8 NF) 1982/3*, 170-174.

Pfendler, C. (1993). Vergleich der Zwei-Ebenen Intensitäts-Skala und des NASA Task Load Index bei der Beanspruchungsbewertung während Lernvorgängen. *Zeitschrift für Arbeitswissenschaft, 47 (19 NF) 1993/1*.

Pitrella, F.D. and Käppler, W.-D. (1988). *Identification and evaluation of scale design principles in the development of the Sequential Judgement, extended range Scale*. Wachtberg: Forschungsinstitut für Anthropotechnik, Report Nr. 80.

Rummel, B. (1992a). *Begleitende psychologische Untersuchung Langzeitfahrt U22*. Kronshagen, Schiffahrtsmedizinisches Institut der Marine. Report 11/92.

Rummel, B. (1992b). Entwicklung einer kognitiven Testbatterie zur Evaluation von Umweltbedingungen. In Der Bundesminister der Verteidigung, P II 4 (Ed.), *Untersuchungen des Psychologischen Dienstes der Bundeswehr 1991/1992*. München: Verlag für Wehrwissenschaften.

Wiegand, D. (1974). Die quantitative Messung der psychischen Beanspruchung während des Fahrens durch eine simultane Nebentätigkeit. *Zeitschrift für experimentelle und angewandte Psychologie, XXI*, 679-690.

Wiegand, D. (1988). Metrological indication of strain under field condition. In H. Aschenbrenner, and D. Wiegand (Eds.), *Proceedings of the Workshop on Psychological Fitness*, Document DS/A/DR89/107 (pp. 107-115). Brussels: NATO-Defence Research Group.

Wiegand, D. (1989). Measuring the mental workload during task related activities including driving a vehicle by means of concurring time interval estimates. In M. Lind and E. Hollnagel (Eds.), *Eighth European Annual Conference on Human Decision Making and Manual Control* (pp. 64-75). Lingby: Technical University of Denmark, Institute of Automatic Control Systems.

3 Workload Measurement Methods

3.1 Workload rating scales

When workload is analysed workload rating scales are utilized more often than all other measurement methods. Workload rating scales are also often used as criteria against which other measurement methods are judged (Hart and Wickens, 1990). The reason is that, "if the person feels loaded and effortful, he is loaded and effortful, whatever the behavioural and performance measures show" (Johannsen et al.,1979). Furthermore, workload rating scales in general have many other advantages compared to other workload measurement methods. That is why they can be recommended especially for test and evaluation and why they are discussed more in detail.

Workload rating scales can be classified according to the level of measurement: nominal, ordinal, interval and ratio scale level (Lysaght et al.,1989). Most rating scales have ordinal or interval scale level. On an ordinal scale

level non-parametric statistics must be used, whereas on interval scale level the parametric methods can be applied, which better use information in the experimental data. Another classification comes from Hart and Wickens (1990), who differentiate between unidimensional, multidimensional and hierarchical rating scales. Unidimensional rating scales (e.g., Overall Workload (OW) Scale, Hill et al., 1992) measure only one dimension and give an overall workload rating. They provide no diagnostic information about the sources of workload (Hart and Wickens, 1990). They can also be used as a first screening test, as they can be administered very easily and do not need much time. Therefore, unidimensional ratings can also be applied easier while performing a task.

To get diagnostic information one has to use multidimensional ratings of workload (e.g., SWAT, NASA-TLX), which have several subscales and can detect the workload drivers. One disadvantage of multidimensional scales is, that the more dimensions they have, the more difficult it is to use the ratings while performing a task, because ratings can interfere with the main task.

Hierarchical rating scales (e.g., Modified Cooper-Harper Scale; Wierwille and Casali, 1983) have the advantage that the rating is done in a step by step procedure with a decision tree, which makes the evaluation process easier. Their disadvantage is, to give no diagnostic information (Hart and Wickens, 1990). But despite of all these differences, Lysaght et al. (1989) came to the conclusion that in all studies with empirical comparisons of rating scales in the same task, there is a good correspondence between the methods and they show in general the same rank orders in respect to the difficulty levels of the tasks.

The advantages common to most workload rating scales are, e.g.:

- High validity: in studies comparing different classes of workload measurement methods, workload rating scales often show the highest validity (e.g., Hicks and Wierwille, 1979; Pfendler, 1982; Schick and Radtke, 1979).
- High face validity.
- Minimal costs. Data may be recorded by observer, video/audio tape, keyboard entry, or paper and pencil by the subject.
- Easy to implement.
- Can be applied to a wide range of tasks (Hart and Wickens, 1990).
- Non-intrusive when used after task completion.
- Can be used prognostically in the developmental phase of a system.
- When safety is threatened in an operational system, videos of a mission can be rated, or operator workload can be evaluated by an observer.

The general disadvantages associated with most workload rating scales are:

- Rating scale results can be influenced by characteristics of respondents, like biases, response sets, errors and pretest attitudes (Dyer et al., 1976). According to Dyer et al. "a bias is a tendency to favour a certain position or conclusion; or an attitude either for or against a certain unproved hypothesis which prevents an individual from evaluating the evidence correctly. A response set or response bias refers to the tendency of a respondent to answer questions in a particular way almost independent from the content of the question. And an error is simply a mistake or departure from correctness". Pretest attitudes are e.g., beliefs and opinions. Although respondents´ characteristics might influence rating results, they might invalidate results only in extreme situations. Furthermore, it is possible to work against such tendencies by appropriate instructions which inform the rater about such pitfalls.
- The use of workload rating scales is not recommended, when it is likely that raters would fake results, have low motivation, have prejudices, guess randomly, etc.
- Between rater variability is relatively high (Hart and Wickens, 1990).
- Raters have to be familiarized or trained with the scales before collecting ratings.

- Ratings do not provide continuous measurements of workload, only periodic measurements at best.
- Negative memory effects are possible. Position effects of high workload events within low workload tasks are discussed. In an experiment it was demonstrated, that rating values increased steadily, when a workload peak approached the end of the task (Thornton, 1985).

To assure proper usage of the workload rating scale methods it is important to consider some general guidelines for their use in test and evaluation. For this purpose test and evaluation has been structured into five different phases:

a) Planning/test design:

- The rating should be made immediately after exposure to each condition rather than after exposure to several or all conditions to reduce dependence on memory.
- Special attention is to devote to the experimental design. Order effects should be avoided by counterbalancing etc. and confounding effects have to be controlled (Wierwille and Casali, 1983).
- The selection of the experimental conditions to be rated should include a good range of the workload dimension even if it is necessary to include items that need not be tested so as to give every rater a common external reference of the dimension. For example, when planning the testing of the handling qualities of 12 different military light trucks with different payloads (Käppler et al., 1988), three reference vehicles representing a range from bad to average to good handling characteristics were included and exposed to all rater-drivers prior to data collection.
- When ratings are administered while performing a task, e.g., after the segments of a mission, one should take care that the ratings are applied after critical workload periods, to prevent task interference and safety problems in operational systems. The more subscales a method has, the more likely interference occurs. It is better to make event related workload measurements than measurements at fixed time intervals to prevent interference. Furthermore, subjects should have a clear understanding of measurement intervals. Intervals should be easily identifiable and mutually exclusive. Recommended length of measurement intervals is 2-15 minutes. The subjects should be alerted when the ratings have to be done.
- In a between-subjects design, raters can not directly compare conditions and make relative or comparative ratings because they are only exposed to one condition. Such raters should be given the means to compare stimuli for each condition to a common frame of reference, i.e., a rather large range of a common reference stimuli dimension in a different but similar task in a pilot experiment preceding the main experiment.

b) Test preparation:

- It is very important that raters are familiarized and trained with the rating scale before the experiment. Subjects should at least perform some tasks varying from easy to difficult and rate these tasks in respect to workload in order to get familiarized with the rating scale. The other extreme is calibration, in which the experimenter presents to the subjects tasks varying in difficulty and shows the corresponding scale values. As the amount of training is not specified there is an influence of the tester on experimental results of rating scales. So, objectivity is limited in the data collection phase. Running a pilot experiment to establish the necessary degree of training can avoid this problem. Calibration was done, e.g., in a tank gunnery control task (Krüger and Gärtner, 1992). Before testing the difficulty of tracking targets from a simulated tank with 3 different control dynamics, subjects were exposed to a tracking task with a range of 13 task difficulty levels ranging from extremely easy to impossible to control. To provide a baseline the subjects were "calibrated" by presenting the easiest, most difficult and medium difficulty task levels and associating those difficulty levels to the ends and middle of the rating scale. This was also done to satisfy the requirements of a between-subject design (see under "Planning").
- Raters have to be familiarized and trained with the tasks before a rating is given. These test conditions have to be repeated in the data collection phase and the ratings should be given immediately after each test condition. For example, in a test of truck handling (Käppler et al., 1988) all drivers first practised a double lane change manoeuvre with all vehicles and load configurations before giving a rating. This had the effect of not only familiarizing drivers with each vehicle they had to drive and the manoeuvre to be performed, but also gave

them an impression of the range of vehicle handling qualities to be rated later. When a system is new for subjects, familiarisation is not sufficient and a systematic training must be conducted before data collection, to prevent that learning effects invalidate the results.

- The subjects should be informed about the most important biases, prejudices, etc. which can influence workload rating scale results in order to avoid any influence as far as possible.

c) Data collection :

- Data may be recorded by observer, video/ audio tape, keyboard entry, or paper and pencil by the subject.
- Subjects should be alerted, when ratings have to be given, especially when ratings have to be made during task completion.
- Assure that the tasks and stimuli are presented properly to subjects in the order specified by the experimental design and that ratings are made by each subject after each stimulus. Identify each rating response correctly in terms of stimuli, subject, group, etc. Computerisation can facilitate this task and assure a proper and consistent presentation of stimuli as well as the rating scale itself.

d) Data reduction :

- When making the statistical analysis the measurement level of the workload measurement method used has to be considered. Most ratings give ordinal or interval data and the appropriate nonparametric or parametric tests have to be selected.
- In addition to the statistical tests, it is of interest to measure the reliability and validity of the scale, discrimination or sensitivity between stimuli, discriminal dispersions around scale points, and any evidence of biases in order to get information about the quality of measurement.

e) Data interpretation:

- No norms are available for data interpretation. The ranges of acceptable or inacceptable workload are not known (Hart and Wickens, 1990). Only some coarse rules of thumb exist for SWAT and the Bedford Scale. That means that objectivity in data interpretation is limited. Usually results are interpreted by comparing test conditions, mission segments, etc.
- To compare rating results from different experiments is critical, as the ratings are relative and the experimental conditions of one experiment have an important influence on the frame of reference for evaluation. That is why the frame of reference can change from experiment to experiment.
- Measurement level of the workload rating scale method must be considered when comparing the results of different experimental conditions. Ordinal data, e.g., do not allow to say: System A is 50% lower in workload than system B.
- Dissociation between different rating scale methods do not necessarily mean that one method is better than the other, it can also mean that they measure different aspects of workload.

In the following chapters specific workload rating scales which are recommended for test and evaluation are described and critically analyzed. The description and analysis of each method is organised into 14 sections (including the references) so that the methods can be evaluated comparatively. The sections are sometimes overlapping in contents, as it was tried to summarize all relevant aspects in each section without reference to others. After the description of a method (1) important specific factors, which have to be considered in addition to the general factors in the different phases of testing and evaluation to assure proper usage of this workload measurement method are mentioned (2). After these recommendations the specific advantages and disadvantages of a workload measurement method are reflected (3, 4). General advantages and disadvantages mentioned above are only repeated in the specific descriptions of the methods, when it seemed to be especially important. In the following sections it is discussed to which extend the specific method fulfils a number of criteria, which are important preconditions for successful workload measurement (5-13). On the basis of these criteria workload measurement methods can be evaluated, compared and selected for test and evaluation.

These evaluation criteria are:

- Independence of the method from tester influence: In respect to data collection, data reduction, and interpretation of the results (5)
- Validity of the method: The degree to which the method measures what it is intended to measure (6)
- Reliability of the method: The degree of consistency and repeatability of scores obtained by the measurement method (7)
- Feasibility of the method: The ability of the method to satisfy implementation demands (8)
- Economy of the method: The monetary costs of the method (9)
- Face Validity of the method: The degree to which the method appears to measure workload to the non-specialist (10)
- Interference of the method: The degree of non-intrusiveness of the method in respect to the main task (11)
- Diagnosticity of the method: The ability of the method to differentiate among different sources of workload (12)
- Generality of the method: The degree of applicability of the method to a wide variety of tasks (13)

These criteria could have been evaluated by rating scales. But for this purpose an appropriate sample of expert raters would have been required as a considerable between subject variability was expected. So, simply the empirical evidence from the literature is presented or conclusions have been drawn from other available information about the methods.

References

Dyer, R.F., Matthews, J.J., Wright, C.E., and Yudowitch, K.L. (1976). *Questionnaire construction manual*. U. S. Army Research Institute for the Behavioral and Social Sciences, Fort Hood, Texas.

Hart, S.G. and Wickens, C.D. (1990). Workload assessment and prediction. In H.R. Booher, (Ed.), *MANPRINT: An approach to systems integration* (pp. 257- 296). New York: Van Nostrand Reinhold.

Hicks, Th.G. and Wierwille, W.W. (1979). Comparison of five mental workload assessment procedures in a moving-base driving simulator. *Human Factors, 21(2)*, 129-143.

Hill, S., Iavecchia, H., Byers, J., Bittner, A.C., Zaklad, A.L., and Christ, R.E., (1992). Comparison of four subjective workload rating scales. *Human Factors, 34 (4)*, 429-439.

Johannsen, G., Moray, N., Pew, R., Rasmussen, J., Sanders, A., and Wickens, C. (1979): Final report of experimental psychology group. In N. Moray (Ed.), *Mental workload, it´s theory and measurement*. New York. Plenum Press, 101-114.

Käppler, W.-D., Pitrella, F.D., and Godthelp, H. (1988). *Psychometric and performance measurement of light weight truck handling qualities*. Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 77.

Krüger, W. und Gärtner, K.-P. (1992). *Untersuchung verschiedener Bediensignalkennlinien für Richtschützen unter Beschleunigungsstörungen*. Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 97.

Lysaght, R.J., Hill, S.G., Dick, A.O., Plamondon, B.D., Linton, P.M., Wierwille, W.W., Zaklad, A.L., Bittner, A.C., and Wherry, R.J. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies* (Technical Report 851). Fort Bliss, TX: U.S. Army Research Institute, Field Unit.

Pfendler, C. (1982). Bewertung der Brauchbarkeit von Methoden zur Messung der mentalen Beanspruchung bei Kfz-Lenkaufgaben. *Z. Arb. wiss*. 36 *(8 NF)* 1982/3, 170-174.

Pitrella, F.D. and Käppler, W.-D. (1988). *Identification and evaluation of scale design principles in the development of the Sequential Judgement, extended range Scale*. Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 80.

Schick, F.V. and Radtke, H. (1979). *Untersuchung der Pulsfrequenzvariabilität als Schätzgröße der Pilotenbeanspruchung bei anthropotechnischen Experimenten*. DFVLR - FB 79-33.

Thornton, D.C. (1985). An investigation of the "Von Restorff" phenomenon in post-test workload ratings. In *Proceedings of the Human Factors Society 29th Annual Meeting* (pp. 760-764). Santa Monica, CA: Human Factors and Ergonomics Society.

Wierwille, W.W. and Casali, J. G. (1983). A validated rating scale for global mental workload measurement applications. *In Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 129-133). Santa Monica, CA: Human Factors and Ergonomics Society.

3.1.1 NASA Task Load Index (NASA-TLX)

(NASA, 1986; Hart and Staveland, 1988)

Description of the method

NASA TLX was derived from the NASA Bipolar-Rating Scale which has ten subscales (Hart and Staveland, 1988). Significant for the selection of the six subscales of NASA-TLX was the sensitivity of the subscales, the statistical independence from other subscales and the subjective importance regarding the individual concepts of workload (Lysaght et al., 1989). NASA recommends the application of NASA-TLX and not of the Bipolar Rating Scale.

NASA-TLX is based on the assumption that workload is a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance. Workload emerges from the interaction between the requirements of a task, the circumstances under which it is performed and the skills, behaviours, and perceptions of the operator (Hart and Staveland, 1988). Furtheron, workload is regarded as a multidimensional construct so that consequently a workload measurement method must be made up of multiple workload dimensions. Another aspect which was important in the development of NASA-TLX was the observation that subjects showed great interindividual differences regarding the importance of the different workload dimensions, when rating tasks. That is why the authors developed an individual weighting procedure for the different workload dimensions, which is assumed to reduce between subject variability of the results (Lysaght et al., 1989).

The six subscales can be divided into three groups. Characteristics of the task: Mental, Physical and Time Demands. Behavioural characteristics: Performance and Effort. Individual characteristic: Frustration. Each of the bipolar subscales of NASA-TLX consists of 20 five point steps from 0-100, the endpoints having verbal descriptors. The NASA-TLX subscale definitions (Hart and Staveland, 1988) are given in Figure 1.

There are two steps when using NASA-TLX: In the first phase the subject rates the task with regard to workload. The scales are scored as follows: if the subject places his/her marks precisely on one of the tic-marks then the value of that tic mark is used, but if the subject´s mark is between two tic-marks then the value of the right tic-mark is used (NASA, 1986). In the second phase each subject makes a paired comparison, deciding with all 15 possible pair combinations of the 6 dimensions which pair element is more important with regard to workload in the rated task. From the results a rank order of the dimensions from 0-5 is derived by which the individual subscale scores of the rated task are weighted. By summing up the weighted subscale scores and dividing them by the sum of the weights (= 15), the Mean Weighted Workload Score is obtained, that indicates workload in per cent (Table 1). If two or more qualitatively different tasks are rated with regard to workload, a separate paired comparison has to be made for each task (NASA 1986). There are no interpretation supports (e.g. about tolerance levels, etc.) available.

NASA-TLX requires an adequate familiarization of the subject with the method. Before event scoring subjects should use the rating scale on a few tasks (NASA, 1986). With sufficient training the subject needs approximately one minute for event scoring and about three minutes for paired comparison.

The weighting procedure of NASA-TLX was criticised by different authors. In a study of Pfendler (1991) with a German version of NASA-TLX a better differentiation and better reliability were achieved with a Mean Unweighted Workload Score of the six scales than with the Mean Weighted Workload Score. An increase in between subject variability could not be noticed in contrast to the assumptions of the authors of NASA-TLX. Unweighted and weighted workload scores correlated with r = 0.94. From the results the conclusion can be drawn that the paired comparison can possibly be dropped. For further generalisation the results have to be verified with a broader range of tasks. Similar arguments come from Nygren (1991), who describes the weighting procedure of NASA-TLX as ineffective and recommends to ignore it simply.

Title	**Endpoints**	**Descriptions**
MENTAL DEMAND	Low/High	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?
PHYSICAL DEMAND	Low/High	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?
TEMPORAL DEMAND	Low/High	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?
PERFORMANCE	good/poor	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?
EFFORT	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
FRUSTRATION LEVEL	Low/High	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Figure 1: NASA-TLX rating scale definitions (Hart and Staveland, 1988)

Table 1: Scoring of NASA-TLX (fictitious example); underlined scale rated as more important by the Subject.

Paired Comparison:			Computation of the mean weighted workload score:			
	Weighting		Subscale	Subscale Rating Score	Weight	Weighted Subscale Rating Score
EF-OP	TD-MD	MD=3	Mental Demand (MD)	25 Completed	3	75
TD-FR	FR-EF	PD=0	Physical Demand (PD)	5	0	0
TD-EF	OP-MD	TD=5	Temporal Demand (TD)	65	5	325
PD-FR	MD-EF	OP=4	Performance (OD)	50	4	200
OP-FR	OP-TD	EF=2	Effort (EF)	50	2	100
PD-TD	MD-PD	FR=1	Frustration (FR)	25	1	25
PD-OP	EF-PD					
FR-MD				Total		725
				Weights		15
				Mean Weighted Workload Score		48

Factors to consider in the phases of test and evaluation
(General factors: see introduction)

- Using the six subscales during task performance might cause interference with the main task and influence safety in an operational system.
- Additional time is needed to make the paired comparison.
- When using NASA-TLX with different tasks, the paired comparison must be repeated.
- It is favourable to have programs for scoring of raw data.
- Parametric tests can be used for data analysis, as NASA-TLX data are assumed to have interval scale level.

Advantages of the method
(General advantages of workload rating scales: see introduction)

- NASA-TLX has a good face validity, and thus a good acceptance by the subjects (Hill et al.,1992).
- NASA-TLX can be applied with a very broad range of tasks in contrast to other workload measurement methods (e.g. secondary tasks) (Hart and Wickens, 1990; see also below).
- The authors assume that NASA-TLX has interval scale level.
- The American version of NASA-TLX proved to be more valid than workload measurement methods like SWAT and the Modified Cooper-Harper Scale (Byers et al., 1988; Battiste and Bortolussi, 1988; Hancock et al., 1989).
- NASA-TLX can also be employed in a prognostic way during the concept development phase, if enough details about the system are known (Beevis, 1992).
- Because of the weighting procedure NASA-TLX is assumed to have less between subject variability than other subjective methods (Hart and Wickens, 1990), but evidence is contradictory (e.g., Pfendler, 1993).

- Unlike with SWAT there are no floor effects with NASA-TLX during low workload situations (Battiste and Bortolussi, 1988).
- Dutch, English and German versions of the scale are available.

Disadvantages of the method

(General disadvantages of workload rating scales: see introduction)

- Translated versions of NASA-TLX (German and Dutch) showed lower sensitivity than the Sequential Judgement Scale from Pitrella and Käppler (1988) (Pfendler, 1993), and a Dutch effort scale (BSMI)(Veltman and Gaillard, 1993). The result might be explained by different connotations of the subscales of NASA-TLX in other languages.
- The original instructions are too extensive (Veltman and Gaillard, 1993) and use partly technical terms which are not comprehensible to every subject.
- The six subscales of NASA-TLX partly show significant intercorrelation (Pfendler and Widdel, 1988).
- Because of its relatively high number of dimensions to be rated, NASA-TLX is less suitable for simultaneous workload evaluations during task completion.
- In some experiments the subscales "frustration" and "physical demands" only show a small relevance for workload (Pfendler and Widdel, 1988; Sepehr, 1988; Veltman and Gaillard, 1993).
- One subscale can be dropped because of the weighting procedure. When two different subscales are eliminated by two different subjects the Mean Weighted Workload Scores of these two subjects are strictly seen no more comparable. Suggestions for improvement with weights from 1-6 instead of 0-5 have been made by Sepehr (1988).

Independence of the method from tester influence

The use of NASA-TLX requires a training of the subject with the method that is not specified more precisely. Therefore, certain limitations concerning objectivity during data collection phase exist, while the objectivity of scoring is given. The objectivity of data interpretation is also limited as there are no norms and tolerance levels available (like with all other workload measurement methods).

Validity of the method

From the many workload studies using NASA-TLX some results relevant for validity are presented:

- Operator Workload was evaluated with four subjective workload rating scales in three weapon systems (remotely piloted vehicle, helicopter, air defence system). A factor analysis supplied one workload dimension with NASA-TLX showing the highest factor loading ($r = 0.899 - 0.942$) (Hill et al., 1992).
- In a flight simulator piloting tasks with different levels of complexity were carried out for the evaluation of cockpit displays. Significant differences were obtained with NASA-TLX and SWAT but not with physiological measurement methods (Nataupsky and Abbott, 1987).
- Flights with and without system failures were carried out in a Boeing 727 flight simulator. NASA-TLX and SWAT were sensitive to differences between these high and low workload flights and to differences among flight segments. NASA-TLX but not SWAT was sensitive to the increase in workload during the cruise segment of the high workload flight (Battiste and Bortolussi, 1988).
- Workload was measured with NASA-TLX and SWAT while learning a tracking task and a significant reduction of workload was observed during the course of learning. NASA-TLX supplied a better differentiation than SWAT (Hancock et al., 1989).
- When learning a pattern detection task a significant workload reduction was noticed with NASA-TLX (Pfendler, 1993). The Sequential Judgement Scale (Pitrella and Käppler, 1988) used in the same experiment was slightly superior in validity. A better differentiation between the learning blocks and a higher validity

coefficient was found with NASA-TLX, when using a Mean Unweighted Workload Score instead of the Mean Weighted Workload Score.

- In a visual pattern recognition task with four different degrees of complexity, two significant workload differences could be found between the two most simple and the most complex pattern (Pfendler and Widdel, 1988).
- When measuring workload of fighter pilots in a fixed base flight simulator, a unidimensional Dutch effort scale (BSMI) proved to have higher sensitivity than NASA-TLX in an intercept and a navigational scenario. The differences between flying curves and flying straight were only significant for the BSMI (Veltman and Gaillard, 1993).
- In experimental comparisons high correlation between NASA-TLX and other rating scales could be measured (Lysaght et al., 1989).

Reliability of the method

Vidulich and Tsang (1987) found average test-retest coefficients of r = 0.42 for the Mean Weighted Workload Score. With a German version of the scale Pfendler and Widdel (1988) found intraclass coefficients of r = 0.84 by analysis of variance. With the same method Pfendler (1991) found an intraclass coefficient of r = 0.56, being lower than that of the Sequential Judgement Scale (r = 0.87) (Pitrella and Käppler, 1988), which was compared to NASA-TLX in this experiment. When using the Mean Unweighted Workload Score instead of the Mean Weighted Workload Score the intraclass coefficient of NASA-TLX could be improved up to r = 0.750 without increasing between subject variability. According to Battiste and Bortolussi (1988) NASA-TLX seems to have higher reliability than SWAT (r = 0.769 vs. r = 0.751)

Feasibility of the method

NASA-TLX requires only little instrumentation and can easily be implemented in any environment without limitations.

