NORTH ATLANTIC TREATY
ORGANISATION

RESEARCH AND TECHNOLOGY
ORGANISATION

AC/323(HFM-121)TP/62

www.rto.nato.int

RTO TECHNICAL REPORT (PART II)

TR-HFM-121-Part-II

Virtual Environments for Intuitive Human-System Interaction

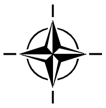
(Environnements virtuels d'interaction
Homme-Système Intuitive)

Human Factors Considerations in the Design,
Use, and Evaluation of AMVE-Technology

Final Report of Task Group TR-HFM-121.

Published July 2007

Distribution and Availability on Back Cover



NORTH ATLANTIC TREATY
ORGANISATION

RESEARCH AND TECHNOLOGY
ORGANISATION

AC/323(HFM-121)TP/62

www.rto.nato.int

RTO TECHNICAL REPORT (PART II)

TR-HFM-121-Part-II

Virtual Environments for Intuitive Human-System Interaction

(Environnements virtuels d'interaction
Homme-Système Intuitive)

Human Factors Considerations in the Design,
Use, and Evaluation of AMVE-Technology

Final Report of Task Group TR-HFM-121.

The Research and Technology Organisation (RTO) of NATO

RTO is the single focus in NATO for Defence Research and Technology activities. Its mission is to conduct and promote co-operative 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 co-ordination 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 co-ordinates RTO's co-operation 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 co-operation.

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 System Analysis and Studies 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 co-operation 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.

The content of this publication has been reproduced
directly from material supplied by RTO or the authors.

Published July 2007

ISBN 978-92-873-0077-7

Table of Contents

Page

HFM-121 / RTG-042 Members viii

Executive Summary and Synthèse ES-1

Introduction 1-1

Perceptual Issues of Augmented and Virtual Environments 2-1
1.0 Introduction 2-1
2.0 User Characteristics 2-2
3.0 System Characteristics 2-3
4.0 Task Characteristics 2-4
5.0 Perceptual Requirements 2-6
  5.1 Visual Requirements 2-6
  5.2 Special Visual Issues 2-7
    5.2.1 Eye Dominance 2-7
    5.2.2 Pupil Adaptation 2-7
    5.2.3 Visual Field 2-8
    5.2.4 Accommodation 2-8
    5.2.5 Sensitivity to Flicker 2-8
    5.2.6 Vision Deficiencies 2-8
  5.3 Audio Requirements 2-9
    5.3.1 Sound Perception 2-9
    5.3.2 Sound Processing 2-10
  5.4 Haptic Feedback 2-12
  5.5 Olfactory Feedback 2-15
6.0 References 2-15

Presence 3-1
1.0 Introduction 3-1
2.0 Theories, Definitions, and Multi-Dimensionality 3-1
3.0 Effects of Presence 3-3
4.0 Measurement of Presence 3-4
  4.1 Subjective Measurement 3-4
  4.2 Objective Measurement 3-6
5.0 Determinants of Presence 3-6
  5.1 Sensors and Displays 3-7
    5.1.1 Coverage and Richness 3-7
    5.1.2 Device Characteristics 3-7

        5.1.3 Obtrusiveness 3-7
        5.1.4 Consistency 3-8
    5.2 Virtual World 3-8
        5.2.1 Situation and Task 3-8
        5.2.2 Virtual World Realism 3-8
        5.2.3 Transfer of Object Permanence 3-9
        5.2.4 Occurrence and Representation of Human Beings 3-9
    5.3 Action 3-9
        5.3.1 Action Types 3-10
        5.3.2 Number and Range 3-10
        5.3.3 Control of Action 3-11
        5.3.4 Lag/Latency 3-11
        5.3.5 Action Realism 3-11
        5.3.6 Illusionary Action 3-11
    5.4 Presentation 3-12
        5.4.1 Visual Presentation Realism 3-12
        5.4.2 Aural Presentation Realism 3-12
        5.4.3 Use of Conventions 3-13
    5.5 Human Mind 3-13
        5.5.1 Individual Characteristics / Personality Type 3-13
        5.5.2 Willingness to Suspend Disbelief 3-13
        5.5.3 Attention 3-14
        5.5.4 Mood 3-14
        5.5.5 Knowledge, Prior Experience and Expectations 3-14
6.0 Concluding Remarks 3-15
7.0 References 3-15