Economy of the method

NASA-TLX has low costs in respect to material and personnel. For scoring of NASA-TLX data it is advantageous to have the appropriate programs. NASA-TLX needs, like all subjective procedures, a certain amount of training for the subjects and the experimenter.

Face validity of the method

NASA-TLX has good face validity. In an experiment comparing NASA-TLX, SWAT, Overall Workload Scale and Modified Cooper-Harper Scale, NASA-TLX proved to have the highest face validity in respect to workload measurement (Byers et al., 1988).

Interference of the method

There is no interference with the main task, when NASA-TLX is administered after task completion. Only when evaluating workload while performing the main task, interference might exist, as it takes some time to make six ratings. But that depends on the difficulty level of the task and the familiarity with NASA-TLX. When system safety is an important aspect, a possible solution to this problem might be to use postmission subjective ratings that were supported by videotapes of the mission (Wierwille and Eggemeier, 1993).

Diagnosticity of the method

NASA-TLX is assumed to have acceptable diagnosticity because it uses 6 subscales, which can be analysed separately. Nevertheless there are limitations, as a study with a German version has shown, that more than 50% of all possible intercorrelation between the subscales are significant and some subscales (Frustration Level, and Physical Demand) might be not so relevant (Pfendler and Widdel, 1988). On the other side, the weights given to each subscale provide additional diagnostic information (Hart and Wickens, 1990).

Generality of the method

NASA-TLX can be used in basic and applied research, in the field of ergonomics, industrial psychology, and man-machine-communication. NASA-TLX has been applied in environments ranging from the laboratory to military and civilian helicopter, general aviation, transport, and military jet simulators and aircraft, and in ground-based systems (Hart and Wickens, 1990). German versions of NASA-TLX were used in the laboratory in visual pattern recognition tasks (Pfendler and Widdel, 1988), and in pattern detection tasks (Pfendler, 1991). Sepehr (1988) applied a German version to error-detection in program-text, to driving tasks in the simulator and in the real system, to tracking tasks and to piloting tasks in a flight simulator.

References

Battiste, V. and Bortolussi, M. (1988). Transport pilot workload: A comparison of two subjective techniques. In *Proceedings of The Human Factors Society 32nd Annual Meeting* (pp. 150-154). Santa Monica, CA: Human Factors and Ergonomics Society.

Byers, J.C., Bittner, A.C., Hill, S.G., Zaklad, A.L., and Christ, R.E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 1145-1149). Santa Monica, CA: Human Factors and Ergonomics Society.

Beevis, D. (1992). *Analysis techniques for man-machine systems design*. Report AC/243 (Panel 8) TR/7 Vol. 2. Brussels: NATO Defence Research Group.

Hancock, P.A., Robinson, M.A., Chu, A.L., Hansen, D.R., and Vercruyssen, M. (1989). The effect of practice on tracking and subjective workload. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1310-1314). Santa Monica, CA: Human Factors and Ergonomics Society.

Hart, S.G. and Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.*), Human mental workload* (pp. 139 - 183). Amsterdam: North-Holland.

Hart, S.G. and Wickens, C.D. (1990). Workload assessment and prediction. In H.R. Booher (Ed.), *MANPRINT. An approach to systems integration* (pp. 257-296). New York: Van Nostrand Reinhold.

Hill, S., Iavecchia, H., Byers, J., Bittner, A.C., Zaklad, A.L., and Christ, R.E., (1992). Comparison of four subjective workload rating scales. *Human Factors*, 34 (4), 429-439.

Lysaght, R.J., Hill, S.G., Dick, A.O., Plamondon, B.D., Linton, P.M., Wierwille, W.W., Zaklad, A.L., Bittner, A.C., and Wherry, R.J. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies* (Tech. Report 851). Fort Bliss, TX: U.S. Army Research Institute, Field Unit.

NASA Task Load Index (TLX): Computerized version, (1986). Moffet Field, CA: NASA-Ames Research Center, Aerospace Human Factors Research Division.

NASA Task Load Index (TLX): Paper and pencil version, (1986). Moffet Field, CA: NASA-Ames Research Center, Aerospace Human Factors Research Division.

Nataupsky, M. and Abbott, T.S. (1987). Comparison of workload measures on computer-generated primary flight displays. In *Proceedings of The Human Factors Society 31st Annual Meeting* (pp. 548-552). Santa Monica, CA: Human Factors and Ergonomics Society,

Nygren, TH. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors, 33 (1)*, 17-33.

Pfendler, C. (1991). *Vergleichende Bewertung der NASA-TLX Skala und der ZEIS-Skala bei der Erfassung von Lernprozessen*. Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 92.

Pfendler, C. (1993). Vergleich der Zwei-Ebenen Intensitäts-Skala und des NASA Task Load Index bei der Beanspruchungsbewertung während Lernvorgängen. *Z. Arb. wiss. 47 (19 NF) 1993/1*, 26-33.

Pfendler, C. and Widdel, H. (1988). *Gedächtnisleistung und Beanspruchung beim Wiedererkennen von farbigen und schwarzweißen Reizmustern auf elektronischen Anzeigen*. Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 81.

Pitrella, F.D. and Käppler, W.-D. (1988). *Identification and evaluation of scale design principles in the development of the Sequential Judgement, extended range Scale*. Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 80.

Sepehr, M.M. (1988). Assessment of subjective mental workload using NASA-Task Load Index. In *Proceedings of the 7th European Annual Conference on Human Decision Making and Manual Control* (pp. 69-75). Paris, Eléctricité de France, Clamart Cedex, France.

Veltman, J.A. and Gaillard, A.W.K. (1993). Indices of mental workload in a complex task environment. *Neuropsychobiology, 28*, 72-75.

Veltman, J.A. and Gaillard, A.W.K. (1993). Measurement of pilot workload with subjective and physiological techniques. In *Proceedings of the Workload Assessment and Aviation Safety Conference*. Royal Aeronautical Society, April 27, 28.

Veltman, J.A. and Gaillard, A.W.K. (1993). *Evaluation of subjective and physiological measurement techniques for pilot workload*. Rapport, IZF 1993 A-5.

Vidulich, M.A. and Tsang, P.S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. In *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 1057-1061). Santa Monica, CA: Human Factors Society.

Wierwille, W.W. and Eggemeier, F.Th. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors, 35 (2),* 263-268.

3.1.2 Subjective Workload Assessment Technique (SWAT)

(Armstrong Aerospace Medical Research Laboratory, 1987; Reid et al., 1981a; Reid et al., 1989)

Description of the method

SWAT is a subjective workload measurement method in which the subjects rate the workload of a task on the basis of the dimensions of "time load", "mental effort load", and "psychological stress load". The SWAT dimensions are derived from a workload definition from Sheridan and Simpson (1979). The method uses conjoint measurement and scaling (Eggemeier et al., 1982; Krantz and Twersky, 1971) and the ratings on these dimensions, which are made on an ordinal scale level can be combined into one overall workload score on an interval scale level. Each of the three subscales has three levels in the form of a category scale with verbal descriptors that outline each dimension (Reid and Nygren,1988):

I. Time Load

1. Often have spare time. Interruptions or overlap among activities occur infrequently or not at all.
2. Often have spare time. Interruptions or overlap among activities occur infrequently.
3. Almost never have spare time. Interruptions or overlap among activities are very frequent, or occur all the time.

II. Mental Effort Load

1. Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention.
2. Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention required.
3. Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.

III. Psychological Stress Load

1. Little confusion, risk, frustration, or anxiety exists and can be easily accommodated.
2. Moderate stress due to confusion, frustration, or anxiety noticeably adds to workload. Significant compensation is required to maintain adequate performance.

3. High to very intense stress due to confusion, frustration, or anxiety. High to extreme determination and self-control required.

There is also a German Version of SWAT (Schick et al., 1989) and furthermore PRO-SWAT (Projective-SWAT) can be used prognostically with systems which are still in the design phase (Kuperman, 1985; Kuperman and Wilson, 1985; Beevis, 1992).

When applying SWAT one can distinguish two phases: scale development phase and event-scoring phase. In scale development the subject has to rank order all 27 combinations of the three levels of each of the three dimensions according to her individual perception of increasing workload. For this purpose the subject uses 27 cards with descriptions of these 27 combinations. After that, it is necessary to test with a computer program if the card sort is in accordance to a series of mathematical axioms so that conjoint measurement and scaling can be applied. But the method allows some axiom violations (Reid and Nygren, 1988). After this procedure the SWAT technique uses conjoint analysis to convert the rank data of the card sort into an interval scale from 0-100 in which each of these 27 combinations has a fixed value.

As in general individual different card sorts are given, each subject also gets an individual scale solution. If there is enough agreement between the card sorts of a group of subjects a single group scale can be constructed (Lysaght et al., 1989). This can be tested by Kendall´s Coefficient of Concordance (W). A group solution is possible if W is .75 or higher (Reid and Nygren, 1988). If W is below this value a procedure called SWAT prototyping is applied. This procedure for grouping subjects into homogeneous subgroups is assumed to provide greater precision of measurement (Reid et al., 1982). From the SWAT dimensions time load (T), mental effort load (E), and psychological stress load (S) six hypothetical prototypes (TES, TSE, ETS, EST, STE, SET) have been developed, characterized by hypothetical card sorts which place different emphasis on the three dimensions. The TES prototype, e.g., places the greatest emphasis on time, the second on effort, and third on psychological stress.

To find out which prototype corresponds best to an individual rank order, raters´ rank orderings are correlated with each of the six prototype orderings, which can be done by Spearman´s Rho (Reid and Nygren, 1988). The highest correlation indicates which prototype corresponds best. Subjects with the highest correlation in one prototype are grouped together and a common scale solution has to be found.

After scale development event scoring can take place, in which the subject evaluates the relevant task with regard to time load, mental effort load, and psychological stress load (e.g., 2, 1, 3). The scale value associated with this combination (e.g., 30.5) is the dependent variable used in the further analysis.

The original SWAT scores, based on conjoint measurement and scaling, have been compared to scores, based simply on summated ratings from the three SWAT subscales or to composite scores, derived from multivariate statistics (Biers and Masline, 1987; Biers and Mc Inerney, 1988). The results showed that the three kinds of scores were equally sensitive and highly correlated. The conclusion is drawn that workload measurement with SWAT can also be done effectively without going through scale development phase.

Factors to consider in test and evaluation
(General factors: see introduction)

- When workload assessment is done with SWAT while performing a task, be aware of interference with the main task.
- SWAT training and card sort requires about 1.5 hours per subject.
- Software to do conjoint scaling is required.
- Requires experimenter decisions to specify scale type and adequacy.
- "Redline" values have been developed (Reid and Colle, 1988). Workload problems causing performance degradation may start at scores between 30-50.
- Keep in mind that SWAT scores are interval and not ratio scale data.

Advantages of the method
(General advantages of workload rating scales: see introduction)

- Measurement on interval scale level.
- As the evaluation is very time economic, SWAT can also be used during task completion without interfering with the main task, when not applied in critical workload periods.
- SWAT can be used prognostically.
- Critical SWAT values for predicting operator overload exist (Reid and Colle, 1988)
- SWAT can be applied with a very broad range of tasks (Hart and Wickens, 1990).
- Low cost and ease of use in real time. So SWAT can be used especially when time and resources are critical.
- Published record of application in laboratory and field settings.

Disadvantages of the method
(General disadvantages of workload rating scales: see introduction)

- SWAT dimensions are derived intuitively not empirically.
- There are significant intercorrelation between all SWAT dimensions (Boyd, 1983).
- Scale development can be problematic. It is time consuming (20-60 minutes) and affords good verbal abilities. The motivation to perform the card sort is sometimes low as it is not very interesting (Lysaght et al., 1989). Nevertheless, it must be done conscientiously. When not done properly this might have severe consequences for the validity of results. The experimenter has to understand the need for card sort and "sell" it to the subjects.
- The authors are not very precise in respect to the number of axiom violations allowed.
- Although computerised, data analysis in respect to conjoint measurement and scaling is extensive in comparison to most other workload measurement methods.
- When analysing the subscales separately, the evaluation is rather coarse as the number of levels (3) in each dimension is limited.
- There are indications of low sensitivity of SWAT at low workload levels (Battiste and Bortolussi, 1988). There have been approaches to solve such problems with a five point scale, but this would increase the difficulty of scale development, as the card sort would be extended to 125 combinations.
- SWAT is less sensitive for short duration intervals.

Independence of the method from tester influence

As SWAT, like all other subjective methods, needs some training, the amount of training, which is not specified, might influence the results. Furthermore it is not exactly specified how many axiom violations in the card sort are tolerable. So the tester has some degrees of freedom to decide which card sort can be accepted or must be rejected. Finally the interpretation of the workload score is not very objective, like with other methods, but at least some "rules of thumb" exist.

Validity of the method

Overviews on validity of SWAT can be found with Reid and Nygren (1988) and Lysaght et al. (1989). According to these authors SWAT has proven to be a valid workload indicator in many laboratory experiments and in applied settings with a strong emphasis on aviation application. In the following some workload studies, demonstrating the validity of SWAT, are summarized:

- Critical tracking and simulated radio communication (Reid et al., 1981b). SWAT differentiated between all three levels of the tracking task.

- Spatial memory (Eggemeier and Stadler, 1984). Complexity of histograms (2 versus 6 bars) and length of the retention interval (16 versus 32 seconds) had a significant influence on SWAT scores. No significant influence had the spatial orientation of the histograms (rotations of 0, 90, 180 degrees).
- Short term memory (Eggemeier et al., 1982). Significant influence of the number of elements (2, 3, 4) that had to be updated in memory and of element presentation rate (interstimulus intervals of 0.5, 2.0, 3.5 and 5.0 seconds) on SWAT scores.
- Probability monitoring (Notestine, 1984). Significant differences between two difficulty levels were found.
- Tracking (Hancock et al., 1989). Workload significantly decreased when learning a tracking task. SWAT proved to have lower sensitivity than NASA-TLX.
- Criterion Task Set (CTS) testbattery (Polzella and Reed, 1987). Consists of the following tasks: Continuous recall, grammatical reasoning, language processing, mathematical processing, memory search, spatial processing, unstable tracking, probability monitoring, interval production. A multidimensional scaling analysis of SWAT ratings of the CTS test battery resulted in two orthogonal dimensions: response time and task effort.
- Simulated landing approaches (Schick et al., 1989). SWAT differentiated significantly between flight segments with and without windshear and between the first five segments of the landing approach. No effects or only small effects could be found with ECG and EMG parameters and with pilot activity.
- Simulated transport flights (Battiste and Bortolussi, 1988). SWAT differentiated between high and low workload flights and also among flight segments. NASA-TLX was more sensitive than SWAT.
- Operation of weapon systems (remotely piloted vehicle, helicopter, air defence system) (Hill et al., 1992). Factor analysis revealed a single workload factor. NASA-TLX had higher factor loadings than SWAT.
- Aviation-like psychomotor dual-task (Kilmer et al., 1988). SWAT was more sensitive in respect to task difficulty (three levels of wind gusts) than the Modified Cooper-Harper Scale.

Reliability of the method

Reliability of card sorts was estimated in two studies. Correlation of four pilots, making the card sort before and after the main experiment, ranged from 0.77- 1.00 (Gidcumb, 1985). Also Reid and Nygren (1988) found high correlation from retests which emphasis stability of card sorts.

Reliability of SWAT in the event scoring phase was estimated by Battiste and Bortolussi (1988). They found average retest coefficients of $r = 0.751$, which were lower than those of NASA-TLX.

Feasibility of the method

The scale development phase is crucial for the scale construction. But low motivation when making the card sort and low verbal abilities can endanger this phase and consequently the successful use of SWAT. In addition, software is needed to analyse card sort results. On the other side scale development phase might be ignored (Biers and Masline, 1987; Biers and Mc Inerney, 1988). Seen away from scale development, SWAT is easy to implement. Event-scoring is very time economic and can also be done during task completion, when it is timed properly. PRO-SWAT can also be applied in the design phase of a system, when enough detailed information about the system is available (Kuperman, 1985; Kuperman and Wilson, 1985; Beevis, 1992).

Economy of the method

Card sort can be time consuming (at least 20 minutes) and, when the scale development phase is not ignored, the following data analysis is also more time consuming than with other methods and has to be done by computer. In all other respects SWAT is very inexpensive and in comparison to scale development the event scoring is performed very quickly.

Face validity of the method

In a study from Hill et al. (1992) subjects rated NASA-TLX and the Overall Workload (OW) Scale better than SWAT in respect to the ability to assess workload. The Modified Cooper-Harper Scale was rated to have the same face validity as SWAT.

Interference of the method

When SWAT is used after task completion the method can not interfere with the primary task. When SWAT is used during task completion there might be some limited interference with the primary task, but that depends on subject´s familiarity with the task and with SWAT. The interference should be lower than with NASA-TLX, as the subjects have to report only three numbers referring to the three SWAT subscales. An event related data collection appears to be a better alternative than data collection at fixed time intervals (Lysaght et al., 1989). Another possible solution, especially when safety is threatened, is to use SWAT with postmission videotapes (Corwin, 1992).

Diagnosticity of the method

As SWAT consists of three subscales a certain amount of diagnosticity can be expected, when the subscales are analysed separately. A restriction exists in so far as the subscales have significant intercorrelations (Boyd, 1983), and as each subscale has only three levels. Nevertheless Reid and Nygren (1988) found examples of good diagnosticity of SWAT. E.g., Potter and Acton (1985) observed that Mental Effort Load scale increased most at the lower levels of a continuous recall task whereas Time Load and Psychological Stress Load scales changed more at the moderate and high difficulty levels of the task. The differential sensitivity of the dimensions was also shown with another study (Potter, 1986).

Generality of the method

Seen away from the restrictions mentioned in this description of SWAT, this method can be applied to a very broad spectrum of tasks. SWAT was used with many laboratory tasks (see under "validity"). Reid and Nygren (1988) have listed the following SWAT application studies in simulation and operational environments:

Category:	**System:**
<u>Simulation</u>	
Aircraft	- F-16/F-15 Air-to-Air - KC-135 Flight Deck Modernisation - A-300 Approach and Landing (Schick and Hann, 1987) - B-52 Long Mission (Skelly and Purvis, 1985) - DC-10 Approach and Landing (Biferno and Reed, 1983) - B-52 CG/Fuel Level Advisory System - Helicopter NOE (HAWORTH et al., 1987) - General Aviation Training (Haskell and Reed, 1985)
Control Center	- Ground Launch Missile (Crabtree et al.,1984; Acton and Crabtree, 1985) - Nuclear Power Plant Training (Beare and Dorris, 1984) - Oil Refinery (Beville Engineering, Inc., 1986)
<u>Operational</u>	
Aircraft	- F-16 Flight Test* - A-10 Flight Test* - Laser Guided Missile Flight Test* (Ossard et al., 1987)
Control Center	- C-1412 Air Drop/Air Land** - KC-10 Boom Operator** (Dodge et al., 1984) - Command and Control Center** (Courtright and Kuperman, 1984)

* Flight Test

** Operational Test & Evaluation (OT&E)

References

Acton, W. H. and Crabtree, M. S. (1985). Workload assessment techniques in system redesign. In *Proceedings of the IEEE National Aerospace and Electronics Conference*.

Armstrong Aerospace Medical Research Laboratory (1987). *Subjective Workload Assessment Technique (SWAT): A users guide*. Dayton, OH: AAMRL, Wright Patterson AFB.

Battiste, V. and Bortolussi, M. (1988): Transport pilot workload: A comparison of two subjective techniques. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 150-154). Santa Monica, CA: Human Factors and Ergonomics Society.

Beare, A.N. and Dorris, R.E. (1984). The effects of supervisor experience and the presence of a shift technical advisor on the performance of two-man crews in a nuclear power plant simulator. In *Proceedings of the Human Factors Society Twenty-Eight Annual Meeting* (pp. 242-246). Santa Monica, CA: Human Factors and Ergonomics Society.

Beevis, D. (1992). *Analysis techniques for man-machine systems design*. Report AC/243 (Panel 8) TR/7 Vol. 2. Brussels: NATO Defence Research Group.

Beville Engineering, Inc. (1986): *Human Factors Analysis of Refinery Consolidation.*

Biers, D.W. and Masline, P.J. (1987). Alternative approaches to analyzing SWAT data.. In *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 63-66). Santa Monica, CA: Human Factors and Ergonomics Society.

Biers, D.W. and Mc Inerney, P. (1988). An alternative to measuring subjective workload: Use of SWAT without the card sort. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 1136-1139). Santa Monica, CA: Human Factors and Ergonomics Society.

Biferno, M. and Reid, G.B. (1983). *DC-10 study - Does the distance of a touch-panel control influence operator performance and/or workload?* (unpublished report).

Boyd, S.B. (1983). Assessing the validity of SWAT as a workload measurement instrument. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 124-128). Santa Monica, CA: Human Factors and Ergonomics Society.

Corwin, W.H. (1992). Inflight and postflight assessment of pilot workload in commercial transport aircraft using the Subjective Workload Assessment Technique. *The International Journal of Aviation Psychology 2(2),* 77-93.

Courtright, J.F. and Kuperman, G. (1984): Use of SWAT in USAF system T & E. In *Proceedings of the Human Factors Society 28th Annual Meeting* (pp. 700 - 703). Santa Monica, CA: Human Factors and Ergonomics Society.

Crabtree, M.S., Bateman, R.P., and Acton, W.H. (1984). Benefits of using objective and subjective workload measures. In *Proceedings of the Human Factors Society 28th Annual Meeting* (pp. 950-953). Santa Monica, CA: Human Factors and Ergonomics Society.

Dodge, D.C., Wong, T.J., and Brown, K.W.(1984*). Boom control system improvement study - phase II - Supplemental indication system.* Report No. MDC J9732 (Douglas Aircraft Company, McDonnell Douglas).

Eggemeier, F.T. and Stadler, M.A. (1984). Subjective workload assessment in a spatial memory task. In *Proceedings of the Human Factors Society 28th Annual Meeting* (pp. 680-684). Santa Monica, CA: Human Factors and Ergonomics Society.

Eggemeier, F.T., Crabtree, M.S., Zingg, J.J., Reid, G.B., and Shingledecker, C.A. (1982). Subjective workload assessment in a memory update task. In *Proceedings of the Human Factors Society 26th Annual Meeting* (pp. 643-647). Santa Monica, CA: Human Factors and Ergonomics Society.

Gidcumb, C. (1985). *Survey of SWAT use in flight test* (BDM/A-85-0630-TR). Albuquerque, NM: BDM Corporation.

Hancock, P.A., Robinson, M.A., Chu, A.L., Hansen, D.R. and Vercruyssen, M. (1989). The effect of practice on tracking and subjective workload. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1310-1314*)*. Santa Monica, CA: Human Factors and Ergonomics Society.

Hart, S.G. and Wickens, C.D. (1990). Workload assessment and prediction. In H.R. Booher (Ed.): *MANPRINT. An approach to systems integration* (pp. 257-296). New York: Van Nostrand Reinhold.

Haskell, B. and Reid, G.B. (1985). An investigation of the subjective perception of workload and performance in low-time private pilots. *Aviation Space and Evironmental Medicine*.

Haworth, L.A., Bivens, C.C., Shively, R.J., and Delgado, D. (1987). Advanced cockpit and control configurations for single pilot helicopter-nap-of-the-earth flight. *Paper presented at the American Helicopter Society Forty-Third Annual Forum and Technology Display* .

Hill, S., Iavecchia, H., Byers, J., Bittner, A.C., Zaklad, A.L., and Christ, R.E., (1992). Comparison of four subjective workload rating scales. *Human Factors, 34 (4),* 429-439.

Kilmer, K.J., Knapp, R., Burdsal, C. jr., Borresen, R., Bateman, R., and Malzahn, D. (1988). Techniques of subjective assessment: A comparison of the SWAT and Modified Cooper-Harper Scales. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp.155-159). Santa Monica, CA: Human Factors and Ergonomics Society.

Krantz, D.H. and Twersky, A. (1971). Conjoint measurement analysis of composition rules in psychology. *Psychological Review, 78,* 151-169.

Kuperman, G.G. (1985). Pro-SWAT applied to advanced helicopter crewstation concepts. In *Proceedings of the Human Factors Society 29th Annual Meeting* (pp. 398-402). Santa Monica, CA: Human Factors and Ergonomics Society.

Kuperman, G.G. and Wilson, D.L. (1985). A workload analysis for strategic conventional standoff capability missions. In *Proceedings of the Human Factors Society 29th Annual Meeting* (pp.635-639). Santa Monica, CA: Human Factors and Ergonomics Society.

Lysaght, R.J., Hill, S.G., Dick, A.O., Plamondon, B.D., Linton, P.M., Wierwille, W.W., Zaklad, A.L., Bittner, A.C., and Wherry, R.J. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies* (Tech. Report 851). Fort Bliss, TX: U.S. Army Research Institute, Field Unit..

Masline, P.J. and Biers, D.W. (1987). An examination of projective versus post-task subjective workload ratings for three psychometric scaling techniques. In *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 77-80). Santa Monica, CA: Human Factors and Ergonomics Society.

Notestine, J.C. (1984). Subjective workload assessment and effect of delayed ratings in a probability monitoring task. In *Proceedings of the Human Factors Society 28th Annual Meeting* (pp.685-689). Santa Monica, CA: Human Factors and Ergonomics Society.

Ossard, G., Amalberti, R., and Poyot, G. (1987). *Evaluation de la charge de travail du pilote induite par un systeme d´arme guide laser,* (Ministère de la Défense: Centre d'Etudes et de Recherches de Médecine Aerospatiale, Laboratoire d'Etudes Médicophysiologiques 16/330 .

Polzella, D.J. and Reid, G.B. (1987). A multidimensional scaling analysis of Subjective Workload Assessment Technique (SWAT) ratings of the Criterion Task Set (CTS). In *Proceedings of the Human Factors Society 31st Annual Meeting* (pp.398-401*)*. Santa Monica, CA: Human Factors and Ergonomics Society.

Potter, S.S. (1986). *Subjective Workload Assessment Technique (SWAT) subscale sensitivity to variations in task demand and presentation rate*. Unpublished Master´s Thesis, Wright State University, Dayton, Ohio.

Potter, S.S. and Acton, W.H. (1985). Relative contributions of SWAT dimensions to overall subjective workload ratings. In *Proceedings of the Third Symposium on Aviation Psychology* (pp. 231-238). Columbus, Ohio, Ohio State University.

Reid, G.B., Potter, S.S., and Bressler, J.R. (1989). *Subjective Workload Assessment Technique (SWAT): A user's guide* (Tech. Report AAMRL-TR-89-023). Wright-Patterson Air Force Base, OH: USAF Armstrong Laboratory.

Reid, G.B., Eggemeier, F.T., and Nygren, T.E. (1982). An individual differences approach to SWAT scale development. In *Proceedings of the Human Factors Society 26th Annual Meeting* (pp. 639-642). Santa Monica, CA: Human Factors and Ergonomics Society.

Reid, G.B. and Colle, H.A. (1988). Critical SWAT values for predicting operator overload. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp.1414-1418). Santa Monica, CA: Human Factors and Ergonomics Society.

Reid, G.B. and Nygren, T.E. (1988). The Subjective Workload Assessment Technique. In: P.A. Hancock and M. Meshkati (Ed.): *Human mental workload* (pp. 185-218). Amsterdam: North- Holland.

Reid, G.B., Shingledecker, C.A., Nygren, T.E., and Eggemeier, F.T. (1981a). Development of multidimensional subjective measures of workload. In *Proceedings of the IEE International Conference on Cybernetics and Society* (pp. 403-406).

Reid, G.B., Shingledecker, C.A., and Eggemeier, F. T. (1981b). Application of conjoint measurement to workload scale development. In *Proceedings of the Human Factors Society 25th Annual Meeting* (pp.522-526). Santa Monica, CA: Human Factors and Ergonomics Society.

Schick, F.V. and Hann, R.L., (1987): The use of Subjective Workload Assessment Technique in a complex flight task. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload, AGARDograph No. 282* (pp.37-41*).* Neuilly sur Seine, France: AGARD.

Schick, F.V., Teegen, U., Uckermann, R., and Hann, R.L. (1989). *Validation of the Subjective Workload Assessment Technique in a simulated flight task.* DFVLR-Forschungsbericht 89-01.

Sheridan, T.B. and Simpson, R.W. (1979*). Toward the definition and measurement of the mental workload of transport pilots* (FTL Report R79-4). Cambridge, MA: Flight Transportation Laboratory.

Skelly, J.J. and Purvis, B.D. (1985). B-52 wartime mission simulation: Scientific precision in workload assessment. In *Proceedings of the 1985 Air Force Conference on Technology in Training and Education* (*TITE)* (pp. 105-109).

3.1.3 Modified Cooper-Harper (MCH) Scale

(Wierwille and Casali, 1983)

Description of the method

The Cooper-Harper Scale (Cooper and Harper, 1969) has been used successfully for evaluation of aircraft handling tasks and other motor tasks. In more advanced complex automatic systems the human operator´s role has changed and he is less involved in active control of the system and more occupied with activities like perception, monitoring, evaluation, communication, and problem solving (Wierwille and Casali, 1983). As the original Cooper-Harper Scale was not developed for evaluation of such activities, Wierwille and Casali (1983) modified this method, so that it can be utilized with tasks, characteristic for modern systems. The derived Modified Cooper-Harper (MCH) Scale (Figure 1) retains the decision tree of the original scale but has changed the wording. The authors assume that the scale gives a reliable overall assessment of workload at least on a relative basis.

Factors to consider in test and evaluation

(General factors: see introduction)

Wierwille and Casali (1983) make some recommendations in respect to the appropriate use of the MCH Scale:

- It is recommended that the MCH Scale is used in experiments where overall workload is evaluated.
- The MCH Scale has been validated in a simulated flight environment with perceptual, mediational (cognitive) and communication tasks. The authors assume that the scale will also be sensitive in other environments with such tasks, typical for modern operator machine systems.
- Good experimental design is necessary. Order of presentation has to be counterbalanced and possible confounding processes has to be controlled.
- Ratings should be given immediately after the task to be rated.

- Subjects need careful instruction and adequate training.
- The scale has to be presented in an adequate size. The magnified form of the original scale (in Wierwille and Casali, 1983) should have approximately 26.5 cm in the horizontal.
- According to Lysaght et al. (1989) the MCH Scale measures on ordinal scale level, so that nonparametric statistical tests have to be used.

Advantages of the method
(General advantages of workload rating scales: see introduction)

- The decision tree makes the rating easier by allowing sequential decisions.
- The rating is very time economic and can also be used while performing a task.

Disadvantages of the method
(General disadvantages of workload rating scales: see introduction)

- The scale is based on the assumption that the two dimensions performance and effort are associated. But the relationship between both dimensions is much more complex than assumed.
- The scale is based on the assumption that low workload is highly desirable. But in low workload situations very often problems of vigilance and monotony arise.
- As the scale is presented in the form of a category rating, the values may result in ordinal scale data (Pitrella and Käppler, 1988).
- The scale has a low diagnosticity.
- Measurement is on ordinal scale level.

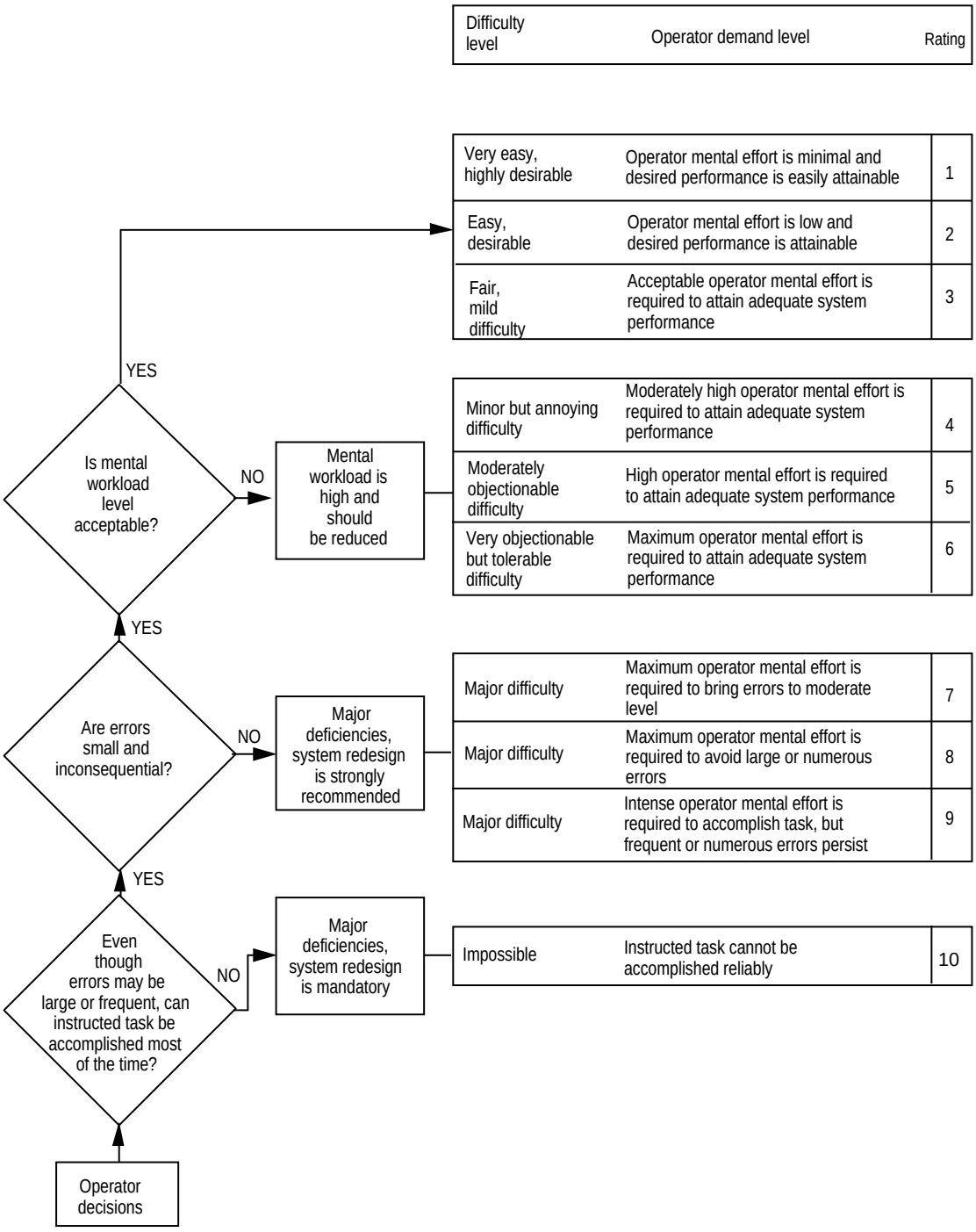


Figure 1: Modified Cooper-Harper Scale

Independence of the method from tester influence

As with all subjective scales, there may be an influence from the tester, as the instructions and the subjects´ training with the scale are not specified. Furthermore, no norms or other aids for data interpretation exist.

Validity of the method

Wierwille and Casali (1983) conducted three experiments varying perceptual, mediational, and communicational load to test validity of the MCH Scale:

- Perceptual experiment: Load in a flight simulator was manipulated in a cross country flight in three levels by varying the rate and number of danger conditions on the engine status displays (e.g. oil temperature). The pilots had to detect and identify the danger conditions. A compensation had not to be made. MCH Scale showed significant differences between all workload levels.
- Mediational (cognitive) experiment: While flying in a simulator pilots had to solve navigational tasks with three levels of difficulty, based on the number and complexity of the problems. The subjects verbalized the results of the problems without implementing them into the flight. The MCH Scale showed two significant differences between workload levels.
- Communications experiment: The task consisted of aircraft control and communications in a flight simulator. For aircraft control pilots had to carry out commands in respect to, e.g., heading, altitude, airspeed given from the tower. The aircraft control task was not changed in difficulty. In addition pilots had to perform a call sign detection task, which was varied in three difficulty levels. The task was assumed to include aspects of communications detection, comprehension, and response execution. MCH Scale showed two significant differences between workload levels.
- Operator Workload was evaluated with four subjective workload rating scales (SWAT, NASA-TLX, OW, MCH) in three weapon systems (remotely piloted vehicle, helicopter, air defence system). A factor analysis supplied one workload dimension with MCH showing the lowest factor loadings (r = 0.799-0.904) (Hill et al., 1992).
- Students performed an aviation like psychomotor dual-task (Kilmer et al., 1988). SWAT was more sensitive in respect to task difficulty (three levels of wind gusts) than the MCH Scale.

Reliability of the method

Skipper et al. (1986) conducted a communication experiment using the MCH scale, like in the study from Casali and Wierwille (1983). The MCH Scale results in both experiments were nearly identical. The conclusion can be drawn that the MCH Scale has a good reliability. Further reliability estimates of the scale are not known.

Feasibility of the method

The MCH Scale can be implemented very easily without any instrumentation. Training with the scale is necessary.

Economy of the method

High economy seen away from training with the scale.

Face validity of the method

In the study from Hill et al. (1992) NASA-TLX and a scale measuring overall workload (OW) were rated better in respect to face validity than the MCH Scale. MCH Scale was rated approximately as good as SWAT.

Interference of the method

Intrusion is only expected when the rating is done during task completion. In this case intrusion depends on familiarity with the task, task difficulty, and familiarity with the MCH Scale.

Diagnosticity of the method

MCH is unidimensional and provides only a measure of overall workload. The method can identify problems but does not give any diagnostic information. For a careful analysis of the problem other methods have to be used.

Generality of the method

According to Hart and Wickens (1990) MCH Scale has demonstrated sensitivity in simulated flight but was less useful in other environments. Main emphasis in the simulated flight experiments was on perceptual, cognitive, and communication load and not on motor load.

References

Casali, J.G. and Wierwille, W.W. (1983). A comparison of rating scale, secondary task, physiological and primary task workload estimation techniques in a simulated flight task emphasising communications load. *Human Factors, 25,* 623-641.

Cooper, G.E. and Harper, R.P. (1969*). The use of pilot rating in the evaluation of aircraft handling qualities*. NASA TN-D-5153. Moffet Field, CA: NASA Ames Research Center.

Hart, S.G. and Wickens, C.D. (1990). Workload assessment and prediction. In H.R. Booher (Ed.): *MANPRINT. An approach to systems integration* (pp. 257-296). New York: Van Nostrand Reinhold.

Hill, S., Iavecchia, H., Byers, J., Bittner, A.C., Zaklad, A.L., and Christ, R.E., (1992). Comparison of four subjective workload rating scales. *Human Factors, 34 (4),* 429-439.

Kilmer, K.J., Knapp, R., Burdsal, C. jr., Borresen, R., Bateman, R., and Malzahn, D. (1988). Techniques of subjective assessment: A comparison of the SWAT and Modified Cooper-Harper Scales. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 155-159). Santa Monica, CA: Human Factors and Ergonomics Society.

Lysaght, R.J., Hill, S.G., Dick, A.O., Plamondon, B.D., Linton, P.M., Wierwille, W.W., Zaklad, A.L., Bittner, A.C., and Wherry, R.J. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies* (Tech. Report 851). Fort Bliss, TX: U.S. Army Research Institute, Field Unit.

Pitrella, F.D. and Käppler, W.-D. (1988). *Identification and evaluation of scale design principles in the development of the Sequential Judgement, extended range Scale*. Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 80.

Skipper, J.H., Rieger, Ch.A., and Wierwille, W.W. (1986). Evaluation of decision-tree rating scales for mental workload estimation. *Ergonomics,* 29, 585-599.

Wierwille, W.W. and Casali, J.G. (1983). A validated rating scale for global mental workload measurement applications. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 129-133). Santa Monica, CA: Human Factors and Ergonomics Society.

3.1.4 Sequential Judgement Scale (Zwei-Ebenen Intensitäts-Skala, ZEIS) (Pitrella and Käppler, 1988)

Description of the method

Corresponding to 14 scale design guidelines compiled from the literature, from ergonomic design principles, and from experiments that indicate a positive influence for a principle on reliability and validity of rating scales, Pitrella and Käppler (1988) constructed the Sequential Judgement Scale in order to rate the difficulty of vehicle handling tasks experienced by drivers. The design of the scale was based on the following principles:

Use continuous instead of category scale formats.
Use both verbal descriptors and numbers at scale points.
Use descriptors at all major scale markings.
Use horizontal rather than vertical scale formats.
Either use extreme or no descriptors at end points.
Use short, precise, and value-unloaded descriptors.
Use empirically determined rank-ordered descriptors.
Select and use equidistant descriptors.
Use psychologically-scaled descriptors.
Use positive numbers only.
Have desirable qualities increase to the right.
Use descriptors free of evaluation demands and biases.
Use 11 or more scale points as available descriptors permit.
Minimise rater workload with suitable aids.

Although designed in response to the need for a better scale to evaluate vehicle handling, the rating scale measures task difficulty and can be applied to a much wider variety of tasks. As subjective experienced task difficulty can be seen as one of the most relevant dimensions of workload, important aspects of workload are measured. The authors found the Sequential Judgement Scale to have interval scale properties as tested by a method from Torgerson (1958) (Käppler and Pitrella, 1988). This permits the use of parametric statistics on rating data. An eleven and a fifteen point version of the scale in German, Dutch and English languages are available.

The English version of the fifteen point scale, that was used in a complex telemanipulation experiment (Spain and Holzhausen, 1991), is shown below (Figure 1). The instructions in the graphics preceeding the scale itself has to be adapted to each task. According to Käppler and Pitrella (1988) the rating is facilitated for the subjects by the two level scale design requiring subjects to make two judgements in sequence, first a coarse judgement, and then a second finer one. The first judgement is made according to three basic categories "difficult", "medium" or "easy". This first level choice graphically switches subjects to specific instructions, and then to a smaller appropriate section of the full continuous scale in order to make a finer rating by marking the line at the appropriate point. Subjects are permitted to cross over to the adjacent section. These three scale sections consist of fewer points so that the workload of subjects is reduced. Consequently the rating can be made with the workload of two short scales but with the precision of a longer scale. Questions and user guidance instructions were integrated into the scale as far as possible to assure that all subjects receive identical instructions and that these are available at the right time. In this way objectivity of the scale is increased. A cognitive model of the internal rating process has been developed by Pitrella (1989a).

Scoring of the scale is done by measuring subjects mark on the scale in millimetres. In order to get a score which is comparable to other scales which increase with subjective task difficulty it is recommended to measure from the scale´s right end point and to transfer the measure into a percentage score of the whole scale. In this way scales of different sizes can also be compared.

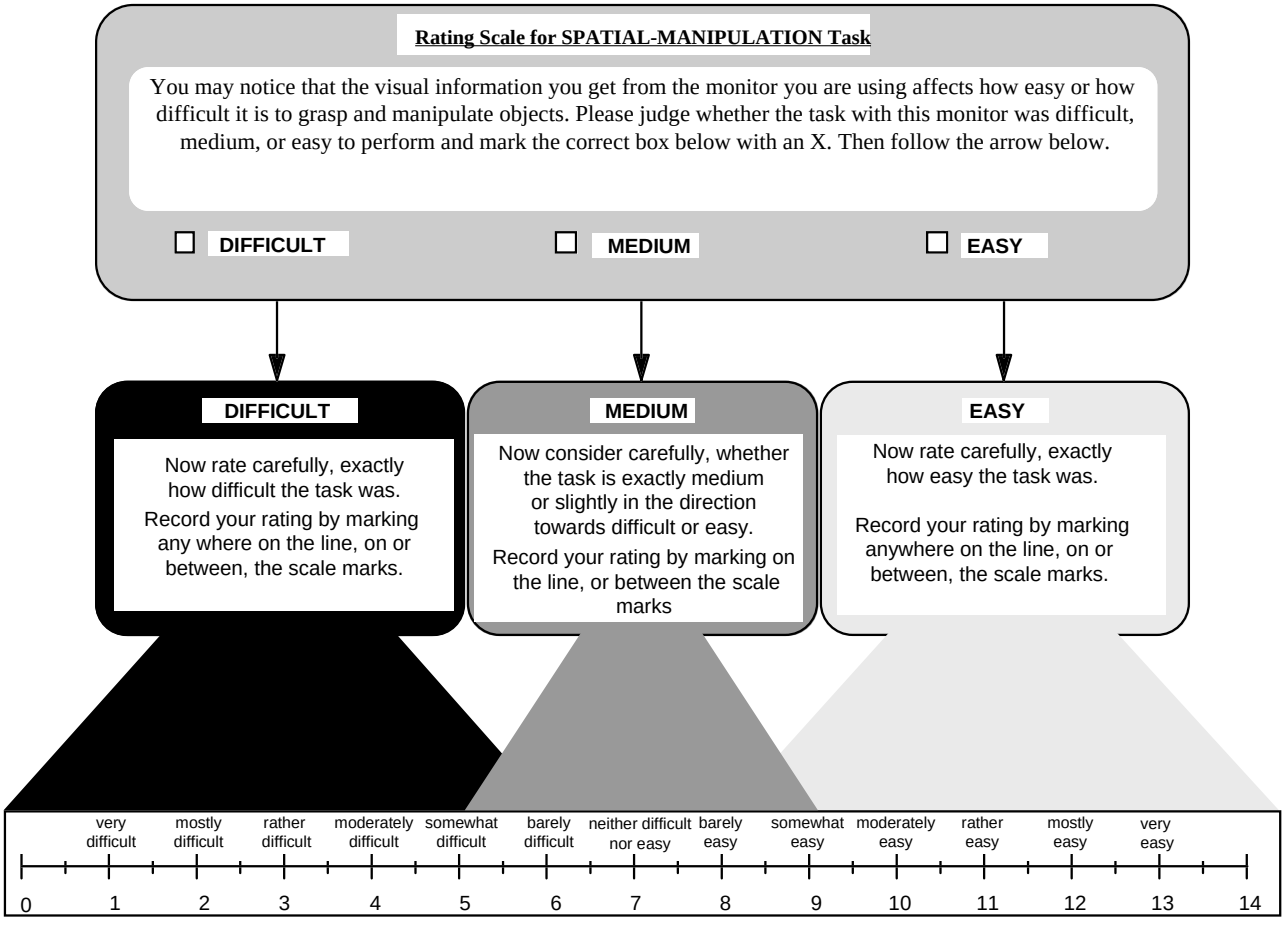


Figure 1: Fifteen point form of the Sequential Judgement Scale

Factors to consider in the phases of test and evaluation
(General factors: see introduction)

- The instructions in the graphics preceeding the scale has to be adapted to each task.
- In the statistical analysis, parametric tests can be used, as the Sequential Judgement Scale was tested to have interval scale properties.
- In contrast to most other workload rating scale methods high rating scores with the Sequential Judgement Scale mean low difficulty level, when not transformed.

Advantages of the method
(General advantages of workload rating scales : see introduction)

- Interval scale level.
- Can be used for measuring workload during task completion as only a little time is needed.
- Dutch, English and German versions of the scale are available.
- Can also be used with different workload dimensions (e.g. the basic scale design could be used for each of the scale dimensions of NASA-TLX).
- Less open to the central tendency bias than other scales.
- Rating scale is based on design principles which have a positive influence on scale performance.
- Good rater acceptance.

Disadvantages of the method
(General disadvantages of workload rating scales: see introduction)

- If only overall workload is measured, rating results will have low diagnosticity.
- Limited information on validity of the scale.

Independence of the method from tester influence

All rating scales need a certain amount of training. The Sequential Judgement Scale has been shown to require only orientation rather than training to achieve reasonable results although rating performance may be improved by training. This reduces tester influence. Also, the incorporation of written instructions as an integral part of the scale eliminates tester influence possible when the tester personally reads or instructs subjects in the use of the scale. Like with most other scales no norms or other interpretation aids exist.

Validity of the method

Some experiments have been made to validate the scale, but the data is as yet limited:

- Truck handling (Käppler et al., 1988): Closed-loop double lane change manoeuvres were made to compare handling qualities of 12 light weight military trucks. The comparison was based on rating scale results. Independent variables: 4 different payload categories, speed of 60 km/h and highest safe speed, loaded and unloaded trucks, 2- and 4-wheel drive. Rating results with the Dutch language Sequential Judgement Scale: 27 different configurations were compared, which resulted in 351 possible pair combinations. From these comparisons 90% were significantly different with respect to subjectively experienced task difficulty. Significant differences in ratings were found between loaded and unloaded trucks and between different models of trucks. No significant differences were found between payload categories nor between 2- and 4-wheel-drive.
- Experimental car handling (Käppler and Godthelp, 1989): The experiment involved closed loop straight lane driving at 100 km/h. Independent variables: Tire pressure variations (3), lane width variations (4). Data from this experiment were used to calculate the first validation correlation of the Dutch language Sequential Judgement Scale. Validity correlation between vehicle handling ratings and tire pressures and lane widths

varied between 0.96 and 1.0. All variations were significant and ratings indicated increasing difficulty with decreasing tire pressure and lane width.

- Manual tracking (Pitrella, 1989b): In a laboratory experiment to validate the 15 point German language Sequential Judgement Scale ten different difficulty levels of a tracking task were presented to the subjects. A validity coefficient was calculated between the ratings of task difficulty and a combination of amplitude and frequency of the tracking forcing function. The correlation varied from 0.88 (the lowest correlation for a single subject) to 0.99 (the highest correlation for a single subject) and reached 0.99 for the group. Sensitivity of the scale permitted a significant discrimination between all 45 combinations of the ten difficulty levels. A comparison of rating results with the objective measure (RMS error) indicated that the rating results equalled the validity and sensitivity of the objective measure.

- Learning (Pfendler, 1993): Validity coefficients (estimates of omega square = o^2) taken when learning a colour pattern detection task give rise to the assumption of the Sequential Judgement Scale ($o^2 = 0.762$) being superior to the German version of NASA-TLX ($o^2 = 0.708$). From the 12 learning blocks, from which 66 multiple comparisons can be derived, 17 comparisons proved to be significantly different with ZEIS and 16 with NASA-TLX at $p<0.05$. Both workload measurement methods were consistent in 15 comparisons.

Reliability of the method

- Truck handling (Käppler et al., 1988): When rating vehicle handling with the Dutch eleven point Sequential Judgement Scale reliability coefficients varied between 0.92 and 0.99 with rater group data and 0.64 for single raters.