A Review of the Mental Workload Literature 4-1

1.0 Introduction 4-1
    1.1 Definitions of Workload 4-1
    1.2 Reasons for Measuring Workload 4-3
    1.3 Criteria for Workload Measurement Methods 4-4
2.0 Notable Reviews of Workload Measurement Techniques 4-5
3.0 Workload Measurement Technique Categories 4-6
    3.1 Rating Scales 4-7
    3.2 Performance Measures 4-13
    3.3 Psychophysiological Measures 4-16
        3.3.1 Electroencephalography 4-19
        3.3.2 Eye Movement 4-19
        3.3.3 Heart Rate 4-20
        3.3.4 Respiration 4-20
        3.3.5 Future Psychophysiological Measure Developments 4-21
4.0 Important Remaining Workload Measurement Issues 4-21
5.0 Recommending Workload Measures 4-22
6.0 Conclusion 4-23
7.0 Abbreviations 4-24

8.0 References 4-26
Appendix 1 – Internet Workload Measurement Technique Search: Hits by Keyword 4-32

Simulator Sickness Research Summary 5-1

1.0 Reviews 5-1
2.0 Selected History 5-2
  2.1 Miller and Goodson 5-2
  2.2 McGuinness, Bouwman, and Forbes 5-3
  2.3 McCauley 5-4
  2.4 Crowley 5-5
3.0 Signs and Symptoms 5-5
4.0 Measurement 5-6
  4.1 Simulator Sickness Questionnaire (SSQ) 5-7
  4.2 Instrumented Physiological Measures 5-8
  4.3 Tests of Postural Equilibrium 5-8
5.0 Incidence 5-9
6.0 Residual Aftereffects 5-11
7.0 Adaptation 5-12
8.0 Susceptibility 5-13
  8.1 Gender 5-14
  8.2 Age 5-14
  8.3 Age and Experience 5-15
  8.4 Experience 5-15
  8.5 Prior History of Motion Sickness 5-16
  8.6 Miscellaneous: Illness, Drugs, Sleep, Fatigue 5-17
  8.7 Simulator Variables 5-18
  8.8 Task Variables 5-18
9.0 Simulator Sickness, Performance, and Training 5-18
  9.1 Performance 5-18
  9.2 Training 5-19
10.0 Treatment 5-21
11.0 Theory 5-21
  11.1 Sensory Conflict Theory 5-21
  11.2 Postural Instability Theory 5-22
  11.3 SS, Age, and Theory 5-22
12.0 Guidelines for Reducing Simulator Sickness and Risks from Aftereffects 5-23
  12.1 General Rules 5-23
13.0 Suggestions for Future Research 5-24
14.0 References 5-24

Dynamic Situations: The Soldier's Situation Awareness 6-1

1.0 Situation Awareness 6-1
2.0 SA and Decision Making 6-1
  2.1 Uniform Battlefield Decision Making 6-2

3.0 Experience in SA 6-3
4.0 Digital Systems to Enhance SA 6-3
5.0 References 6-4

Team and Collective Performance Measurement 7-1
1.0 Introduction 7-1
2.0 Team Performance 7-1
  2.1 Example of Measure of Team Outcomes – UPAS 7-2
  2.2 Measures of Team SA and Shared Mental Models 7-2
  2.3 Example Measure of Team SA – SALIENT 7-2
  2.4 The Role of Mental Models in Team Effectiveness 7-3
3.0 Communication Analysis 7-4
  3.1 An Example of Communication Analysis – The IOC Model 7-5
4.0 Distributed VR systems – Event Based Training 7-6
  4.1 The TARGET Checklist 7-6
  4.2 The Teamwork Observation Measure 7-6
5.0 Collective Performance Assessment 7-7
  5.1 Implications for Collective Training Assessment Techniques 7-8
  5.2 Collective Performance Assessment and Mission Phases 7-8
6.0 Objective Measures 7-10
7.0 Individual Performance 7-11
  7.1 FOI Approach to Performance Measurement 7-11
  7.2 The FOI Pilot Performance Scale 7-12
  7.3 Objective Measures of Individual Performance 7-12
8.0 Summary 7-13
9.0 References 7-14

After Action Review in Simulation-Based Training 8-1
1.0 Simulation-Based Training Environments 8-1
2.0 Intrinsic Feedback 8-2
3.0 Extrinsic Feedback 8-2
4.0 Extrinsic Feedback Methods for Collective Training 8-3
5.0 How Realistic Battlefield Simulations Set the Stage for an AAR 8-4
6.0 The Roles of AAR Aids in Supporting Feedback 8-4
7.0 A Recently Developed AAR Tool for Virtual Environments 8-5
  7.1 Evaluation and Utilization 8-8
8.0 Current and Future Research Needs 8-9
9.0 Summary 8-10
10.0 References 8-10