- Manual Tracking (Pitrella, 1989b): When rating difficulty levels in target tracking tasks with the German fifteen point Sequential Judgement Scale reliability correlation were 0.99 with rater group data while single rater reliability was 0.87.

- Learning (Pfendler, 1993): A reliability coefficient derived from an analysis of variance (intraclass correlation) showed a correlation of 0.87 for single raters in an experiment involving the learning of patterns in a visual recognition task. With the German version of NASA-TLX a correlation coefficient of 0.56 was calculated.

Feasibility of the method

A paper and pencil version and a computerized version of the scale are available which have nearly no implementation restrictions at all.

Economy of the method

Minimal costs.

Face Validity of the method

High face validity.

Interference of the method

When used after task completion no interference exists. There is little interference, if any, when used periodically during a task, as only one rating has to be given. But this effect depends on familiarity of subjects with the scale, the task to be rated, and the difficulty level of the task. With a low task difficulty level no interference is expected.

Diagnosticity

Low diagnosticity as only one dimension is measured in general. But: the scale format could also be applied with different dimensions, e.g., with the six dimensions of NASA-TLX. This would provide higher diagnosticity.

Generality

The Sequential Judgement Scale can be applied to any task that can be measured with a difficulty intensity scale. In addition to the above mentioned studies the scale was also applied, e. g., in a comparison of three different tank gunnery control systems (Krüger and Gärtner, 1992). The scale can also be adapted to other dimensions if a suitable set of descriptors can be found.

References

Käppler, W.-D. and Godthelp, H. (1989). *Design and use of the Two-Level Sequential Judgement Rating Scale in the identification of vehicle handling criteria: I. Instrumented car experiments on straight lane driving*. Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 79.

Käppler, W.-D. and Pitrella, F.D. (1988). Evaluation of vehicle handling: Design and test of the Two-Level Sequential Judgement Rating Scale. In *Proceedings of the 23rd Annual Conference on Manual Control.* Massachusetts Institute of Technology, Cambridge, Mass.

Käppler, W.-D., Pitrella, F.D., and Godthelp, H. (1988). *Psychometric and performance measurement of light weight truck handling qualities*. Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 77.

Krüger, W. und Gärtner, K.-P. (1992). *Untersuchung verschiedener Bediensignalkennlinien für Richtschützen unter Beschleunigungsstörungen.* Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 97.

Pfendler, C. (1993). Vergleich der Zwei-Ebenen Intensitäts-Skala und des NASA Task Load Index bei der Beanspruchungsbewertung während Lernvorgängen. *Z. Arb. wiss. 47 (19 NF) 1993/1*, 26-33.

Pitrella, F.D. and Käppler, W.-D. (1988). *Identification and evaluation of scale design principles in the development of the Sequential Judgement, extended range Scale*. Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 80.

Pitrella, F.D. (1989a). *A cognitive model of the internal rating process*. Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 82.

Pitrella, F.D. (1989b). *Validation of the 15 point German language Sequential Judgement Scale.* (unpublished)

Spain, E.H. and Holzhausen, K.-P. (1991). Stereoscopic versus orthogonal view displays for performance of a remote manipulation task. In J.O. Merrit and S.S. Fisher (Eds.), *Stereoscopic displays and applications, Proceedings of the Society of Photo-Optical Instrumentation Engineers* (pp.103-110). Bellingham: SPIE.

Torgerson, W.S. (1958). *Theory and methods of scaling*. New York: John Wiley.

3.1.5 Subjective Workload Dominance (SWORD) Technique

(Vidulich, 1989)

Description of the method

The Subjective Workload Dominance (SWORD) technique (Vidulich, 1989) was derived from a technique called the Analytic Hierarchy Process (AHP), developed by Saaty (1980), to evaluate any dimension. Lidderdale (1987) was the first who applied the method to subjective workload assessment. In contrast to the original AHP, the method was modified. Whereas AHP uses the eigenvector matrix calculation algorithm to transform the raw data into the rating scale values, Lidderdale applied the geometric mean method, which is easier to understand and to calculate and comes to comparable results. Lidderdale also applied a different rating scale with less scale points than AHP. SWORD uses the rating scale recommended by Saaty (1980) combined with the geometric mean transformation. Vidulich (1989) points out that the results with this transformation are essentially identical with the results, when using the eigenvector calculation algorithm. There are also some important differences between SWORD and most other workload rating scale techniques (Vidulich, 1989). In contrast to most other techniques, which are based on absolute comparisons (without referring to the other task conditions), SWORD uses relative comparisons, i.e., each task is individually compared to all other tasks. Furthermore, when using SWORD tasks are evaluated retrospectively after having completed all tasks. With other methods (e.g., SWAT, NASA-TLX) tasks are evaluated immediately, one after the other, to avoid negative memory effects. Finally SWORD ratings

have ratio scale level (Vidulich et al., 1991), whereas most other approaches have only interval or ordinal scale level. Consequently there are no restrictions in statistical data analysis when using SWORD.

The SWORD method is based on a paired comparison and uses three main steps: (1) Collecting the raw judgement data. (2) Constructing the judgement matrices, and (3) calculating the SWORD ratings. The following summarizes the description of Vidulich et al. (1991).

(1) Collecting the raw judgement data: After performing all tasks the rater has to make a paired comparison between all possible task combinations in respect to workload. For each comparison a rating scale with 17 discrete steps is presented. One task appears on the left endpoint of the scale, the other one on the right hand side. When the rater experiences the two tasks as equal in workload, he makes a mark on the "Equal" in the centre of the rating scale. The more dominant one task is in respect to workload, the closer to the endpoint of the dominant task the mark will be placed by the rater in the corresponding category of the scale. To characterize the relative workload dominance of one of both tasks, nine verbal descriptors with "Equal" in the centre and "Weak", "Strong", "Very Strong" and "Absolute" at both sides of the centre are added to the scale (Figure 1).

(2) Constructing the judgement matrix: In the judgement matrix the rows and columns correspond to the tasks which are compared two by two. Each cell in the matrix shows the result of the paired comparison of the corresponding task in the row and the task in the column. The diagonal from the upper left to the lower right shows the comparison of each task with itself and contains the values one. The upper right part of the matrix contains a value of one, when the compared tasks have approximately the same workload level, a value from 2, 3, 4, 5, 6, 7, 8 or 9, when the task on the left hand side of the rating scale is dominant (2, meaning the lowest dominance level and 9 the highest), or the reciprocal of these numbers (1/2, 1/3.....1/9, where 1/9 means the highest dominance level), when the task on the right hand side of the rating scale is dominant in respect to workload. The lower left side of the matrix is filled with the reciprocals of the corresponding upper right cells (Figure 2).

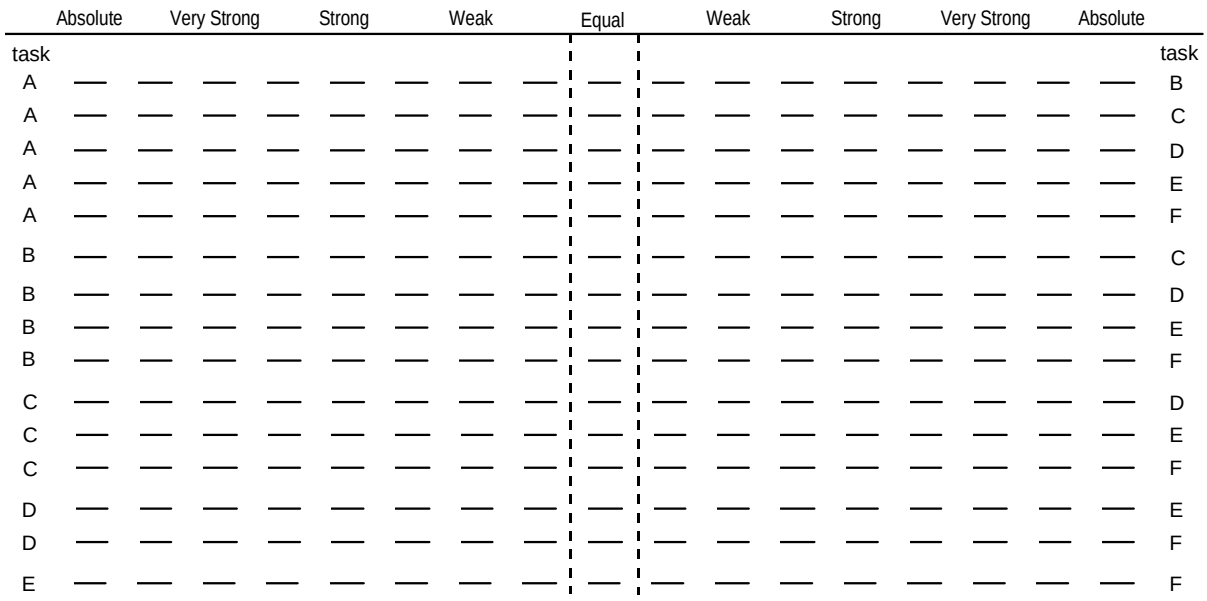


Figure 1: SWORD evaluation form

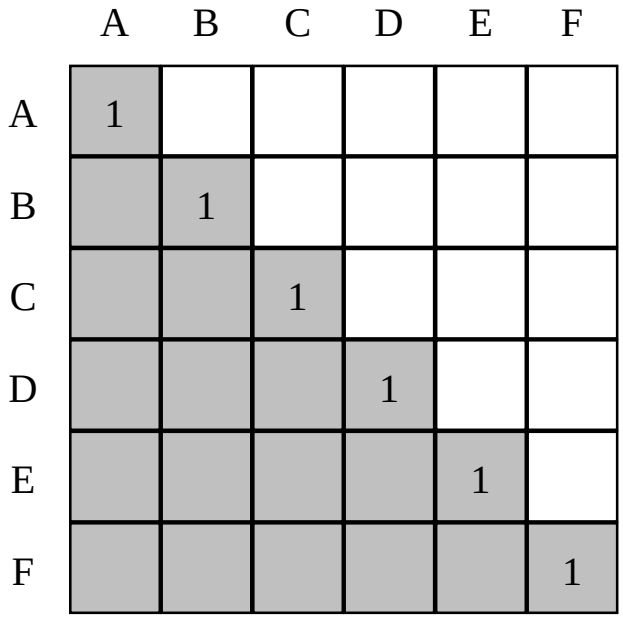


Figure 2: SWORD judgement matrix

(3) Calculating the ratings: For each row of the matrix the geometric mean is calculated. These values are normalized. The resulting values are the workload ratings of each task. Problems can arise, when ratings are not consistent. "If task A is rated twice as hard as task B, and task B is rated three times as hard as task C, than task A should be rated six times as hard as task C" (Vidulich et al., 1991). The authors also describe a measure to evaluate consistency (Vidulich et al., 1991).

Factors to consider in test and evaluation
(General factors: see introduction)

- Evaluation is done after performing all tasks.
- The number of tasks which can be compared is limited.
- Retrospection is necessary.
- The tasks has to be identifiable in the evaluation form for the paired comparison.
- SWORD has ratio scale properties (Vidulich et al., 1991) and therefore no limitations for data analysis and interpretation.

Advantages of the method
(General advantages of workload rating scales: see introduction)

- SWORD has ratio scale level.
- Relative comparisons are easier to perform than absolute comparisons (e.g. SWAT), as the subject can focus on the relationship between only two conditions (Vidulich and Tsang, 1987).
- Pro-SWORD can be used for projective purposes (Vidulich et al., 1991).
- As SWORD is based on retrospective ratings it has better retest reliability than immediate rating techniques like, e.g., NASA-TLX (Vidulich, 1989).
- Demonstrated higher sensitivity than NASA-TLX and OW (a unidimensional workload scale; Hill et al., 1992) (Vidulich, 1989).
- Consistency of rater judgements can be measured.

Disadvantages of the method

(General disadvantages of workload rating scales: see introduction)

- As SWORD is based on paired comparison the number of tasks which can be compared is limited. The number of comparisons is 1/2n(n-1). When too many tasks have to be compared (e.g., 45 comparisons with 10 tasks) memory is overloaded.
- SWORD cannot be done during task completion.
- As SWORD is based on retrospection, memory effects can influence the results (e.g., position effects).
- The tasks have to be identifiable in the evaluation form for the paired comparison.

Independence of the method from tester influence

The use of SWORD requires a training of the subjects with the method that is not specified more precisely. Therefore, certain limitations concerning objectivity during data collection phase exist, while the objectivity of scoring is given. The objectivity of data interpretation is also limited as there are no norms and tolerance levels available (like with most other workload measurement methods).

Validity of the method

- Subjective workload was measured with an auditory and visual single axis tracking task, performed under single and dual task conditions (Vidulich and Tsang, 1987). AHP (which differs from SWORD only in transformation of raw data, but comes to essentially the same results) was compared to NASA-TLX and the Overall Workload (OW) Scale and demonstrated the highest validity, measured by the variance accounted for.
- Projective validity of Pro-SWORD was tested in a comparison of six head-up displays, while measuring workload with SWORD in a projective and a retrospective way (Vidulich et al., 1991). The projective group consisted of F-16 pilots, who were familiar with only one format of these displays, whereas the retrospective group consisted of F-16 pilots, who had used all display formats in a simulator study. The correlation between the ratings of both groups was highly positive (r = 0.94).

Reliability of the method

In the study of Vidulich and Tsang (1987) test-retest reliability was measured and AHP was observed to have the highest mean correlation coefficients (r=0.896) in comparison to NASA-TLX (r = 0.421) and OW (r = 0.274).

Feasibility of the method

SWORD can be implemented very easily and needs only few instrumentation.

Economy of the method

SWORD is not as economic as other unidimensional workload measurement methods, especially when there are many experimental conditions. Data collection with SWORD takes more time than with other unidimensional workload measurement methods, as all tasks have to be compared with each other. Furthermore, raw data transformation can be cumbersome when no software is available.

Face validity of the method

Although there are no empirical data about face validity available, SWORD is assumed to have good face validity.

Interference of the method

SWORD does not interfere with the main task as the rating cannot be done during task completion.

Diagnosticity of the method

As SWORD is unidimensional, the method cannot differentiate between different sources of workload.

Generality of the method

SWORD is applicable when not too many tasks, experimental conditions etc. have to be compared in respect to workload, as all tasks have to be compared with each other. E.g., with ten tasks 45 comparisons have to be made. Then, also the problems of memory overload and of position effects arise. Furthermore, the tasks to be compared must have such characteristics, that they can be easily verbalized by the experimenter and identified easily by the subject so that the paired comparison can be done. Consequently experimental conditions differing in quantity (e.g., tracking tasks differing in difficulty level) are difficult to compare, as it is difficult to verbalize the experimental conditions. Conditions differing in quality (e.g., different display formats) can be compared much more easier. Also mission segments can only be compared when they can be identified easily by the subjects.

References

Hill, S., Iavecchia, H., Byers, J., Bittner, A.C., Zaklad, A.L., and Christ, R.E., (1992). Comparison of four subjective workload rating scales. *Human Factors, 34 (4),* 429-439.

Lidderdale, I.G. (1987). Measurement of aircrew workload during low-level flight. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload, AGARDograph No. 282* (pp. 78-82). Neuilly sur Seine, France: AGARD.

Saaty, T.L. (1980). *The analytic hierarchy process*. New York: McGraw-Hill.

Vidulich, M.A., (1989). The use of judgement matrices in subjective workload assessment: The Subjective Workload Dominance (SWORD) Technique. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1406-1410). Santa Monica, CA: Human Factors and Ergonomics Society.

Vidulich, M.A. and Tsang, P.A. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. In *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 1057-1061). Santa Monica, CA: Human Factors and Ergonomics Society.

Vidulich, M.A., Ward, G.F., and Shueren, J. (1991). Using the Subjective Workload Dominance (SWORD) Technique for projective workload assessment. *Human Factors, 33(6),* 677-691.

3.1.6 Bedford Scale

(Ellis and Roscoe, 1982)

Description of the method

The Bedford Scale was derived from the Cooper-Harper (CH) Scale (Cooper and Harper, 1969) by Ellis and Roscoe(1982) and uses the same decision tree method. Whereas the CH Scale is used to rate experienced handling qualities of aircraft, the Bedford scale is centred on workload measurement by assessing experienced spare capacity of pilots (Figure 1). In contrast to the other hierarchical scales half ratings are allowed (e.g., 5.5). According to Roscoe (1987) "Pilot workload is the integrated mental and physical effort required to satisfy the perceived demands of a specified flight task". The scale has been generally welcomed by military and civil airline pilots as this interpretation of workload can be accepted by most pilots (Roscoe, 1987).

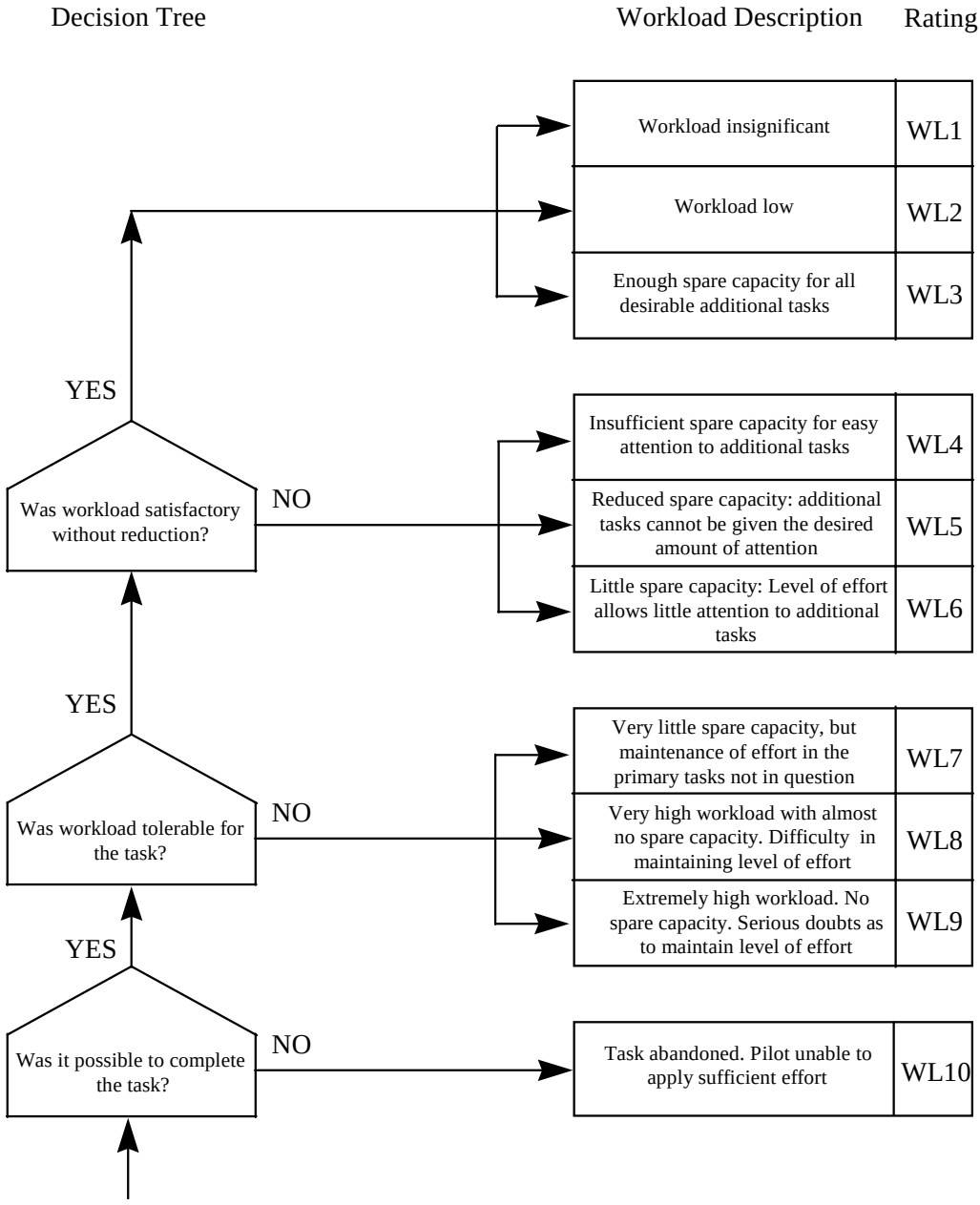


Figure 1: The Bedford Scale (Roscoe, 1987)

Factors to consider in test and evaluation
(General factors: see introduction)

- In-flight ratings with single-seat aircraft are critical, as the rating can interfere with the main task (Roscoe, 1987, Lidderdale, 1987)).
- The Bedford Scale is most valuable when the pilot is flying manually and less valuable when he is only monitoring (Roscoe, 1987).
- The scale is inappropriate for postflight debriefings (Lidderdale, 1987).
- Knee pads with ten push buttons can be used to make in-flight ratings. The scale contents can be remembered easily (Lidderdale, 1987).
- Scale training needs less than half an hour.

- Ratings are given within 15 minutes of each task completion.
- Method requires data gathering sheets designed for the mission that are used with the Bedford Scale. Subject comments are used to fine tune the questionnaires.
- Extensive notes and comments are needed (either recorded by the evaluator, an audio recorder, or on the audio track of a video recorder).
- As the Bedford Scale is not linear (Roscoe, 1987), nonparametric statistics have to be used.
- Different experiences of pilots influence rating results which are therefore not absolute workload ratings. Consequently results from different pilots cannot be compared (Roscoe, 1987).
- Rating values between about 1-3 are considered as tolerable (Wainwright, 1987)

Advantages of the method
(General advantages of workload rating scales: see introduction)

- The Bedford Scale works for global workload assessment.
- The scale is simple. It also works well for explaining test results to the uninitiated.
- The Bedford Scale is well accepted by most pilots.
- The decision tree makes the rating easier by making sequential decisions.
- The rating is very time economic and can also be used while performing a task.

Disadvantages of the method
(General disadvantages of workload rating scales: see introduction)

- Database for validity and reliability of the method is insufficient.
- The scale is simple. Each respondent attaches different meanings to the ratings and their descriptions.
- Pilots tend to provide low ratings on any flying task they perform.
- Poor sensitivity in the lower range of the scale. It does not distinguish workload well in the 1 to 3 rating range.
- Diagnosticity is a problem. Once a problem is identified, it is difficult to pinpoint what the exact problem may be. Additional techniques are required.
- Measurement is on ordinal scale level.

Independence of the method from tester influence

As with all subjective scales, there may be an influence from the tester, as the instructions and the subjects´ training with the scale are not exactly specified. As well, no norms for data interpretation exist. But Wainwright (1987) gives some aids for interpretation of results. According to him "a satisfactory workload is not only demonstrated by all ratings falling in the range 1 to 3, but also when mean workload is in that range, with some deviations into the acceptable bracket".

Validity of the method

The Bedford Scale was mostly utilised in applied settings (Lysaght et al., 1989). Lidderdale (1987), e.g., measured workload of aircrew during a low level flight in an advanced combat aircraft at night or in poor weather. The mission was subdivided into ten sections. In-flight measurement of workload was done with the Bedford Scale, whereas the Analytic Hierarchy Process method (Saaty, 1980) was used for post-flight measurement. There was a high correlation between the results from both methods. Seen away from one experiment (Tsang and Johnson, 1987) having a too small data base, Lysaght et al. (1989) could not find other studies using the Bedford Scale in controlled settings with defined levels of task difficulty. This means that there are not enough data on validity of the Bedford Scale available.

Reliability of the method

No data on reliability of the Bedford Scale could be found.

Feasibility of the method

According to Lidderdale (1987) the Bedford Scale "has proved to be a practical solution to in-flight measurement even during the most demanding tasks", but is not suitable for single-seat operations, as there is no second crew member to ask for an assessment.

Economy of the method

High economy seen away from training with the scale.

Face validity of the method

The Bedford Scale is based on a workload definition which can be accepted by most pilots (Roscoe, 1987). In this respect the Bedford Scale has a high face validity.

Interference of the method

Lidderdale (1987) demonstrated that the rating did not interfere significantly with the primary tasks of the pilot and the navigator even under the most demanding conditions. The method is not assumed to be suitable with single seat operations as task interference is expected.

Diagnosticity of the method

The Bedford Scale is unidimensional and provides only a measure of overall workload. The method can identify problems but does not give any diagnostic information. For a careful analysis of the problem other methods have to be used.

Generality of the method

According to Hart and Wickens (1990) the Bedford Scale was explicitly developed for in-flight measurement of workload and has been used widely in Europe to evaluate pilot workload in military and civilian aircraft and in simulation and flight research in the United States, but has received limited use in non-aviation environments. But when the word "pilot" is replaced by "operator", the Bedford Scale can also be applied to other operational environments (Lysaght et al., 1989).

References

Cooper, G.E. and Harper, R.P. (1969). *The use of pilot rating in the evaluation of aircraft handling qualities*. NASA TN-D-5153. Moffet Field, CA: NASA Ames Research Center.

Ellis, G.A. and Roscoe, A.H. (1982). *The airline pilot´s view of flight deck workload: A preliminary study using a questionnaire*. Royal Aircraft Establishment Technical Memorandum No FS (b) 465.

Hart, S.G. and Wickens, C.D. (1990). Workload assessment and prediction. In H.R. Booher (Ed.), *MANPRINT. An approach to systems integration* (pp. 257-296). New York: Van Nostrand Reinhold.

Lidderdale, I.G. (1987). Measurement of aircrew workload during low-level flight. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload, AGARDograph No. 282* (pp. 69-77). Neuilly sur Seine, France: AGARD.

Lysaght, R.J., Hill, S.G., Dick, A.O., Plamondon, B.D., Linton, P.M., Wierwille, W.W., Zaklad, A.L., Bittner, A.C., and Wherry, R.J. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies* (Tech. Report 851). Fort Bliss, TX: U.S. Army Research Institute, Field Unit.

Roscoe, A.H. (1987). In-flight assessment of workload using pilot ratings and heart rate. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload, AGARDograph No. 282* (pp.78-82). Neuilly sur Seine, France: AGARD.

Saaty, T.L. (1980). *The analytic hierarchy process*. New York: McGraw-Hill.

Tsang, P.S. and Johnson, W. (1987). Automation: Changes in cognitive demands and mental workload. In *Proceedings of the Fourth Symposium on Aviation Psychology*. Columbus, OH: Ohio State University.

Wainwright, W.A. (1987). Flight test evaluation of crew workload. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload, AGARDograph No. 282* (pp.60-68). Neuilly sur Seine, France: AGARD.

Aside from subjective methods there are also analytical techniques, secondary tasks, physiological techniques and measures of primary task performance which can be used for workload evaluation. Analytical techniques are not discussed here, as the goal of the research study group was to mention only methods which can be applied with existing systems. Measures of primary tasks will be discussed in the chapter of performance measurement. Physiological methods and secondary tasks are not as important as rating scales for test and evaluation and that is also why they are described only in short overviews. The reasons will be explained more in detail in the corresponding chapters (e.g., controversial results, limited applicability, high implementation demands, etc.). In most cases these methods should be used only as additional tools, when there are special reasons to apply them. The overviews on these methods will follow the same outline as the descriptions of the rating scale methods, as the same criteria are valid also for the evaluation of these approaches to workload measurement.

3.2 Secondary Task Method

Description of the method

Secondary tasks as workload measurement methods were developed as tools to assess the work capacity and limitations of the human operator with respect to primary task performance (Bornemann, 1942; Lysaght et al., 1989). In this paradigm the operator is required to perform a secondary task while performing the primary task at the same time. The method is based on the assumption that the human operator as an information processing system has only limited channel capacity and the spare capacity left by the primary task is filled up by the secondary task. The higher the performance decrement in the secondary task the lower is the remaining spare capacity of the operator in the particular primary task and the higher is operator workload (Brown, 1962; O'Donnel and Eggemeier, 1986). Lysaght et al. (1989) distinguish about 26 different classes of secondary tasks (e.g., choice reaction time tasks, tracking tasks, monitoring tasks etc.). In contrast to the external secondary tasks the embedded secondary tasks (e.g., radio communication when flying an aircraft) are part of normal system functions (Wierwille and Eggemeier, 1993). They appear as a natural and integral part of the task of interest (Hart and Wickens, 1990) and reduce the main problems of conventional external secondary tasks: task intrusion, poor operator acceptance, and high implementation demands (Lysaght et al., 1989).

Factors to consider in the phases of test and evaluation

- Workload measurement is secondary task specific: which aspect of workload is measured depends on the selected secondary task. When using different kinds of secondary tasks with the same primary task, the results can dissociate.
- Selection of secondary tasks: Such secondary tasks should be selected, which require the same resources necessary for performance of the primary tasks (Hart and Wickens, 1990).
- The use of embedded secondary tasks is recommended as they avoid the general problems encountered with conventional secondary tasks.
- Some tasks produce such a high workload (e.g., mental mathematics) that a secondary task cannot be added (Lysaght et al., 1989).
- Secondary tasks require a sufficient amount of training without and with the simultaneously performed primary task.
- The possibility of secondary task intrusion on primary task should always be considered carefully. Safety aspects should not be neglected in operational systems.
- Secondary tasks measure on interval scale level. Use of parametric statistics is possible.

Advantages of the method

- Objective workload assessment with minimal influence of subjective factors.
- Good diagnosticity when using different secondary tasks.
- Measurement of continuous workload is possible with some secondary tasks, e.g., with tracking tasks (Hart and Wickens, 1990).
- Measurement on interval scale level allows better use of information. E.g., calculation of means and standard deviations is possible. This allows to derive several workload scores from one secondary task.

Disadvantages of the method

- Studies with secondary tasks show contradicting results. Ogden et al. (1979) compared 144 secondary tasks. For most tasks there is one study showing an improvement, another degradation, and a third no change of the secondary task performance with increasing workload (Lysaght et al., 1989).
- The secondary task can interfere with the primary task and can cause a performance decrement in it. This can threaten safety in an operational system.
- Legal aspects have to be considered when using secondary tasks in field experiments.
- Secondary tasks require a sufficient amount of training.
- Operator acceptance is often a problem with secondary tasks.
- Attention allocation between secondary task and primary task can be unstable and invalidate the results.
- The implementation costs can be high, if not an embedded secondary task is used.
- The recommendations for selection of secondary tasks are contradictory. Some authors insist on interference of secondary tasks with primary tasks, others want to avoid interference.
- The results can depend on the subjects ability to perform two tasks at the same time.
- Secondary tasks measure task specific. Results from different main tasks but with the same secondary task cannot be compared, as interaction between main tasks and secondary tasks is different.

Independence of the method from tester influence

As all secondary tasks need a certain amount of training they are not completely independend from the tester. The tester must decide when performance in the secondary task is asymptotic and when the training can be finished. If the training is finished too early there can be a confounding of learning and of the experimental conditions.

Validity of the method

Many validation studies with secondary tasks have been conducted but results are contradictory and no general conclusions about their validity can be drawn (Ogden et. al., 1979). Only some examples of these studies will be presented here. An excellent overview can be found in Lysaght et al. (1989).

Some experiments have been done with car driving tasks under field conditions and validity of some secondary task approaches could be demonstrated. Brown (1962, 1965) and Brown et al. (1966) used an auditory secondary task to measure workload during car driving and found significant differences between driving conditions.

In the studies of Wiegand (1974, 1989) soldiers with different amounts of driving experience drove military trucks with and without traffic and workload was measured with a secondary task. In the first experiment the subtest "digit span" (reproduction of digits) of the "Wechsler Adult Intelligence Scale" was used as a secondary task. Lists with 20 items of three, four and five digits had to be reproduced backward while driving. Only the lists with three and four digits differentiated significantly between driving with and without traffic ($p<0.05$) with the inexperienced drivers. Due to traffic law demands, a new predictor had to be selected. Therefore, time estimation technique was used, in which Subjects had to produce time intervals of 20s. Measures of central tendency were not sensitive, but all measures of variability (standard deviation, coefficient of variation, sum of differences)

differentiated in accordance to workload between levels of traffic and between levels of driving experience of military truck drivers.

In a tracking task with simplified car dynamics Pfendler (1982) used monitoring as a secondary task. Subjects had to monitor a display where a pointer randomly moved between three sections. The "alarm" sections were on the left and right side and the "normal" section was in the center of the display. Subjects had to press a key, as long as the pointer was moving in one of the alarm sections of the display. Percent of missed signals was calculated and two of three comparisons between the difficulty levels of the tracking task were significant ($p<0.01$).

Michon (1966) used tapping as a secondary task and demonstrated an increase in irregularity with workload using different laboratory tasks. Pfendler and Johannsen (1977) used the same task for workload measurement with simulated STOL landing approaches without finding any significant differences between flights with three difficulty levels (flights with and without gusts, and with gusts and use of the stability augmentation system of the aircraft).

Reliability of the method

There were only a few studies available in which reliability of secondary tasks was measured. Pfendler and Johannsen (1977) measured retest reliability of the tapping task and found coefficients of $r_{tt} = 0.8$. In another study retest reliability of a monitoring secondary task was tested (Pfendler, 1982) and coefficients of $r_{tt} = 0.8$ were reached again.

Feasibility of the method

When not using an embedded secondary task implementation demands can be high, depending on the kind of secondary task selected. There might also be space problems in operational systems. Furthermore a training is necessary.

Economy of the method

Secondary tasks cause considerably higher costs than, e.g., rating scales. But costs can be reduced with embedded secondary tasks.

Face validity of the method

Face validity of conventional secondary tasks is low, as the measurement principle is not always clearly understandable for the subjects. Face validity of embedded secondary tasks is assumed to be much more higher than with conventional secondary tasks and consequently also operator acceptance.

Interference of the method

Some authors assume that there must be interference between the primary and the secondary task so that the measurement principle can work, others refuse this assumption. Empirically, interference with corresponding performance decrements in the primary task has been observed in many experiments and it can be a real problem in operational systems, when safety is threatened. Embedded secondary tasks avoid this problem. Extensive training can also reduce interference.

Diagnosticity

When different secondary tasks are used with the same primary task they do not demonstrate all the same sensitivity. When a secondary task responds very sensitive, this shows, that there is considerable overlap in the processing resources of the primary and the secondary task (Wierwille and Eggemeier, 1993). That is why secondary tasks can be used as diagnostic tools. From such studies conclusions can be drawn about the resources which should be used, e.g., for additional functions which are intended to be implemented into a system or about the special workload drivers of a task.

Generality

The application of external secondary tasks is restricted to tasks, where interference with the primary task is irrelevant and safety aspects of the system are not threatened. But there is no generally applicable secondary task, which can be used with all kinds of primary tasks. The secondary task should always be selected carefully so that it uses the same resources as the primary task, although there is some controversy about this point. For a primary tracking task the best secondary task is also tracking, but tracking cannot be used to measure workload with a monitoring task. With imbedded secondary tasks safety and interference problems are reduced.

Secondary tasks can also be used to measure mental degradation after sustained operations or to measure detrimental effects from environmental factors on operator behaviour.

References

Bornemann, E. (1942). Untersuchungen über den Grad der geistigen Beanspruchung. II. Teil: Praktische Ergebnisse. *Arbeitsphysiologie, 142, 12*, 173-191.

Brown, I.D. (1962). Measuring the spare "mental capacity" of car drivers by a subsidiary auditory task. *Ergonomics, 5,* 247-250.

Brown, I.D. (1965). A comparison of two subsidiary tasks used to measure fatigue in car drivers. *Ergonomics, 8,* 467-473.

Brown, I.D., Tickner, A. H., and Simmonds, D. C. (1966). Effects of prolonged driving upon driving skill and performance of subsidiary tasks. *Industrial Medicine of Surgery, 35,* 760-765.

Hart, S.G. (1975). Time estimation as a secondary task to measure workload. In *Proceedings of the 11th Annual Conference on Manual Control*. NASA, TM X-62.

Hart, S.D. and McPherson, D. (1976). Airline pilot time estimation during concurrent activity including simulated flight. In *Proccedings of the 47th Annual Meeting of the Aerospace Medical Association* (pp. 41-45).

Hart, S.G. and Wickens, C.D. (1990). Workload assessment and prediction. In H.R. Booher (Ed.): *MANPRINT: An approach to systems integration* (pp.257-296). New York: Van Nostrand Reinhold.

Knowles, W.B. (1963). Operator loading tasks. *Human Factors*, 5, 155-166.

Lysaght, R.J., Hill, S.G., Dick, A.O., Plamondon, B.D., Linton, P.M., Wierwille, W.W., Zaklad, A.L., Bittner, A.C., and Wherry, R.J. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies* (Tech. Report 851). Fort Bliss, TX: U.S. Army Research Institute, Field Unit..

Michon, J.A. (1966). Tapping regularity as a measure of perceptual motor load. *Ergonomics, 5,* 401-412.

O´Donnell, R.D. and Eggemeier, F.T. (1986). Workload assessment methodology. In K. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and human performance, Vol. 2, Cognitive processes and performance* (pp. 42/1-42/49). New York: Wiley & Sons.

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

Pfendler, C. (1982). Bewertung der Brauchbarkeit von Methoden zur Messung der mentalen Beanspruchung bei Kfz-Lenkaufgaben. *Zeitschrift für Arbeitswissenschaft, 36 (8 NF) 1982/3,* 170-174.

Pfendler, C. and Johannsen, G. (1977). *Beiträge zur Beanspruchungsmessung und zum Lernverhalten in simulierten STOL-Anflügen*. Wachtberg: Forschungsinstitut für Anthropotechnik, FAT Report No. 30.

Wiegand, D. (1974). Die quantitative Messung der psychischen Beanspruchung während des Fahrens durch eine simultane Nebentätigkeit. *Zeitschrift für experimentelle und angewandte Psychologie, XXI,* 679-690.

Wiegand, D. (1989). Measuring the mental workload during task related activities including driving a vehicle by means of concurring time interval estimates. In M. Lind and E. Hollnagel, (Eds.), *Eighth European Annual Conference on Human Decision Making and Manual Control* (pp. 64-75). Lingby: Technical University of Denmark, Institute of Automatic Control Systems.

Wierwille, W.W. and Eggemeier, F.T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors, 35 (2),* 263-281.

Wilson, G.F. and Eggemeier, F.T. (1991). Psychophysiological assessment of workload in multi-task environments. In D. L. Damos (Ed.), *Multiple task performance* (pp. 329-360). London: Taylor & Francis.

Wilson, G.F. and Fullenkamp, P. (1991). A comparison of pilot and WSO workload during training missions using psychophysiological data. In *Proceedings of the Western European Association for Aviation Psychology, Vol. II: Stress and error in aviation* (pp. 27-34). Western European Association for Aviation Psychology, England.

Wilson, G.F. and O'Donnell, R.D. (1988). Measurement of operator workload with the neurophysiological workload test battery. In P.A. Hancock and N. Meshkati (Eds.), *Human mental workload* (pp. 63-100). Amsterdam: North Holland.

3.3 Physiological Workload Measurement Techniques

Description of the methods

Measurement of operator workload with physiological techniques is based on the assumption that workload is reflected in the physiological responses of the central nervous system, the autonomic, the somatic and the endocrine system. According to Hart and Wickens (1990) there are two classes of physiological measures: Measures of emotional and physical activation like heart rate and pupil size and measures of perceptual and mental processing like event related potentials and direction of gaze. What has been said in respect to secondary tasks is also valid for physiological measures of workload: every technique has been shown to be sensitive to workload and almost every technique has been shown to have failures (Lysaght et al., 1989). From this the conclusion should be drawn that physiological methods should only be used in addition to subjective workload measurement methods. They can be applied when implementation demands under field conditions are low and when unobtrusive and continuous measures of workload are needed (Wierwille and Eggemeier, 1993).

Factors to consider in the phases of test and evaluation

- Different physiological methods (and also measures) measure different aspects of workload.
- When using independent groups relative scores have to be used with some methods (e.g., with ECG measures) which relate baseline measures and loaded measures. Consequently baseline measurements have to be made.
- Physiological measures might have artefacts (e.g., the R-peaks are not registered correctly in the ECG).
- Physical activity can influence physiological measures (e.g., heart rate).
- Physiological measurement can restrict mobility of the operator (e.g. through wires, headrests etc.).
- The operator should have the opportunity to adapt to the research environment before the experiment begins (otherwise, e.g., heart rate etc. can be increased).

Advantages of the methods

Partly depending on the selected method

- With many methods continuous workload measurement is possible.
- Objective assessment of workload without involving subjective aspects is possible.
- Anticipatory workload reactions can be assessed as well as workload consequences.
- Assessment of long-term effects of workload after sustained operations is possible.
- With some methods (e.g., ECG) different measures can be derived simultaneously, which assess different aspects of workload (heart rate and heart rate irregularity).
- Many different methods can be applied simultaneously (Hart and Wickens, 1990).

- The subject cannot voluntarily influence the result (like with rating scales).
- The methods do not interfere with task performance (Hart and Wickens, 1990).

Disadvantages of the methods

Partly depending on the selected method

- Possibility of artefacts in resulting data.
- Some physiological methods have very high implementation demands (e.g., evoked cortical potentials).
- Literature shows contradicting results in respect to validity of most physiological methods.
- Physiological Methods can be influenced by physical and emotional factors.
- Restrictions on operator movements (e.g., using of a headrest when measuring eye movements).
- Data analysis can be demanding (e.g., with evoked cortical potentials).

Independence from tester influence

The objective approach prevents any influence of the tester on the results.

Validity of the methods

Measures of heart rate and heart rate variability

The results on heart rate are controversial (Lysaght et al., 1989). Heart rate has been shown to reflect mainly operator stress (Wierwille and Connor, 1983; Roscoe, 1987; Kramer, 1991), whereas validity is limited when emotional factors are not involved, e.g., with tracking tasks (Pfendler, 1982). Results on heart rate variability are also not consistent. In some studies certain measures of heart rate variability proved to be valid indicators of cognitive workload (Luczak and Laurig, 1973; Strasser, 1981; Mulder and Mulder, 1987; Wilson and O´Donnel, 1988; Kramer, 1991) whereas in other experiments they were less sensitive (Pfendler, 1982; Schick and Radtke, 1979).

Measures of eye activity

Measures of eye functions, such as eye movements and eye scanning, blink rate and blink duration, pupil diameter and pupil dilation provide measures directly related to visual workload (Radl, 1969; Beatty, 1982; Hallet, 1986; Hart and Wickens, 1990; Kramer, 1991). The dwell times give important diagnostic information about the sources of workload. The longer the dwell time, the more difficult to read an instrument. According to Lysaght et al., (1989) measuring of eye movements has the highest potential as a workload analysis technique among the physiological techniques. Pupil diameter has also demonstrated sensitivity to workload. But as there are strong measurement restrictions in respect to eye movements and to variations in light level the method is not recommended for field applications.

Measures of brain activity

Spectral analysis of the electroencephalogram (EEG) cannot be recommended for workload analysis (O`Donnel and Eggemeier, 1986; Hecker et al., 1980). In contrast, a series of studies could demonstrate that the Evoked Cortical Potentials (P 300 amplitude of the ECP) are relevant to perceptual task demands and are sensitive to aspects of cognitive processing (Israel et al., 1980; Kramer, 1991; Wilson and Fullenkamp, 1991). But the application of the method is as yet mainly restricted to the controlled laboratory environment (Lysaght et al., 1989).

Measures of body fluid analyses

Body fluid analysis is one of the few techniques available for the assessment of sustained or long-term effects of workload (Lysaght et al., 1989). Recent work has concentrated on salivary cortisol responses which follow levels of stress and workload (Kirschbaum and Hellhammer, 1989). Salivary cortisol levels significantly differentiated

between different levels of stress and workload. The method can also be used to assess anticipatory workload and workload consequences. Salivary cortisol level can be measured easily also under field conditions.

Other physiological workload measurement methods

Other additional physiological methods have been used to measure workload. Methods to measure blood pressure, electromyography (EMG), galvanic skin response (GSR), critical flicker fusion frequency (CFF), etc. are mentioned in the literature but the methods are not recommended for workload measurement (Lysaght et al., 1989).

Reliability of the methods

Estimation of reliability of workload measurement methods has generally been neglected. There are also only few data available on reliability of physiological measures of workload which will be reported in the following. Other empirical studies were not available.

Measures of heart rate and heart rate variability

Blitz et al. (1970) measured test half and retest reliability of heart rate and a measure of heart rate irregularity with a choice reaction task. They found test half reliability coefficients of r = 0.93-0.99 and retest coefficients of r = 0.31-0.79 for both kinds of measures which did not differ very much in respect to reliability. Retest reliability of five measures of heart rate and of heart rate irregularity was also measured by Pfendler (1982) in a tracking task resulting in very high reliability coefficients ($r > 0.9$), with the exception of one measure of heart rate irregularity.

Measures of brain activity

In a study from Hecker et al. (1980) reliability of measures from the spectral analysis of the electroencephalogram was calculated resulting in very low reliability coefficients ($r < 0.3$).

Feasibility of the methods

When physiological standard methods are used (e.g., ECG) the implementation demands are not so high, but much more higher than, e.g., with rating scale methods. For some physiological methods like those which measure the activity of the brain and of the eye implementation demands are extremely high exceeding all other methods in respect to instrumentation and training of the experimenter.

Economy of the methods

Physiological methods have much more higher costs than rating scales and secondary tasks, especially methods to measure the EEG, the ECP and the activity of the eye.

Face validity of the methods

Face validity is not a very relevant aspect for physiological workload measurement methods. Subjects attitudes in respect to the methods are not assumed to influence the results of physiological methods.

Interference of the methods

Physiological workload measurement methods may interfere with primary tasks as electrodes, wires etc. may restrict movements of the operator. Operator movements may also be restricted as movements can lead to artefacts (ECP). ECP measurement can also interfere with the main task when a secondary task is used to elicit the P 300 amplitude. When measuring eye movements or pupil size a headrest may be required which allows only limited movements of the head.

Diagnosticity

Measurement of eye movements shows good diagnosticity as the method can detect sources of workload when measuring high dwell times (e.g., for certain instruments) (Lysaght et al., 1989). All other measures are assumed to have low diagnosticity.

Generality

Application of most physiological workload measurement methods is limited. Their use is restricted, when operator movements are important for task performance: movements can produce artefacts (EEG, ECP), effects of movements can be confounded with effects of the experimental conditions (ECG, EMG) or restrictions are necessary to prevent movements (e.g., headrests when measuring the activity of the eye). Standard ECG technique can be implemented relatively easily and heart rate can be derived and used when physical workload is absent and emotional load is a main aspect of the study. Heart rate variability can be measured when cognitive load is varied. As the results of both methods are controversial it is recommended to use them only in addition to rating scale methods. Methods to measure eye movements are promising (Lysaght et al., 1989) but the implementation can be difficult in an operational environment. Another successful method is measurement of evoked cortical potentials (ECP) but its use can only be recommended in the controlled laboratory environment (Lysaght et al., 1989). Most of the other physiological methods are more useful for measurement of longer term effects of workload (e.g. CFF, salivary cortisol secretion).

References

Bartenwerfer, H. (1960). *Beiträge zum Problem der psychischen Beanspruchung.* Forschungsberichte des Landes Nordrhein-Westfalen, Nr. 808. Westdeutscher Verlag, Köln und Opladen.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Human Perception and Performance, 2*, 556 - 566.

Blitz, P.S., Hoogstraaten, J. and Mulder, G. (1970). Mental load, heart rate and heart rate variability. *Psychol. Forsch. 33*, 277-288.

Hallet, P.E. (1986). Eye movements. In K.R. Boff, L. Kaufmann and J.P. Thomas (Eds.), *Handbook of perception and human performance, Vol. I., Sensory processes and perception* (pp.10/1 - 10/112). New York: Wiley & Sons.

Hart, S.G. and Wickens, C.D. (1990). Workload assessment and predicition. In H.R. Booher (Ed.), *Manprint: An approach to systems integration* (pp. 257-296). New York: Van Nostrand.

Hecker, R., Schmidtke, H., and Wegener, H. (1980). *Reliabilität und Validität spektraler EEG-Parameter als Indikatoren der psychischen Beanspruchung*. Psychologia Universalis. Bd. 48. Meisenheim: Hain.

Israel, J.B., Chesney, G.L., and Donchin, E. (1980). The event related brain potential as an index of display-monitoring workload. *Human Factors, 22*, 211-224.

Kirschbaum, C. and Hellhammer, D. (1989). Salivary Cortisol in psychobiological research: An overview. *Neuropsychobiology, 22*, 150-169.

Kramer, A.F. (1991). Physiological metrics of mental workload: A review of recent progress. In D. L. Damos (Ed.), *Multiple task performance* (pp. 279-328). London: Taylor & Francis.

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

Lysaght, R.J., Hill, S.G., Dick, A.O., Plamondon, B.D., Linton, P.M., Wierwille, W.W., Zaklad, A.L., Bittner, A.C., and Wherry, R.J. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies* (Tech. Report 851). Fort Bliss, TX: U.S. Army Research Institute, Field Unit.

Mulder, I.J. and Mulder, G. (1987). Cardiovascular reactivity and mental workload. In R. I. Kitney and O. Rompleman (Eds.), *The beat-to beat investigation of cardiovascular function*. New York: Oxford.

O`Donnell, R.D. and Eggemeier, F.T. (1986): Workload assessment methodology. In: K.R. Boff, L Kaufman, and J. Thomas (Eds.), *Handbook of perception and human performance: Volume II. Cognitive processes and performance* (pp. 42/1-42/49). New York: Wiley.

Pfendler, C. (1982). Bewertung der Brauchbarkeit von Methoden zur Messung der mentalen Beanspruchung bei Kfz-Lenkaufgaben. *Zeitschrift für Arbeitswissenschaft, 36 (8 NF) 1982/3*, 170-174.

Radl, G. (1969). *Untersuchung zur Quantifizierung der psychischen Beanspruchung bei simulierten Fahrzeugführungsaufgaben*. Anthropotechnische Mitteilung Nr. 8/69. Forschungsinstitut für Anthropotechnik, Meckenheim.

Roscoe, A.H. (1987). In-flight assessment of workload using pilot ratings and heart rate. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload, AGARDograph No. 282* (pp.78-82). Neuilly sur Seine, France: AGARD.

Schick, F.V. and Radtke, H. (1979). *Untersuchung der Pulsfrequenzvariabilität als Schätzgröße der Pilotenbeanspruchung bei anthropotechnischen Experimenten*. DFVLR - FB 79-33.

Strasser, H. (1981). *Arbeitswissenschaftliche Methoden der Beanspruchungsermittlung. Beanspruchungsprofile unter dem Aspekt der Ausführbarkeit und Erträglichkeit menschlicher Arbeit.* Stuttgart: Gentner.

Wierwille, W.W. and Connor, S.A. (1983). Evaluating of twenty workload assessment measures using a psychomotor task in a moving-base aircraft simulator. *Human Factors, 25*, 1-16.

Wierwille, W.W. and Eggemeier, F.T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors, 35 (2),* 263-281.

Wilson, G.F. and Eggemeier, F.T. (1991). Psychophysiological assessment of workload in multi-task environments. In D. L. Damos (Ed.), *Multiple task performance* (pp. 329-360). London: Taylor & Francis.

Wilson, G.F. and Fullenkamp, P. (1991). A comparsion of pilot and WSO workload during training missions using psychophysiological data. In *Proceedings of the Western European Association for Aviation Psychology, Vol. II: Stress and error in aviation* (pp. 27-34). Western European Association for Aviation Psychology, England.

Wilson, G.F. and O'Donnell, R.D. (1988). Measurement of operator workload with the neurophysiological workload test battery. In: P. A. Hancock and N. Meshkati (Eds.), *Human mental workload* (pp. 63-100). Amsterdam: North-Holland.

4 Executive Summary

4.1 Measurement of Operator Workload

Workload can be defined as the human resources an operator expends when performing a specified task. Since the human resources are limited, a system must not overload the operator, otherwise severe performance decrements can occur. Furthermore, two systems with the same level of overall performance can impose quite different levels of workload on the operators. Therefore, it is necessary to measure workload as well as performance. Workload data can also aid decisions and trade-offs needed to insure that workloads are at acceptable levels across time and are optimized both for short term tasks and sustained operations.

Workload can be assessed by subjective methods, secondary tasks, physiological techniques and measures of primary task performance. Workload can also be evaluated by analytical techniques in the design phase of a system. Analytical techniques are not discussed here, since the goal of this Research Study Group is limited to empirical methods only. Measures of primary tasks will be discussed in the chapter on performance measurement. Since current physiological and secondary task methods are not considered as practical for most applications they will only be briefly discussed. The reasons for this are described in more detail in the Workload chapter (e.g., controversial results, limited applicability, high implementation demands, etc.) It is recommended that these methods should be used only as supplementary tools in workload measurements (e.g. for continuous workload measurement).

Subjective methods are used in human engineering testing and evaluation to assess the operator's, or observer's, rating of a task. These methods, especially those with rating scales, have many advantages and few disadvantages in measuring operator workload relative to other approaches.

Of the many available workload measurement approaches, four rating scales were especially recommended for application: (1) NASA-TLX (NASA Task Load Index), (2) SWAT (Subjective Workload Assessment Technique), (3) MCH (Modified Cooper-Harper Scale), and (4) ZEIS (Sequential Judgement Scale). In addition, the four methods are used by more than one nation. More important, however, is that they satisfactorily meet a number of necessary workload measurement standards. A brief description of each of the four methods and their respective applications are given below.

Obviously, other scales can be useful for workload measurement. As an example, two other promising approaches, SWORD (Subjective Workload Dominance Technique) and the Bedford Scale, are added. All the recommended workload measurement methods and their comparisons to specific standards are described in more detail in the Workload Chapter, which also includes the references relevant for the executive summary.

4.2 NASA Task Load Index (NASA-TLX)

Description of the method

NASA-TLX is based on the assumption that workload is a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance. Workload results from the interaction between the requirements of a task, the circumstances under which it is performed, and the skills, behaviours, and perceptions of the operator. Furthermore, it is assumed that a workload measurement method must be made up of multiple workload dimensions. This is why NASA-TLX consists of six subscales: Mental, Physical and Time Demands, Performance, Effort, and Frustration. Each of these bipolar subscales consists of a hundred point scale divided into twenty 5-point interval steps. The endpoints have verbal descriptors. Another important aspect of NASA-TLX is the development of an individual weighting procedure for combining the results of the different subscales to reduce between-subject variability.

Like other approaches, NASA-TLX requires the operator to have an adequate familiarization with the method. Before rating the real tasks the operator should apply the scale to a few practice tasks. After familiarization NASA-TLX consists of two steps. In the first step the operator rates the task with regard to workload. In the second step each operator makes a paired comparison for all 15 possible paired combinations of the 6 dimensions. This is done by deciding which of each paired element is more important with regard to workload for a given task. A Mean Weighted Workload Score is obtained by using the results of the paired comparison (weights from 0-5) to weight the 6 individual subscale scores of the rated tasks. It is recommended that software be used for scoring of raw data. There are no supports available for interpretation of results (e.g. information about tolerance levels, etc.). If two or more qualitatively different tasks are rated with regard to workload, a separate paired comparison has to be made for each task.

The weighting procedure of NASA-TLX has been criticised. In one study better differentiation and better reliability were achieved using a Mean Unweighted Workload Score. In another study the weighting procedure of NASA-TLX was described as ineffective with a recommendation to simply ignore it.

***Advantages of the method*:**

- Good face validity, and good rater acceptance.
- General Applicability
- The American version of NASA-TLX proved to be more valid than most other workload measurement methods.
- Interval scale level is assumed. Parametric tests can be used.
- NASA-TLX can also be employed in a prognostic way.
- There are no floor effects with NASA-TLX.
- Dutch, English and German versions of the scale are available.

***Disadvantages of the method*:**

- Translated versions of NASA-TLX (German and Dutch) showed lower sensitivity than a German (ZEIS) and a Dutch scale (BSMI).
- The original instructions are too extensive.
- The six subscales of NASA-TLX partially show significant intercorrelation.
- In some experiments the subscales "frustration" and "physical demands" only show a small relevance for workload.
- When different subjects give a weight of zero to different subscales the Mean Weighted Workload Scores of these subjects are strictly seen as no longer comparable.
- During task performance NASA-TLX might cause interference with the main task.

Applications:

NASA-TLX can be applied to a very broad range of tasks. It should be used when there is a special interest in diagnosticity, i.e., in detecting the sources of workload. This is possible by using the six subscales of the method. NASA-TLX has good face validity and construct validity. In comparative evaluations with other rating scales and with lower workload levels NASA-TLX proved to be the most sensitive technique. When used while performing tasks NASA-TLX might interfere with the main task, especially in high workload situations. More detailed information, references and comparisons with additional factors (validity, reliability, etc.) are provided in the Workload Chapter.

4.3 Subjective Workload Assessment Technique (SWAT)

Description of the method

SWAT is a subjective workload measurement method in which the operators rate the workload of a task with respect to the subscales of "time load", "mental effort load", and "psychological stress load". Each of the three subscales has three levels in the form of a category scale. Verbal descriptors anchor each level of each of the three dimensions.

There are two phases in the application of SWAT: the scale development phase and the event-scoring phase. In scale development, the operator has to rank order all 27 possible combinations of the three levels of each of the three dimensions according to his perception of the increase in workload. For this purpose he uses 27 cards, from which each gives a description of one of the possible combinations. After that, a test is performed with a computer program to see if the card sort is in accordance with the mathematical axioms required for using the conjoint measurement and scaling procedure. Some axiom violations are allowed however. After testing, the SWAT technique uses conjoint analysis to convert the rank data of the card sort into an interval scale from 0-100. This results in each of the 27 combinations having a fixed value on the 100 point scale. The rather complex data processing procedure of SWAT requires the use of software. In some studies, where the scale development phase has been dropped, the results show that SWAT can also be applied effectively without this procedure.

If there is enough agreement among the card sorts within a group of subjects then a single group scale can be constructed. If this is not the case a procedure called SWAT prototyping must be applied to provide greater precision of measurement. The 6 prototypes (TES, TSE, ETS, EST, STE, SET) are characterized by hypothetical card sorts which place different emphasis on the three SWAT dimensions of time load (T), mental effort load (E), and psychological stress load (S). For the TES prototype, e.g., time load is the most important workload driver, followed by effort, and then psychological stress. Subjects with the highest correlation in one prototype are grouped together and a common scale solution is found.

In the event scoring phase, the subject evaluates the relevant task with regard to time load, mental effort load, and psychological stress load (e.g., 2, 1, 3). The scale value associated with this combination (e.g., 30.5) is the dependent variable used in the subsequent data analysis.

***Advantages of the method*:**

- SWAT measures on interval scale level.
- Event scoring can be done while performing the main task but not during critical workload periods.
- SWAT can be used prognostically.
- SWAT can be applied with a very broad range of tasks.
- SWAT can be used when time and resources are critical.
- Records of application in laboratory and field settings are published.
- "Redline" values have been developed (scores where performance tends to degrade).

Disadvantages of the method:

- SWAT dimensions are derived intuitively.
- There are significant intercorrelations between all SWAT dimensions.
- SWAT training and card sort requires about 1.5 hours per subject.
- Card sort can be problematic: It requires good verbal abilities. Motivation is sometimes low.
- The specifications are imprecise in respect to number of axiom violations allowed.
- SWAT has low sensitivity at low workload levels.
- SWAT requires experimenter decisions to specify scale type and adequacy.

***Applications*:**

SWAT can be applied with most tasks. But in contrast to NASA-TLX the method has been found to be not so sensitive in low workload situations. Since SWAT has only three subscales it can also be used during task performance with low risk of interference with the main task. Information about the sources of workload is limited. Critical SWAT values for predicting operator overload exist. Validity and face validity of SWAT are rated lower than those of NASA-TLX. Studies demonstrate that the complex and cumbersome data analysis of SWAT can possibly be dropped without any disadvantages. More detailed information, a list of SWAT applications, references and comparisons with additional factors (validity, reliability, etc.) are provided in the Workload Chapter.

4.4 Modified Cooper-Harper (MCH) Scale

***Description of the method*:**

The Cooper-Harper Scale has been used successfully for evaluation of aircraft handling tasks and other motor tasks. In more advanced complex automated systems the human operator is more occupied with activities like perception, monitoring, evaluation, communication, and problem solving. Therefore, the original Cooper-Harper Scale has been modified so that it can be utilized with tasks in today’s modern systems. The derived Modified Cooper-Harper (MCH) Scale retains the decision tree of the original scale but has changed the wording. The rating is based on subjective evaluation of mental effort and attained performance. The authors assume that the scale gives a reliable overall assessment of workload on a relative basis.

***Advantages of the method*:**

- The decision tree makes the rating easier by allowing sequential decisions.
- The rating is not very time consuming and can be used while performing a task.

***Disadvantages of the method*:**

- The MCH Scale measures on ordinal scale level.
- The scale's implication that the two dimensions of performance and effort are directly related is problematic.
- The scale's implication that low workload is highly desirable is problematic.
- The scale has no diagnosticity .

Applications:

MCH has been designed for workload assessment of tasks where the operator is involved with activities like perception, monitoring, evaluation, communication, and problem solving. The scale has demonstrated sensitivity in such tasks during simulated flight but has been less useful in other environments. It gives an estimate of overall workload but no diagnostic information. In comparative studies the sensitivity of MCH has been rated lower than that of NASA-TLX or SWAT. More detailed information, references and comparisons with additional factors (validity, reliability, etc.) are provided in the Workload Chapter.

4.5 Sequential Judgement Scale (Zwei-Ebenen Intensitäts-Skala, ZEIS)

Description of the method

The Sequential Judgement Scale has been designed in order to rate the difficulty of vehicle handling tasks experienced by drivers. The scale is constructed according to 14 scale design guidelines compiled from theories, ergonomic design principles, and experimental results that point to factors having a positive effect on reliability and validity of rating scales.

The rating scale measures task difficulty and can be applied to a very wide variety of tasks. The scale also is useful as a measure of workload, since the subjective experience of task difficulty is an important dimension of workload. The rating is facilitated for the operators by the two level scale design requiring them to make two judgements in sequence, first a coarse judgement, and then a second finer one. The first judgement is made according to three basic categories "difficult", "medium" or "easy". The first level choice graphically switches subjects to specific instructions, and then to a smaller appropriate section of the continuous scale of the second level in order to make a finer rating by marking the line at the appropriate point. However, operators are also permitted to cross over to the adjacent section. Questions and user guidance instructions were integrated into the scale as far as possible to assure that all operators receive identical instructions and that they are available at the right time. In this way independence of the scale from tester influence is increased. The instructions in the graphics that precede the scale have to be tailored to each task. Scoring of the scale is done by measuring the operator's mark from the scale's right end point and by transferring the measure into a percentage score of the whole scale. When the operator's mark on the scale is measured from the scale's left end point (like with other scales), high rating scores mean low difficulty level, which is in contrast to other scales.

***Advantages of the method*:**

- Workload can be measured during task performance since event scoring requires little time.
- Dutch, English and German versions of the scale are available.
- ZEIS has good rater acceptance.

***Disadvantages of the method*:**

- ZEIS ratings have no diagnosticity.
- Only limited information on validity of the scale is available.

Applications:

ZEIS is assumed to be applicable to a wide variety of tasks where task difficulty is a relevant dimension. The scale gives an overall estimate of workload without any diagnostic information. In comparative evaluations the validity of ZEIS has been relatively high, but, the validations have been, thus far, restricted mainly to motor tasks. Because it is a unidimensional method, ZEIS can possibly be used during task performance without interfering with the main task. More detailed information, references and comparisons with additional factors (validity, reliability, etc.) are provided in the Workload Chapter.

Chapter 4

Human Task Performance Measurement

1. The application of human task performance data

There are three major categories of human engineering test data: (1) user acceptance, (2) engineering measurement and (3) task performance. Of these, user acceptance data can tell us of the degree to which users accept and think they perform well with the equipment; engineering measurement can tell us whether applicable design standards have been met and good design practices employed; but only measurement of task performance can tell us whether the ultimate human engineering design goal -- an effective and efficient human machine interface -- has been achieved. Success in meeting that goal must ultimately be expressed in terms of successful performance of tasks.

A. Influence system design

The primary utility of human task performance data is to influence system design throughout the development process, including before testing and evaluation is done. If developers know in advance what tasks will be exercised in testing and what standards must be met, they can concentrate upon design characteristics that influence performance of those tasks. When they also are persuaded that the customer is serious about meeting the design goal, is willing to pay for it and to impact the program if it is not met, there is an incentive to place appropriate emphasis on human engineering design.

B. Verify adequate performance under appropriate conditions

If that incentive results in the appropriate emphasis, testing and evaluation may simply serve to verify the success of the design efforts. If it does not, testing and evaluation data reveal design characteristics that degrade task performance, and point the way to areas in which redesign can improve human task performance and, consequently, total system performance.

C. Identify human performance effects on total system performance

When testers and evaluators neglect measuring task performance and rely only upon measures of overall system performance, we give up the ability to learn whether the limiting factor on system performance is equipment performance, human performance, or both. There are known, well-documented limits on certain human performance parameters. Those limits are not efficiently exceeded by personnel selection or training. Task performance data are necessary if we are to know whether we have challenged the limits. They are also needed to identify the human contribution to the performance budget, and conversely to the error and latency budgets. It is only with that information that we can know what approach to take in improving system performance to meet criteria.

D. Provide an objective, valid basis for making acquisition decisions and for modeling and simulation to support all the above

It is always desirable for task performance criteria to have been established before testing, making determining whether they have been met a simple go or no-go decision. The decision-maker's task is quite straightforward if human task performance criteria are satisfied, and overall system performance criteria have been met. But even without predetermined criteria, knowledge of the performance achieved is a useful basis for decision-makers to use in determining how good is good enough. It can help them to avoid the pitfall of deciding, for example, to "train around" inadequate performance if task performance data illuminate the fact that further improvement in task performance is unlikely because the recognized limits for performance of certain tasks have already been reached or approached. This would point developers in the direction of a human engineering design or other improvement rather than training as a solution.

Performance data on well-defined tasks is also useful -- even necessary -- as inputs to modeling and simulation exercises to predict future performance and to decide upon future testing and evaluation priorities. More and more, the reductions in military budgets in all the NATO nations drive us to cut acquisition costs, including costs for testing and evaluation. One response to this pressure is to use modeling and simulation technology to sharpen the focus of our selection of tasks to exercise in testing and evaluation. Modeling of human performance requires validity and precision in our data on human performance of tasks of interest.

2. What data to collect

A. Identify and define tasks

If tasks are defined in terms of their initiating event, their terminating event and a criterion for success, the tester and evaluator have all information necessary to evaluate performance. Data describing human task performance fall into two categories: performance time and error rate. Performance time is defined as the time between the occurrence of the initiating even and the terminating event for the task; error is defined as failing to meet the criterion for success. Several parties may have inputs to the selection of tasks to measure, but identifiable tasks should be ranked according to their importance to successful system performance along such dimensions as: frequency of occurrence, consequences of failure, consumption of the system performance time budget, contribution to the overall system error budget, and other dimensions. It is rare that the luxury of measuring performance on all the tasks of interest is available, so participation by the end users of the evaluation report -- acquisition decision makers -- is very important.

B. Agree upon task performance criteria

In an acquisition program in which the proper initial planning and analysis have been done, criteria for human task performance will have been established before testing, and determining whether they have been met can be a simple go or no-go decision. If human task performance criteria are satisfied, and overall system performance criteria have been met, the decision-maker's task is unambiguous. But even without predetermined criteria, knowledge of the performance achieved is a useful basis for decision makers to use in determining how good is good enough. Agreement upon appropriate performance criteria is often difficult. The user's representative is sometimes unable to specify how much performance is required and asks for something like, "as much as is possible," a difficult criterion to which to test. The materiel developer feels pressure to accept too ambitious performance criteria to keep the customer happy, but may underestimate what is feasible to maximize the chances of meeting requirements. Thus, the requirement for involvement of the decision authority is essential.

C. Exercise tasks

Determining a scenario within which to exercise the equipment functions and the task performance that are of interest to all parts of the testing and evaluation community is often a compromise among the parties involved, limited by the test costs that are sustainable. But, in particular for human tasks, it is important that the conditions under which the data are collected represent the conditions in which the system is expected to operate, and enough iterations of exercise of the tasks must be available to satisfy statistical analysis requirements.

D. Measure performance

As mentioned before, the two major categories of human performance data are performance time and error rate. It is critically important that both be measured because of the self-correcting nature of most human task performance. That is, given infinite availability of time, perfect or near-perfect performance, expressed as a zero error rate, can be expected; given a requirement to perform as quickly as possible a higher rate of error can be expected. Stated another way, one aspect of performance can almost always be traded off for the other. In some situations, the tester encounters a series of sequential tasks in which the availability of one task requires correct performance of a preceding task, thus "error" cannot easily be measured. In this situation, one might argue that the error rate has been established as a constant, zero, and the only remaining variable is time. That is a useful argument, in that it allows testing to proceed, but what probably is happening is that any errors being made simply require repetition of the attempt to perform, and the number of repetitions is reflected in performance time.

If it is important to know what is causing the performance to be slower than expected or required, the tester may still be obliged to attempt to gauge error rate.

1. Time, latency, speed

We can measure and express the temporal part of the requirement as the simple difference in time between the start of the task and its end, but may also include such things as latency between presentation of a stimulus and a response, or repetition rate on repetitive tasks.

2. Error rate, size (rms error on tracking tasks), frequency, probability

We can measure and express the error part of the requirement as the error rate per unit of time, number of repetitions, probability of occurrence, or any and all other ways that satisfy the evaluator's and decision authority's needs for information.

E. Compare with criteria in B.

If available before the test, properly stated task performance criteria can make analysis and evaluation a simple and straightforward process. If they are not available, then the performance achieved is reported to the decision authority, who must then decide whether what has been achieved is good enough.

F. Describe relation to total system performance

This may be one of the most difficult requirements placed on the human engineering tester and evaluator, but is certainly one of the most important if the human engineering evaluation is to impact the decision process. For the military decision-maker, information about human performance achieved on certain tasks may be of academic interest only, unless the effect of their performance on total system performance can be demonstrated. Where human-machine interaction occurs in system performance, it can be of paramount importance to know whether the success of the function in which they interface is limited by human performance or by equipment performance, particularly if that function is identified as one which limits overall system performance.

A (not entirely) mythical exchange between a small caliber weapons designer and the human engineering tester may be helpful as an example. In attempting to improve the accuracy of a sniper rifle, the hardware engineer has data that indicate that thickening the metal barrel of the rifle increases its stiffness and its ability to retain a straight bore and the alignment of its sights. This, the modelers predict, will reduce the round-to-round dispersion by about 30 seconds of arc at a weight penalty of two kilograms or about four and a half pounds. When the prototype is tested, the rounds fired from a fixed machine rest confirm the model's prediction, but when fired by expert snipers who work for the proving ground, the round-to-round dispersion increased by about one and a half minutes of arc above that of the baseline rifle. With his knowledge of human perception limits and performance characteristics, the human engineering tester points out that the resolution of the human foveal (best) vision is limited to about one minute of arc, and that the addition of the extra weight and the disturbance of the balance of the rifle added at least another half minute of human contribution to the round-to-round dispersion due to increased jitter. Thus the hardware contribution to the error budget already was so much less than the human contribution that the design improvement to the hardware resulted in poorer overall human-machine performance.

Another, perhaps more important requirement for relating human performance to overall system performance is that the increasing budget pressure on the acquisition process forces us to compete with other disciplines not only for the ear of the acquisition decision makers but for their money, as well. Unless we can show value added in terms of improving system performance, controlling or reducing life-cycle cost, or both, we may not only fail to influence system design, but cease even to be funded to participate.

3. How to collect it

A. Task performance time

1. Manual input

The earliest attempts to measure task performance were part of the time-and-motion studies conducted by industrial psychologists using manual means of recording such as stop watches and clipboards. There are situations in which technology no more sophisticated than this still is used and is used appropriately and effectively. However, its temporal resolution is limited to the reaction time of the human data recorder, and determination of error rate is limited to his single opportunity observation of task performance. For most purposes, we can do better than that.

2. Event recorders

Electronic event recorders that record events either as an excursion of a pen on a strip chart or an incremental rotation of a counter wheel can sense their data electrically, thus escape the reaction time limitation of the human-with-stopwatch. Because a strip chart recording can be analyzed after the event with the convenience of a time scale, some of the single opportunity observation errors can be corrected.

3. Digital data bus recording

On a large proportion of major systems being developed over the past few years, digital electronic system control features are built into the system, usually employing a high-speed digital data bus for communication among sensors, effectors and displays. Tapping into that data stream for recording test data can provide a very rich source of data on what controls were operated and when and how, and what digital information provided a stimulus or represented the beginning of a task for humans to perform. Used to its greatest extent, it can provide information that is very detailed on a large number of events in system operation, with time resolution down to the microsecond level. The capability it offers can easily overwhelm the tester with the sheer volume of data he can record. Unfortunately, its greatest asset -- this large quantity and rate of recording of high-resolution data -- is often also its greatest liability. If the tester does not identify in advance the set of events on which he needs data and write appropriate software to capture only those data points, he may later find it difficult and expensive to extract the data of interest from those in which it is embedded. It is sometimes difficult for an evaluator to resist requiring a tester to "record everything we can" just to ensure that he doesn't discover after the test is concluded that he left out collection of data of interest. That requirement has often left testers searching almost endlessly through irrelevant data for the small fraction of it that answers valid test issues.

4. Time-tagged video

Color television imagery requires a wide frequency response and a large dynamic range. The video cassette recorder's (VCR) capability to capture that range offers testers an unprecedented opportunity for recording several other high-resolution, time-synchronized channels of audio and other analogue data beyond video imagery. Current camera technology that fits within a package about the size of a matchbook will provide 380 to 450 lines of resolution in a full-color image. In light levels down to 2.5 lux, using "pinhole" lenses that give a depth of field that is sharp from 1.5 inches to infinity, it requires no adjustment of any kind, and can provide a field of view of 90 degrees or more. If monochrome imagery is satisfactory, even lower light levels can be recorded successfully. And unlike a human eye, sensing surfaces can be selected that record energy well outside the visible spectrum at either end, allowing recording of imagery that a human observer could not see directly.

There are many readily identifiable examples of the advantages of using recorded video as a substitute for direct human observation. One is the ability of repeated observation by multiple observers to determine subtle characteristics of performance that would never be noticed by direct observation. The tester can record unobtrusively under conditions that would necessarily degrade a human observer's performance and can sense activities that would be entirely invisible to the human eye, using energy outside the human visible spectrum.

With time-tagging, in which a master clock display can be recorded within the imagery frame, precise determination of when certain events took place, and synchronization with other sources of data such as the high-speed digital data bus, is available. One proper use of the capabilities of the digital data bus would be to identify events of interest which are part of its data stream and use the digital data stream to record a time tag on the video tape to identify events of interest on the video. Like the digital data bus, the video recording capability can readily overwhelm testers with the sheer volume of information recorded, unless provisions are made to identify for reduction and analysis only those parts of the video recording that answer test issues. Experience has shown that reduction of video data can be very time-consuming. It takes about three hours of reduction time for each hour of recording. The amount of time and effort consumed by data reduction can be controlled by proper pretest planning and judicious use of time-tagging from the digital data bus and other sources, so that only the events that answer test issues are reduced.

The popular formats of video tape also provide several channels for recording of different audio channels or any other analogue source that is of interest. There are very reasonably priced VHS recorders which make available at least four analogue channels. These are available via the stereo pair of linear sound tracks that can be recorded on one edge of the tape, and the two high-fidelity stereo channels multiplexed onto the video signal and recorded by the same helical scan heads that record video information. These latter two channels have particularly high information recording capacity because they have the same frequency bandwidth and dynamic range potential required for recording color video imagery. That yields sonic performance that rivals that of digital audio technology. In addition, the part of the imagery signal used to provide second audio programming (SAP) for bilingual households and that used to provide closed captioning for the hearing impaired also offer time-synchronized recording opportunities using VCR technology. This promises a large array of options for recording several time-tagged, synchronized channels of analogue information along with the video image. This permits the tester to capture a wide variety of events of interest all on a single tape cassette with two to eight hours of recording time at the cost of as little as $2 - $3. The increasing availability and decreasing cost of digital video recording technology promises even more options and flexibility.

B. Error rate

Since the tester should be recording both performance time and error rate on tasks of interest, the same technologies described above can record both kinds of data; it is essentially only the reduction and analysis that may differ.

4. Interpretation and Evaluation

Ideally, the tester and evaluator will have been an integral part of the development process from the beginning. If that has happened, an opportunity has been afforded for stating criteria for performance of tasks in terms of performance time and error rate for each task that addresses a test issue. If so, interpretation of the data is unambiguous, and the evaluator has only to evaluate the impact of the performance of the task upon total system performance. With thorough planning at the beginning of the development process, the expected impact of performance of certain tasks on overall system performance may also have already been determined as a part of the total performance time and error budget of the human-machine system, giving even the evaluation a predictable outcome. However, in the (current) real world of testing and evaluation, it is unlikely that such predetermined criteria will exist. Testers may be required simply to report their data to evaluators and decision-makers, with the latter left to their own wisdom to decide how good is good enough. Unless the tester or evaluator estimate the impact of performance of tasks on system performance based upon analysis of the human-machine system's performance time and error budget, the evaluation standards become quite subjective.

This page has been deliberately left blank

Page intentionnellement blanche

Chapter 5

User Opinion

Some background information is in order to avoid confusion on what is meant by user opinion for the purposes of this chapter. In the early drafts of the Exploratory Group's Terms of Reference document, the label given to this category of data was user acceptance. In later drafts it was referred to as subjective judgment. The presently used term was adopted to avoid a connotation of the latter that was perceived by some members of RSG.24 to be negative. The present term, user opinion, is used to avoid confusion, since there is at least one member nation, all of whose field testing activities are called user acceptance trials. For the purposes this chapter, what is meant is better described by the current label: we want to describe how to measure test participants' judgments and feelings about a system and its performance, such as might be collected via questionnaires, interviews and so forth.

This chapter briefly summarizes the different types of questionnaires and ways that questionnaires may be administered. Detailed guidelines regarding what to do in a given situation are included in subsequent sections. Issues to consider when deciding whether to use a structured interview or some other type of questionnaire are discussed. Both structured interviews and other types of questionnaires have their place; each has strengths and limitations that must be taken into account when identifying which instruments to use.

Types of Questionnaires

There are a number of techniques of data collection that can be used to measure human attributes, attitudes, opinions, and behavior. Attitude and opinion are closely aligned if not overlapping. Opinions are restricted to verbalized attitudes. Attitudes are sometimes unconscious or nonverbalized. Some of the methods of data collection are observation, personal and public records, specific performances, sociometry, interviews, questionnaires, rating scales, pictorial techniques, projective techniques, achievement testing, and psychological testing. For this chapter, however, attention has been restricted to a more limited number of data collection techniques: certain paper-and-pencil types of instruments broadly classed as questionnaires, and including only some of the techniques mentioned above. A distinction has also been made in this manual between open ended questionnaire items and closed-end items. Open-ended items are those which permit respondents to express their opinions in their own words, and to indicate any qualifications they wish. The amount of freedom the respondent will be given in expressing an answer to an open-ended item is partly determined by the questionnaire designer. Closed-end items use response alternatives. Respondents are directed to select one or more of the response alternatives from a closed set. Closed-end items frequently used are multiple choice, true-false, checklist, rating scale, and forced-choice. Survey items have been roughly classified into two groups: open-ended items and closed-end items.

It is common to use interview surveys to ask questions and record answers. Structured interviews are included within the definition of questionnaires used, since typically an interview form is developed and used by an interviewer both for asking questions and recording responses, much like a self-administered questionnaire. On the other hand, the unstructured interview makes no use of structured data collection forms. The interviewers are permitted to discuss the subject matter as they see fit with no particular order or sequence. Of course, other interviews fall somewhere between these two extremes. In any case, unstructured interviews, where no structured response forms are used, are <u>not</u> included within the definition of questionnaires used in this chapter.

Ways That Questionnaires Can Be Administered

There are a number of respects in which questionnaire administrations may vary. However, in the usual field test settings, the typical questionnaire administration situation involves paper-and-pencil materials with the author/test officer administering the questionnaire face-toface with a group of test players or evaluators.

Group Versus Individual Administration

Given a printed questionnaire, calendar time is saved by group administration. Group administration allows the opportunity for a questionnaire administrator to explain the survey and answer questions about items. The task of statistical analysis can be initiated with less delay than if one were waiting on a series of individual administrations. An important determinant of group vs. individual is the time at which people complete their participation in the test. Most often all participants are through at the same time. All would be available for questionnaire administration as soon as they could be brought to an appropriate place or places. Prompt group administration gives the same short amount of time for forgetting about test events by those who become the respondents. Group administration generally has a high cooperation rate. If there is an administrator, his/her time is conserved directly in proportion to the number of respondents he/she has in each administrative session. An advantage of group administration is low cost.

Author-Administered Questionnaires

When the test officer or administrator who is familiar with the content of the questionnaire and the test's purposes and objectives can administer the questionnaire, some advantages can be gained. The administrator's instructions and appeals may increase the number of respondents having desirable motivation to complete the questionnaire by giving appropriate consideration to each item. If one employs a self-administration procedure, such as might occur in a mailed-out questionnaire, or if a poorly prepared stand-in plays the role of administrator, then the respondents must derive their instructions and some of their motivation from printed instructions (or from the poorly prepared stand-in). More things usually can end up going wrong when questionnaires are self-administered than when they are administered by a test administrator.

Remote Administrations

From the test officers' point of view, remote administration refers to a questionnaire administration event that they cannot conduct because of its distance from them and/or other demands on their time. This dimension, remote versus face-to-face, is similar but not identical to the previously noted dimension, self-administered versus author administered.

To avoid the possible disadvantages of self-administered questionnaires, the test officer must be able to afford another administrator, train him/her in the knowledge and skills associated with effective administration, and transport him/her to the "remote" administration location. If multiple administrations having location or timing differences which preclude the same administrator from handling them are required, it would appear that the chances are increased that more respondents will experience more "difficulties" in answering the questions. For this type of questionnaire administration, the questionnaire itself would require careful design associated with items and instructions.

Other Materiel Modes

Providing the respondents with a printed questionnaire form, and a pencil to mark/write their responses, is the most common questionnaire administration procedure in field evaluations. In addition, other presentation modes have been used. In a card-sorting procedure that has been used with individuals and groups, each respondent reads statements of candidate problems and then places the card into the appropriate pile according to his/her judgment of the severity of the "problem." Rarer because of expense and logistics problems is the setting up of a computer terminal where each respondent enters (types in) answers to questions that are displayed on a cathode ray tube (or other computer display device).

Structured Interviews Versus Other Types of Questionnaires

Issues to Consider

When deciding whether to use a structured interview or another type of questionnaire, a number of issues should be considered.

Included are the following:

a. To develop questionnaire items, a focus group may be interviewed. Their comments can be used to develop hypotheses and refine questions. This information can be adapted to an interview guide and interview items.

b. Interview items should not use a dichotomous response set. Multiple choice and open-ended questions provide the opportunity for probing.

c. If a structured interview is used, there must be enough qualified interviewers to expeditiously process all interviewees. Sometimes there are only a few personnel to be interviewed, or there is plenty of time available for interviews, so only one or two interviewers will be necessary. In other situations, maybe only an hour or so may be available per interviewee; in these cases, a large number of qualified interviewers must be available.

d. Face-to-face interviews have a higher response rate than mail surveys.

e. In most cases, respondents have a greater tendency to answer open-ended questions in an interview than when response is by paper and pencil.

f. It is possible to adapt face-to-face interview guides for telephone surveys. Oral labeling of the scale points should be assessed on a pilot survey to be sure that the responses are not biased by the oral presentation of the scale.

g. Telephone interviews are faster to perform than mail surveys.

h. Interviews conducted by telephone require an interview structure that promotes a high interaction between the interviewer and respondent.

i. Group-administered paper-and-pencil questionnaires may be less expensive, more anonymous, and completed faster than the same number of interviews.

j. Respondents seem to be less likely to report unfavorable things in an interview than in an anonymous questionnaire. Typically, questionnaires are also more likely than interviews to produce self-revealing data.

k. Issues involving socially acceptable or unacceptable attitudes and behaviors will elicit more response bias.

1. During interviews, respondents often have a tendency to try to support the norms that they assume the interviewer adheres to.

m. Interviewers with biases on the issues under discussion may reflect them in the content they record, as well as in what they fail to record.

n. Ethnic background differences between interviewer and respondent probably will not influence the survey results unless the items have a racial content or are found to be threatening.

o. Although a structured interview using open-ended questions may produce more complete information than a typical questionnaire containing the same questions, empirical research seems to indicate that responses to the

typical questionnaire are more reliable; i.e., more consistent. Structured interviews using closed-end questions appear to be as reliable as paper-and-pencil questionnaires.

p. It may be difficult to code a combination of open-ended and closed-end items for interview surveys. (See Section XIII-B, Scoring Questionnaire Responses.)

Combinations of Methods

There are some situations where a combination of methods of questioning might be used:

a. An interview might be used to obtain information for designing a paper-and-pencil questionnaire.

b. Personal interviews or telephone interviews might be used for respondents who do not return questionnaires administered remotely (such as mail questionnaires).

c. When respondents are unable to give complete information during an interview, they can be left a copy of a questionnaire to complete and mail in, so that the necessity for a call-back is eliminated.

Conclusion

Both structured interviews and other types of questionnaires appear to have their advantages and disadvantages. The choice of which to use may well depend upon costs, which are generally lower for the typical questionnaire. The typical questionnaire is apparently more reliable, while the structured interview may provide more unique and more abundant information. If the dimensions of a problem have not been explored before, the best compromise would appear to be to use the interview approach with open-ended items to uncover the dimensions, and follow this by the use of the paper-and-pencil questionnaire with closed-end items to obtain more specific information.

Content of Questionnaire Items

The recommended general steps in preparing a questionnaire include preliminary planning, determining the content of questionnaire items, selecting question forms, wording of questions, formulating the questionnaire, and pretesting. As part of preliminary planning, the information required has to be determined, as do procedures required for administration, sample size, location, frequency of administration, experimental design of the field test, and analyses to be used. Selecting question forms is a function of the content of the questionnaire items and requires knowledge of types of questionnaire items and scaling techniques. The wording of questions is the most critical and most difficult step. Formulating the questionnaire includes formatting, sequencing of questions, consideration of data reduction and analysis techniques, determining basic data needed, and insuring adequate coverage of required field test data. Pretesting involves using a small but representative group to insure that all questions are understandable and unambiguous
Types of Questionnaire Items

There are many types of questionnaire items: open ended items, multiple choice items, rating scale items, behavioral scale items, ranking items, forced choice and paired-comparison items, card sorting items and tasks, and semantic differential items.

It may be noted that a number of ways have been utilized in the professional literature for differentiating and classifying item types. Which types are special cases of other types could be debated at length. Unanimous agreement with the definitions used in this section cannot, therefore, be anticipated.

Attitude Scales and Scaling Techniques

At times, questionnaire developers will wish to treat the total group of items on a questionnaire as a single measuring scale, and from them obtain a single overall score on whatever they are in interested in measuring. This is a common practice, especially with the measurement of attitudes. A typical attitude scale is composed of a

number of questions or statements selected and put together from a much larger number of questions or statements according to certain statistical procedures. Some of these procedures, called scaling techniques, are discussed in this section.

A distinction is needed, however, between two ways in which the term scale is used. An attitude scale could be constituted of items each one of which employs a response scale. A component of score could be achieved on each item. Adding these item scores together - which means considering the whole set of items as a scale - produces a total attitude score for the individual respondent.

There are, generally speaking, two general methods for the construction of scales such as attitude scales. The first method makes use of a judging group and one of the psychological scaling methods developed by Thurstone. It results in a set of statements being assigned scale values on a psychological continuum. The continuum may be favorableness-unfavorableness, like-dislike, or any other judgment. The psychological scaling methods, therefore, have considerably greater application than for the scaling of attitudes. They can be used to scale statements or objects. They have been used, for example, to determine the perceived favorableness of words and phrases commonly used as rating scale response alternatives.

The second general method is based on the direct responses of agreement or disagreement with attitude statements and does not result in a set of statements being assigned scale values on a psychological continuum. Both the Likert and Guttman scales are examples of this latter method.

Preparation of Questionnaire Items

Once a decision has been made regarding the type or types of items that are to be used in a questionnaire, attention must be given to the actual development of the items. This involves the following development topics: mode of questionnaire items; wording of items for both question stems and response alternatives; difficulty of items; length of question stem; order of question stem; number of response alternatives; and order of response alternatives.

A distinction has been made between a questionnaire item, a question stem, and response alternatives. A questionnaire item has both a question stem and response alternatives. The response alternatives are the answer choices for the question. (They are sometimes called "options.") The question stem is that part of the item that comes before the response alternatives.

Considerations Related to Questionnaire Administration

Questionnaire administration issues include: anonymity for respondents, motivational factors related to questionnaire administration, administration time, characteristics of administrators, administrative conditions, the training of raters and other evaluators, and other factors related to questionnaire administration.

Pretesting of Questionnaires

Even the most careful screening of a questionnaire by its developer or by questionnaire construction experts will usually not reveal all of its faults. Pretesting is an important and essential procedure to follow before administering any questionnaire. Its purpose is, of course, to find those overlooked problems and faults that would otherwise reduce the validity of the information obtained from the questionnaire responses.

Pretesting may seem to some uninformed individuals to be a waste of time, especially when the author may have asked several people in his or her own office to critique the questions, or perhaps even asked a questionnaire specialist to critique it. However, pretesting is an investment that is well worthwhile. It is crucial if the decision that will result from the questionnaire is of any importance.

Evaluating Questionnaire Results

An extended discussion on evaluating questionnaire results is outside the scope of this chapter. There are some factors related to the evaluation of questionnaire results that should be noted since they may influence how questionnaires are designed and developed. One is the scoring and coding of questionnaire responses, and the other concerns data analyses.

Scoring Questionnaire Responses

Practical Considerations

a. Planning the questionnaire in line with scoring and tabulation requirements can save both time and money. The phrasing of questions and their sequencing and layout affect tabulation time. For example, it is advantageous to have data coded and entered for analysis directly from edited questionnaires. Questionnaires consisting of only closed-end items will have a lower level of error for data entry than open-ended items. This is a more cost-effective approach. However, there are some drawbacks such as greater difficulty in verifying the coding and greater data entry time than when using a coding sheet.

b. A decision should be made ahead of time regarding whether the data will be tabulated by hand or machine.

c. Response alternatives should be precoded whenever possible. Codes for open-ended items are more difficult to construct than codes for closed-end items. To develop open-ended item codes, list out possible responses to the item. Pretest the questionnaire to classify responses to open-ended items. Construct a classification system and code. Pretest the code and revise as necessary. Develop a separate code for responses that were not possible to fit into the classification system above.

d. Codes need to be developed which guide coders in assigning code numbers to each answer. This includes the following: codes for missing data for item nonresponse, codes for item responses that are uncodable due to poor respondent performance, and a code for the "Don't know" response alternative.

e. Codebooks are constructed to define, clarify, and amend codes used during the coding process. Codes that have caused difficulty for the coders should be noted, such as classification systems and codes for open-ended items. Coders require training on specifics of the classification system and codes used for the study, and for the general principles of coding.

f. Since it does not seem to matter if items are scrambled or in blocks according to content, blocking may be preferred due to greater hand scoring ease.

g. Telephone surveys now use Computer Assisted Telephone Interviewing (CATI). These systems are still in experimental stages, and they require extensive programming. Items are read off the CRT screen, and telephone interviewers type respondent answers into a terminal for direct data entry.

Other Considerations

a. There may be a justification for scoring rating scale items dichotomously according to the direction of response. It is sometimes done when bipolar scales are analyzed in terms of the proportion of responses in either direction of the basic dichotomy. The justification is based upon results that seem to indicate that composite scores reflect primarily the direction of responses and only to a minor extent their intensities.

b. One investigator found that many Likert-type rating scales consisting of 2 through 19 steps may be collapsed into two or three measurement categories for analysis with no loss of precision.

c. When working with paired comparison items with a "No preference" option, the "No preference" responses can often be either divided proportionate to the preference responses, or disregarded altogether. The basis for this

suggestion is that respondents who claim neutrality appear to exhibit the same preference patterns as those who express a preference.

d. By using any one of several methods of scoring or transforming self-rating scale raw scores, it is usually possible to approximate dichotomous forced choice results with considerable saving in administration time, and a small gain in test-retest reliability.

e. Investigators sometimes use intensity scores as well as rating scale content scores. One way of obtaining an intensity score is to follow each question with the query, "How strongly do you feel about this?" A second way involves weighting extreme responses (positive and negative) as 2, moderate responses as 1, and neutral responses as 0. These weights can then be summed for an intensity score.

Data Analyses

A detailed discussion of data analyses is beyond the scope of this chapter; however, the following points are noted:

1. Analyses of questionnaire responses are chiefly of two types: summary tabulations and statistical analyses. Tabulations are used primarily for the presentation of results. Statistical tests are used to determine whether the differences in the results are significant. Statistical literature is available which presents numerous tests usable in such analyses.

2. As part of the questionnaire development process, tentative (dummy) analysis tables should be developed to assure that the data to be obtained are appropriate.

3. Weights can be assigned to questionnaires when there is a probability that the selection of respondents is not representative of the population as a whole. For example, a sample distribution drawn from a list of service personnel receiving training, and enrolled in various courses, may result in unequal probability sampling. Since the subjects may be enrolled in more than one course, the more courses they take, the greater the chance they will be selected into the sample.

Weights are also used in making adjustments for total nonresponse and in poststratification. They are able to assign greater importance to some sampled elements than to others in the data analysis. Poststratification conforms the sample distribution to the known population distribution. The sample distribution is adjusted across the strata. This is useful when the population is known, but the stratified sample elements cannot be determined at the selection stage. In such situations, prior stratification is not employable, although poststratification may be applied later. When a sample is weighted to a known population, it will adjust for the sampling fluctuations, as well as for nonresponse. For example, if nonresponse is higher for a specific age group, the sample will conform to the known age distribution when weighted. The development of weights is a difficult task. Standard computer programs for weighted data can be applied in data analysis.

4. Four kinds of measurement scales have been identified: nominal, ordinal, interval, and ratio. Appropriate statistical analyses are associated with each. Hence, the data analysis limitations of various forms of questionnaires should be considered before an instrument is designed. For example, less can be done statistically with open-ended questions than with ranking questions.

Interview Considerations

If properly used, the interview is an effective means of obtaining data. It is a technique in which an individual is questioned by a skilled and trained interviewer who records all replies, preferably verbatim in most cases. Most of the principles of questionnaire construction discussed in previous chapters apply to the interview as well. This section, however, notes some issues specifically related to interviews.

Some of these issues are: the distinction between structured and unstructured interviews, interviewer's characteristics relative to the interviewee, situational factors, training interviewers, data recording and reduction, and special problems. There is, unfortunately, little that can be recommended to avoid some of the problems noted. The questionnaire developer should, in any case, be aware of them.

Structured and Unstructured Interviews

The term "structured" when applied to interviews is intended to emphasize that the interviewer employs a script of all the questions to be asked. In the unstructured interview, the interviewers may know many of the topics to be covered but they need to learn more about the subject overall, so they are willing to be led by the interviewee even into digressions. Unstructured interviews may occur as a preliminary to preparing either a questionnaire or a structured interview script. One could use a questionnaire as the script for a structured interview if one already had the questionnaire developed, but not enough time to convert it to a more convenient format. The main difference between the structured interview and questionnaire is procedural.

The degree of proficiency required of interviewers in conducting an unstructured interview is generally not available during Army field test evaluations. A structured interview requires the interviewer to have only moderate skill and proficiency, and hence is usually preferred. The advantages of the structured interview include: the opportunity to probe for all the facts when the respondent gives only a partial or incomplete response; a chance to ensure that the question is thoroughly understood by the respondent; and an opportunity to pursue other problem areas which may arise during an interview. The structured interview is almost always preferable to a questionnaire when the test group is small (10 to 20), and when time and test conditions permit.

Interviewer's Characteristics Relative to Interviewee

More research is needed to identify how characteristics of an interviewer affect the respondent. Some areas of concern are presented below.

1. Rank, Grade or Status of the Interviewer

It is recommended that the interviewer should be of similar rank or grade to the individuals being interviewed. A difference in rank or grade introduces a bias in the data that has been found to substantially influence test results. Interviewees tend to give the answer they perceive the higher-ranking interviewer favors. When the interviewer is of lower grade, the interviewee may not show respect and may not cooperate.

Evidence indicates that the greater the disparity between the status of the interviewer and that of the respondent, the greater the tendency for biased responses. Respondents tend to provide answers that will be more favorably received by the interviewer.

Data suggest that in the interview situation the respondent tends to support the norms adhered to by the interviewer. Lower socioeconomic respondents may defer to the norms represented by a higher-status interviewer. The effect, however, is related to the types of questions asked. Sensitive issues involving socially accepted or rejected answers will effect more bias.

2. Sex of the Interviewer

Differences in response patterns according to the interviewer's sex depend on subject matter as well as on the composition of the respondent populations and other characteristics of the specific survey situation. Subject matter that tends to be most sensitive to differences in male/female response patterns deals with gender stereotypes. Interview items used in performance appraisals may be sensitive to sex role stereotypes. It is recommended that this type of item be investigated for rating differences between males and females. Interview items that are relevant to technical background experience (not usually obtained by females) also show gender response differences.

3. Race of the Interviewer

The effects of the race of the interviewer on the respondent should probably be viewed as the result of interaction between interviewer and respondent characteristics, or the result of the item content. Respondents often give socially desirable answers to interviewers whose race differs from theirs, particularly if the interviewee's social status is lower than that of the interviewer and the topic of the question is threatening.

Nonsensitive, nonracial items appear to be relatively immune to interviewer effects for racial background. Therefore, racial background of the interviewer does not usually seem to affect survey results. It would be possible to assign interviewers of different racial background regardless of the respondent's racial background. An interviewer's race can probably establish different frames of reference for items with racially related content. For threatening items or items with racially related content, more valid results might be expected when the interviewer is of the same race as the respondent.

4. Experience of the Interviewer

There may be no significant differences between interview completion rates for experienced and inexperienced interviewers who have received sufficient interviewer training for face-to-face interviews and telephone interviews. However, it has been found that experienced interviewers may have different error rates than inexperienced interviewers. This error rate has been associated with the age of the interviewer, and the amount of interviewer training. Interviewer error is usually controlled through selection and training. Older interviewers (age 55 and over) have been known to frequently deviate from interviewer guides. Younger interviewers were found to follow the interview guides more closely. Nonstandardized administration of the interview could jeopardize the overall standardization of the survey procedures.

Other evidence indicates that field interviewers trained for less than a day produce more survey errors than more highly trained interviewers. Individuals responsible for developing interview items, guides, and training require sufficient development time prior to administration of the interview. Interview techniques to increase standardization have been known to improve through training. Response rates for telephone interviews may also be increased through training.

Situational Factors

Among the situational factors that should be considered when interviews are used are the following:

1. It helps greatly if the interviewees perceive the interviewer as interested in hearing their comments, as willing to listen, and (if the situation requires) as willing to protect them from recrimination for being adverse in their evaluations.

2. Interviews should be conducted in a quiet, temperature-controlled environment where the respondent can be comfortable and relaxed. Each respondent should be interviewed in private, separate and apart from all others, so that no other person hears or is biased by his/her responses.

3. The reinforcing behaviors of the interviewer have an influence on the responses collected, and at times may cause respondents to change their preferences. Such comments as "good" or "fine" and such actions as smiling and nodding can have a decided effect on test results. Praised respondents normally offer more answers than unpraised ones. Praising respondents may also tend to reduce "Don't know" answers without increasing insincere or dishonest responses.

4. Interested respondents seem to be more subject to interviewer effects than uninterested ones.

5. Interview questions that are read slowly indicate to respondents that they can take their time in carefully and thoughtfully answering the question. Rushing through an interview may reduce accuracy.

6. Use a "focus" group or pilot screening as a way to develop hypotheses and refine questions for establishing an interview guide and interview items. Interview guides are to be followed so questions are asked without any wording changes. This promotes standardization across interviews.

7. Incomplete answers to survey questions require nondirective probing. When asking for clarification regarding an incomplete answer, the respondent is not to be directed toward any one response. Instead, phrases such as "tell me more" would be useful to employ.

8. Recording answers to interviews that use closed-end questions requires only that the interviewer mark the answer that the respondent selects.

9. When recording answers to open-ended questions, use a tape recorder if the respondent agrees or write down the answers verbatim. It is possible to combine open-ended and closed-end items for interview questionnaires, although coding and recording may be more difficult for the open-ended items.

10. For telephone surveys, use an interview structure and interview guide that promotes a high interaction between the interviewer and the respondent. This may be useful in increasing response rate.

11. Response cards can be adapted from face-to-face interviews for telephone surveys. Oral labeling of the scale points should be assessed on a pilot survey to be sure that the responses are not biased by the oral presentation of the scale.

Training Interviewers

Generally, interviewers require a certain amount of training. Army personnel may check with the Army Research Institute-Field Unit closest to them for help in this area. Some of the factors which should be considered when training interviewers are the following:

1. Training sessions for interviewers usually range between two days and five days. Interviewers conducting field interviews require more training than individuals who conduct telephone interviews. Two days minimum up through five days training are recommended for face-to-face interviews.

2. Sometimes researchers provide interviewers with information about the general research goals, sampling procedures, data analysis, and reports that will result from the survey.

3. Interviewer training requires general information in the course content such as how to introduce the study, as well as more specific information. Interviewers need to be familiar with the wording used in the survey, and any branching instructions. Standardization of the study through asking questions, probing incomplete answers, and recording answers are important aspects of the course content.

4. Interviewer training usually incorporates a demonstration of the standardized interview, and exercises where trainees role-play both the respondent and the interviewer. Practice sessions may also be tape-recorded.

Data Recording and Reduction

In the structured interview, both questions and answers are orally communicated. The interviewer may encode the answers on paper, or tape record the responses for later encoding (but only if the interviewee agrees to the taping and does not seem influenced by the presence of a recording device).

Other topics related to interview data recording and reduction are outside the scope of this chapter.

Special Interviewer Problem

When interviews are used, the qualified interviewer will avoid leading, pressuring, or influencing the direction of an interviewee's evaluations. If potential interviewers have strong preferences regarding the system(s) being tested, they should probably be disqualified.

Many studies have been conducted that show other biasing effects on the interviewer. Factors leading to significant effects of the interviewer upon results include: relatively high ambiguity in the wording of the inquiry; interviewer "resistance" to a given question; and resistance to additional questioning or probing. Interviewer bias can exist without being apparent, and the direction of bias is not necessarily uniform. The least interviewer bias is probably found with questions that can be answered "Yes" or "No." The bias can result from differences in interviewing methods, differences in the degree of success in eliciting factual information, and differences in classifying the respondent's answers. Interviewers' expectations may have a more powerful effect on the results than their ideological preferences.

Some interviewers have a tendency not to transmit printed instructions word for word Hence, total phrases may be eliminated and key words originally intended to focus the respondent's attention on some specific point are omitted or changed. Key ideas are lost, mainly through omission. Variability of interviewer performance seems to vary both across interviewers and within individuals.

An interviewer's attitude toward a question can communicate itself sufficiently to the respondent so that the meaning of the question is altered. When training interviewers to deliver a questionnaire in a standardized fashion, they need to rehearse the questions for tone of voice and body language to reduce any interviewer bias.

This page has been deliberately left blank

Page intentionnellement blanche

Chapter 6

Engineering Measurement of Hardware Characteristics

RSG.24's Terms of Reference stated that we would address data collected to support evaluation both of new systems being acquired and of improvements to existing systems. This includes data in hardware measurement characteristics, as size, weight, light levels, noise levels, crew workspace layout, ingress and egress provisions, temperature, vibration, the brightness, legibility and labeling of displays, and the placement, configuration and force requirements of controls.

RSG.24's Program of Work required that we identify these hardware characteristics and the procedures for measuring and describing them.

The environmental and physical attributes of a system may have positive or negative influence on human performance. Therefore, it is important to measure physical characteristics of the system such as size, weight, light levels, noise levels, crew workspace layout, ingress and egress provisions, temperature, vibration, the brightness, legibility and labeling of displays, and the placement, configuration and force requirements of controls. These measurements are then compared to design requirements or best engineering practices to determine if they are within acceptable ranges.

This document is to be used by a wide variety of people, who may be human factors people and others whose primary expertise is not the human factors domain.

We must provide all these people understandable data and means which must be consistent and pragmatic, in order for them to conduct tests and evaluations in the best conditions, during the materiel acquisition process. They shall be also able to find in this document enough precise references, if needed, to help them to reach their initial goal.

Various kinds of data are available:

- data from direct observation, which may be also subjective, but don't need any specific device,
- data from indirect observation, using, for example, video recording devices, that may induce important post treatment,
- measured data, with data acquisition and treatment devices, for physical parameters for example.

To achieve the testing goal, different procedures and investigation techniques may be used. For example, we can notice:

- testing procedures, well defined or standardized,
- evaluation grids,
- questionnaires,
- rating scales,
- semi directive interviews,
- assessment interviews,
- measuring devices like various sensors,
- etc.

Frequently, a detailed and complete job analysis, before the test and evaluation period, will be necessary to help the analyst in determining the most critical points to be observed, in order to:

- choose what is to be evaluated or tested,
- define which means he will have to use to perform this analysis.

It seems to be unrealistic to commit measurements without any well defined goal.

Some work has been done in the past, more specifically for wheeled, then tracked vehicle by a NATO group (issued from the Western Europe Union). A lot of testing procedures have been gathered in a document called "Allied Vehicle Testing Publications", which contains a specific chapter about "Ergonomics", including subchapters related to hardware engineering measurements and other areas.

These "AVTP" are very well suited for measurement of hardware characteristics for all kind of vehicles, but also for many kinds of workstations anywhere else. However, for some specific activities, like those related to the individual in the battlefield (e.g. antitank weapons), some items may lack. Studies on this particular point are conducted in different countries at the present time (France, Great Britain and United States).

We propose to define the way to do the measurements of hardware characteristics as following. Three major items are to be taken into account:

(1) Dimensions,
(2) Environmental constraints,
(3) Human-machine interfaces.

These items are to be considered through the three same questions each time, which are:

(1) WHY do we measure?
(2) WHAT is to be measured?
(3) HOW do we measure?

The third point might include specification of the kind of devices to be used, and all related characteristics they must verify.

The answers to these three questions provide data which then may be evaluated against the requirements of military and other standards or system specifications. At this step, human factors people can provide the program managers and decision makers with comprehensive and consistent data on the consequences in terms of system efficiency and/or performance related to various choices (technical, organization, costs, etc.).

MEASUREMENT OF HARDWARE CHARACTERISTICS

1. DIMENSIONS

1.1. WHY?

We make the measurements in order to answer evaluation issues and to verify good Human Machine System performance. For example, good posture diminishes fatigue and consequently improves the global performance. It is important to provide as detailed as possible measurements related to that topic, to be able to compare various solutions in a short period of time, and to determine what proportion of the targeted population will be accommodated appropriately.

1.2. WHAT?

What characteristics of the system will be measured? For example: reaching distances (optimal and maximal), viewing distances, etc.

1.3. HOW?

The answer to this question includes the procedures and methodologies for measurements and the tools used for the measurements, such as instrumentation and sensors.

The tools used range from the very simple (e.g. centimeter ruler) to the relatively sohisticated (e.g. digital recording of the data stream from a high speed data bus).

Each nation may provide anthropometric data of their users populations. Those data may be extracted from anthropometric databases, or calculated (i.e. estimated) for future user population during the total life cycle of the system (which may length more than thirty years for some of them). A useful reference is the recently published NATO Soldier Target Audience Description, RTO-TR-22 AC/323(HFM)TP/20, which describes anthropometric, demographic, and physiological characteristics of the NATO Soldier population.

2. ENVIRONMENTAL VARIABLES

Environmental variables are a major concern with system efficiency. They must be measured using generally recognized methods, in order to be able to compare data across nations. If not, it will be difficult to give usable information for system evaluation.

2.1. WHY?

It is well established that the soldier's performance and system performance is affected by different environmental conditions. This fact justifies paying special attention to that kind of measurement.

2.2. WHAT?

Examples of variables to be measured are:

- Noise
    - Vibrations
    - Shock
- Light
- Toxic substances
    - Temperature
    - Humidity
    - Air speed

2.3. HOW?:

The measures have to be made at the users location, and depend on the activity of those users. Most of the variables are measured physically or chemically, and require instrumentation.

3. HUMAN-MACHINE INTERFACE

3.1. WHY? A good human-machine interface supports the operator in doing his or her job, enhancing system performance and crew interaction.

3.2. WHAT?

- Guidance:

* Prompting: it refers to the means available in order to lead the users to making specific actions whether it is data entry or other tasks. This criterion also refers to all the means that help users to know the alternatives when several actions are possible, depending on the contexts. Prompting also

concerns status information, that is information about the actual state or context of the system, as well as information concerning help facilities and their, accessibility.

* Grouping and discrimination of items (by location and by format): this criterion concerns the visual organization of the information items in relation to one another. It takes into account the topology (location) and some graphical characteristics (format) in order to indicate the relationships between the various items displayed, to indicate whether or not they belong to a given class, or else to indicate differences between classes. This criterion also concerns the organization of items within a class.

* Immediate feedback: it concerns system responses to users' actions. These actions may be simple keyed entries or more complex transactions such as stacked commands. In all cases computers responses must be provided, and they should be fast, with appropriate and consistent timing for different types of transactions. In all cases, a fast response from the computer should be provided with information on the requested transaction and result.

* Legibility: It concerns the lexical characteristics of the information presented on the screen that may hamper or facilitate the reading of this information (character brightness, contrast between letter and background, font size, interword spacing, line spacing, paragraphs spacing, line length, etc.). By definition, the criterion Legibility does not concerns feedback or error messages.

- User workload: The criterion Workload concerns all interface elements that play a role in the reduction of the users' perceptual or cognitive load, and in the increase of the dialogue efficiency.

* Brevity: The criterion Brevity concerns the perceptual and cognitive workload both for individual inputs and outputs, and for sets of inputs (i.e., sets of actions needed to accomplish a goal or a task). Brevity corresponds to the goal of limiting the reading and input workload and the number of actions steps.

+ Conciseness: This criterion concerns workload from a perceptual and cognitive workload for individual inputs or outputs. By definition, Concision does not concern feedback or error messages.

+ Minimal actions: This criterion concerns workload with respect to the number of actions necessary to accomplish a goal or a task. It is here a matter of limiting as much as possible the steps users must go through.

* Information density: This criterion concerns the users' workload from a perceptual and cognitive point of view with regard to the whole set of information presented to the users rather than each individual element or item.

- Direct user control: This criterion refers to the relationship between the computer processing and the actions of the users. This relationship must be explicit, i.e. the computer must process only those actions requested by the users and only when requested to do so.

* User situational awareness: The user must have sufficient knowledge of system actions and status to exercise effectively his or her direct control. The system should provide appropriate information and options.

- Adaptability: The adaptability of a system refers to its capacity to behave contextually and according to the users' needs and preferences.

* Flexibility: This criterion refers to the means available to the users to customize the interface in order to take into account their working strategies and/or their habits, and the task requirements. Flexibility is reflected in the number of possible ways of achieving a given goal. In other words; it is the capacity of the interface to adapt to the users' particular needs.

* Users' experience management: This criterion refers to the means available to take into account the level of user experience.

- Error management: This criterion refers to the means available to prevent or reduce errors and to recover from them when they occur. Errors are defined in this context as invalid data entry, invalid format for data entry, incorrect command syntax, etc.

* Error protection: This criterion refers to the means available to detect and prevent data entry errors, or actions with destructive consequences.

* Quality of error messages: This criterion refers to the adaptability, and specificity about the nature of errors syntax, format, etc. and the actions needed to correct them.

* Error correction: This criterion refers to the means available to the users to correct their errors.

- Consistency: This criterion refers to the way interface design choices (codes, naming, formats, procedures, etc.) are maintained in similar contexts, and different when applied to different contexts.

- Significance of codes: This criterion qualifies the relationship between a term and/or a sign and its reference. Codes and names are significant to the users when there is a strong semantic relationship between such codes and the items or actions they refer to.

- Compatibility: This criterion refers to the match between users' characteristics (memory, perceptions, customs, skills, age, expectations, etc.) and task characteristics on the one hand, and the organization of the output, input, and dialogue for a given application, on the other hand. It also concerns the coherence between environments and between applications.

3.3. HOW?

Hardware measurement alone is normally not a sufficient basis for evaluation of the human-machine interface of a system. Testing normally will include (1) verification of compliance with existing standards and with good engineering design practice, (2) measurement of performance of tasks believed to be affected by human-machine interface design, and (3) collection of information from users and subject matter experts. Item one is engineering measurement; the other two items are treated in other chapters of this document.

As mentioned earlier, the AVTP is a good source of guidance on testing procedures primarily for vehicles, but useful for some other systems as well, and is well known within NATO. The tester should consider thoughtfully which techniques, procedures and methods are most appropriate to meet each unique testing need.

This page has been deliberately left blank

Page intentionnellement blanche

Chapter 7

Conclusions and Recommendations

Human engineering testing and evaluation are conducted to ensure that the intended users of a system can operate and maintain it. This type of testing determines if the system's equipment meets human engineering, safety and other criteria relevant to human use, while meeting all mission performance requirements. Analyses to be carried out must verify that effective, efficient and safe operation of the total human-machine system is supported by the overall architecture, workspace design and workload imposed by the system's design.

Human engineering testing and evaluation is an important component of any test effort as it is generally the only activity that looks at the influence of human performance on system performance. The need for ensuring accurate human performance is obvious when operators are in the control loop of a system as they directly affect performance of the system. However, even in the most automated system, humans are often assigned critical functions such as enabling, programming, initializing, calibrating, verifying, validating, designating, and authorizing. Indeed, accurate human performance becomes more critical with increased automation because of the absence of a human operator. For example, once a cruise missile is in flight, the operator is out of the missile's control loop and cannot compensate for any errors that may have escaped detection when the crew loaded the missile onto a launch platform, programmed it with a flight plan and launched it.

The purpose of this report is to document the efforts of RSG-24 and to present its recommended guidelines for accomplishing human engineering test and evaluation. The goal was standardization of test content, procedures and conditions /sequence of test events. The intent is not to impose standardization of system engineering design. The guidelines are expected to facilitate the sharing of data and evaluations which will cut test costs by reducing duplication and the quantity of test data required to support decisions.

The intended audience for these guidelines includes all persons who conduct human engineering testing on military systems. The professional tester benefits from using the guidelines because he will be able to compare his test results to those gathered on other systems, under other conditions, or by other nations. The novice tester gains the additional benefit of having a readily available reference to proven techniques.

This document offers widely accepted techniques and methods for accomplishing human engineering tests. The approach taken to accomplishing human engineering testing generally depends on the organization's tradition and upon the individual specialist's education, experience, and intuition. In the absence of a common approach and data set, human engineering test results are difficult to compare for different test conditions, systems, armed services, or nations. As a result, duplicative testing becomes necessary, and important lessons learned and generality of findings are lost.

Some readers may note the absence of a treatment of the human-software interface. EG.K discussed attempting to include that in the present document, but at the time the planning for this effort was under way, no common or widely used techniques or methods were recognized. It was decided to leave that task to a subsequent group to address when such techniques and methods were known. Both EG.K and RSG-24 recognized that they were dealing with a process in motion and that commonly accepted and used techniques and methods would continue to emerge, and should be added to the toolkit. Revision and updating of any document of this kind is necessary if it is to continue to be useful.

It is therefore recommended both that (1) the techniques and methods described in this document should be used to the maximum extent possible in NATO testing and evaluation, because they will work to mitigate those costs and losses, and (2) that they be revised and updated periodically to include newly recognized common tools in the recommended list, and to benefit from advances in technology.

This page has been deliberately left blank

Page intentionnellement blanche

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's References	3. Further Reference	4. Security Classification of Document
	RTO-TR-021 AC/323(HFM-018)TP/19	ISBN 92-837-1068-1	UNCLASSIFIED/ UNLIMITED

5. Originator Research and Technology Organization
North Atlantic Treaty Organization
BP 25, 7 rue Ancelle, F-92201 Neuilly-sur-Seine Cedex, France

6. Title
NATO Guidelines on Human Engineering Testing and Evaluation

7. Presented at/sponsored by
the Human Factors and Medicine Panel (HFM) Research and Study Group 24.

8. Author(s)/Editor(s)	9. Date
Multiple	May 2001
10. Author's/Editor's Address	**11. Pages**
Multiple	108

12. Distribution Statement There are no restrictions on the distribution of this document. Information about the availability of this and other RTO unclassified publications is given on the back cover.

13. Keywords/Descriptors

Human factors engineering
Guidelines
NATO
Man computer interface
Tests
Evaluation
Task analysis
Workloads
Performance evaluation
Measurement
Surveys
Systems engineering
Humans

14. Abstract

Testing and Evaluation I (T&E) is an integral part of the system development process. Human Engineering T&E addresses the quality and effectiveness of the interface between the humans who participate as part of a human-machine system and the hardware and other non-human components. In the interest of supporting among NATO nations the co-development, co-production of systems, and shared use of T&E resources as a means of sharing more effective and less expensive systems, this document describes techniques and methods that are recommended for common use in NATO.

The measurement categories addressed are:

1. Description of test participants
2. Measurement of operator workload
3. Human task performance measurement
4. User opinion
5. Engineering measurement of hardware characteristics

It is recommended that the techniques and methods described be used to the maximum extent possible, and that the list be revised periodically to support currency and continued usefulness.

This page has been deliberately left blank

Page intentionnellement blanche

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION
BP 25 • 7 RUE ANCELLE
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

DIFFUSION DES PUBLICATIONS
RTO NON CLASSIFIEES

L'Organisation pour la recherche et la technologie de l'OTAN (RTO), détient un stock limité de certaines de ses publications récentes, ainsi que de celles de l'ancien AGARD (Groupe consultatif pour la recherche et les réalisations aérospatiales de l'OTAN). Celles-ci pourront éventuellement être obtenues sous forme de copie papier. Pour de plus amples renseignements concernant l'achat de ces ouvrages, adressez-vous par lettre ou par télécopie à l'adresse indiquée ci-dessus. Veuillez ne pas téléphoner.

Des exemplaires supplémentaires peuvent parfois être obtenus auprès des centres nationaux de distribution indiqués ci-dessous. Si vous souhaitez recevoir toutes les publications de la RTO, ou simplement celles qui concernent certains Panels, vous pouvez demander d'être inclus sur la liste d'envoi de l'un de ces centres.

Les publications de la RTO et de l'AGARD sont en vente auprès des agences de vente indiquées ci-dessous, sous forme de photocopie ou de microfiche. Certains originaux peuvent également être obtenus auprès de CASI.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE
Streitkräfteamt / Abteilung III
Fachinformationszentrum der Bundeswehr, (FIZBw)
Friedrich-Ebert-Allee 34
D-53113 Bonn

BELGIQUE
Coordinateur RTO - VSL/RTO
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evère, B-1140 Bruxelles

CANADA
Services d'information scientifique pour la défense (SISD)
R et D pour la défense Canada
Ministère de la Défense nationale
Ottawa, Ontario K1A 0K2

DANEMARK
Danish Defence Research Establishment
Ryvangs Allé 1, P.O. Box 2715
DK-2100 Copenhagen Ø

ESPAGNE
INTA (RTO/AGARD Publications)
Carretera de Torrejón a Ajalvir, Pk.4
28850 Torrejón de Ardoz - Madrid

ETATS-UNIS
NASA Center for AeroSpace Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320

FRANCE
O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc
BP 72, 92322 Châtillon Cedex

GRECE (Correspondant)
Hellenic Ministry of National Defence
Defence Industry Research & Technology General Directorate
Technological R&D Directorate
D.Soutsou 40, GR-11521, Athens

HONGRIE
Department for Scientific Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ISLANDE
Director of Aviation
c/o Flugrad
Reykjavik

ITALIE
Centro di Documentazione Tecnico-Scientifica della Difesa
Via XX Settembre 123a
00187 Roma

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

PAYS-BAS
NDRCC
DGM/DWOO
P.O. Box 20701
2500 ES Den Haag

POLOGNE
Chief of International Cooperation Division
Research & Development Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL
Estado Maior da Força Aérea
SDFA - Centro de Documentação
Alfragide
P-2720 Amadora

REPUBLIQUE TCHEQUE
Distribuční a informační středisko R&T
VTÚL a PVO Praha
Mladoboleslavská ul.
197 06 Praha 9-Kbely AFB

ROYAUME-UNI
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

TURQUIE
Millî Savunma Başkanliği (MSB)
ARGE Dairesi Başkanliği (MSB)
06650 Bakanliklar - Ankara

AGENCES DE VENTE

NASA Center for AeroSpace Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320
Etats-Unis

The British Library Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
Royaume-Uni

Canada Institute for Scientific and Technical Information (CISTI)
National Research Council
Document Delivery
Montreal Road, Building M-55
Ottawa K1A 0S2, Canada

Les demandes de documents RTO ou AGARD doivent comporter la dénomination "RTO" ou "AGARD" selon le cas, suivie du numéro de série (par exemple AGARD-AG-315). Des informations analogues, telles que le titre et la date de publication sont souhaitables. Des références bibliographiques complètes ainsi que des résumés des publications RTO et AGARD figurent dans les journaux suivants:

Scientific and Technical Aerospace Reports (STAR)
STAR peut être consulté en ligne au localisateur de ressources uniformes (URL) suivant:
http://www.sti.nasa.gov/Pubs/star/Star.html
STAR est édité par CASI dans le cadre du programme NASA d'information scientifique et technique (STI)
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
Etats-Unis

Government Reports Announcements & Index (GRA&I)
publié par le National Technical Information Service
Springfield
Virginia 2216
Etats-Unis
(accessible également en mode interactif dans la base de données bibliographiques en ligne du NTIS, et sur CD-ROM)

Imprimé par St-Joseph Ottawa/Hull
(Membre de la Corporation St-Joseph)
45, boul. Sacré-Cœur, Hull (Québec), Canada J8X 1C6

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION
BP 25 • 7 RUE ANCELLE
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Telefax 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

DISTRIBUTION OF UNCLASSIFIED RTO PUBLICATIONS

NATO's Research and Technology Organization (RTO) holds limited quantities of some of its recent publications and those of the former AGARD (Advisory Group for Aerospace Research & Development of NATO), and these may be available for purchase in hard copy form. For more information, write or send a telefax to the address given above. **Please do not telephone**.

Further copies are sometimes available from the National Distribution Centres listed below. If you wish to receive all RTO publications, or just those relating to one or more specific RTO Panels, they may be willing to include you (or your organisation) in their distribution.

RTO and AGARD publications may be purchased from the Sales Agencies listed below, in photocopy or microfiche form. Original copies of some publications may be available from CASI.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Coordinateur RTO - VSL/RTO
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evère, B-1140 Bruxelles

CANADA
Defence Scientific Information Services (DSIS)
Defence R&D Canada
Department of National Defence
Ottawa, Ontario K1A 0K2

CZECH REPUBLIC
Distribuční a informační středisko R&T
VTÚL a PVO Praha
Mladoboleslavská ul.
197 06 Praha 9-Kbely AFB

DENMARK
Danish Defence Research Establishment
Ryvangs Allé 1, P.O. Box 2715
DK-2100 Copenhagen Ø

FRANCE
O.N.E.R.A. (ISP)
29 Avenue de la Division Leclerc
BP 72, 92322 Châtillon Cedex

GERMANY
Streitkräfteamt / Abteilung III
Fachinformationszentrum der Bundeswehr, (FIZBw)
Friedrich-Ebert-Allee 34
D-53113 Bonn

GREECE (Point of Contact)
Hellenic Ministry of National Defence
Defence Industry Research & Technology General Directorate
Technological R&D Directorate
D.Soutsou 40, GR-11521, Athens

HUNGARY
Department for Scientific Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ICELAND
Director of Aviation
c/o Flugrad
Reykjavik

ITALY
Centro di Documentazione Tecnico-Scientifica della Difesa
Via XX Settembre 123a
00187 Roma

LUXEMBOURG
See Belgium

NETHERLANDS
NDRCC
DGM/DWOO
P.O. Box 20701
2500 ES Den Haag

NORWAY
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

POLAND
Chief of International Cooperation Division
Research & Development Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL
Estado Maior da Força Aérea
SDFA - Centro de Documentação
Alfragide
P-2720 Amadora

SPAIN
INTA (RTO/AGARD Publications)
Carretera de Torrejón a Ajalvir, Pk.4
28850 Torrejón de Ardoz - Madrid

TURKEY
Millî Savunma Başkanliği (MSB)
ARGE Dairesi Başkanliği (MSB)
06650 Bakanliklar - Ankara

UNITED KINGDOM
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

UNITED STATES
NASA Center for AeroSpace Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320

SALES AGENCIES

NASA Center for AeroSpace Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320
United States

The British Library Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
United Kingdom

Canada Institute for Scientific and Technical Information (CISTI)
National Research Council
Document Delivery
Montreal Road, Building M-55
Ottawa K1A 0S2, Canada

Requests for RTO or AGARD documents should include the word 'RTO' or 'AGARD', as appropriate, followed by the serial number (for example AGARD-AG-315). Collateral information such as title and publication date is desirable. Full bibliographical references and abstracts of RTO and AGARD publications are given in the following journals:

Scientific and Technical Aerospace Reports (STAR)
STAR is available on-line at the following uniform resource locator:
http://www.sti.nasa.gov/Pubs/star/Star.html
STAR is published by CASI for the NASA Scientific and Technical Information (STI) Program
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
United States

Government Reports Announcements & Index (GRA&I)
published by the National Technical Information Service
Springfield
Virginia 22161
United States
(also available online in the NTIS Bibliographic Database or on CD-ROM)

Printed by St. Joseph Ottawa/Hull
(A St. Joseph Corporation Company)
45 Sacré-Cœur Blvd., Hull (Québec), Canada J8X 1C6

ISBN 92-837-1068-1