Technical Evaluation Report 9-1
1.0 Introduction 9-1
  1.1 Background 9-1
  1.2 Purpose and Scope of this Workshop 9-2

1.3 Program Workshop 9-2
1.4 Attendees 9-5
2.0 Technical/Scientific Situation of Military VR Applications 9-5
2.1 Introduction 9-5
2.1.1 Technological Developments 9-5
2.1.2 Scientific Disciplines Involved 9-7
2.1.3 Evolved Application Fields 9-7
2.1.4 Requirements Perspective 9-10
2.1.5 Application Development Approach 9-11
2.2 Bottlenecks and Opportunities 9-11
2.3 Recommended Actions 9-12
3.0 Conclusions 9-13
3.1 Future Meetings 9-14

Conclusions 10-1

1.0 Future Consideration 10-3

HFM-121 / RTG-042 Members

CHAIRMAN

Dr. Thomas Alexander
Research Institute for Communication,
Information Processing, and Ergonomics (FKIE)
Neuenahrer Strasse 20
53343 Wachtberg
GERMANY
Phone: +49 228 9435 480
Fax: +49 228 9435 508
Email: alexander@fgan.de

CO-CHAIRMAN

Dr. Stephen Goldberg
US Army Research Institute
12350 Research Parkway
Orlando, FL 32826
UNITED STATES
Phone: +1 407 384 3980
Fax: +1 407 384 3999
Email: Stephen.Goldberg@us.army.mil

NATIONAL REPRESENTATIVES

CANADA

Dr. Lochlan Magee
Head, Simulation and Modelling for Acquisition
Rehearsal and Training (SMART) Section
DRDC Toronto
P.O. Box 2000, 1133 Sheppard Avenue West
Toronto, Ontario M3M 3B9
Phone: +1 416-635-2149
Fax: +1 416-635-2184
Email: Lochlan.Magee@drdc-rddc.gc.ca

DENMARK

Ms. Lisbeth M. Rasmussen
Senior Adviser, M.Sc.
Applied Research Section (TDBE)
Danish Defence Acquisition and Logistics
Organization
Lautrupbjerg 1-5
DK-2750 Ballerup
Phone: +45 7257 1552
Fax: +45 7257 5120
Email: lr@mil.dk

FRANCE

MC. Paul Gorzerino
DGA/DCE.ETC4/ETAS
Chef de la Division Facteurs Humains
Etablissement Technique d'Angers
Route de Laval, BP 36
49460 Montreuil-Juigne
Phone: +33 2 4193 6673
Fax: +33 2 4193 6703
Email: paul.gorzerino@dga.defense.gouv.fr

SWEDEN

Mr. Jonathan Borgvall
Researcher
Swedish Defence Research Agency
Division of C2-Systems
Department of Man-System Interaction
P.O. Box 1165
SE-581 11 Linkoping
Phone: +46 13 378 232
Fax: +46 13 124 938
Email: jonathan.borgvall@foi.se

Mr. Patrik Lif
Researcher
Swedish Defence Research Agency
Division of C2-Systems
Department of Man-System Interaction
P.O. Box 1165
SE-581 11 Linkoping
Phone: +46 13 378 179
Fax: +46 13 124 938
Email: patand@foi.se

THE NETHERLANDS

Dr. Nico J. Delleman
Simulation-Based Design Ergonomics
Department of Performance & Comfort
TNO Human Factors
Kampweg 5, P.O. Box 23
3769 ZG Soesterberg
Phone: +31 346 356 347
Fax: +31 346 353 977
Email: Delleman@tm.tno.nl

UNITED KINGDOM

Ms. Heather Mcintyre
DSTL
Rm 126, Bldg 115
Bedford Technology Park
Thurleigh, Bedforshire MK44 2FQ
Phone: +44 1234 716 446
Fax: +44 1234 716 440
Email: hmmcintyre@dstl.gov.uk

Ms. Ebb Smith
DSTL
Policy and Capability Studies
Rm 126, Bldg 115
Bedford Technology Park
Thurleigh, Bedfordshire MK44 2FQ
Phone: +44 (0) 1234 225 037
Fax: +44 (0) 1234 225 041
Email: mesmith@dstl.gov.uk

UNITED STATES

Lt. Joseph Cohn
Lead, Training Requirements
and Operational Evaluation
Navy Center for Applied Research
in Artificial Intelligence
Naval Research Laboratory
Code 5510
4555 Overlook Avenue, SW
Washington, DC 20375-5337
Phone: +1 202-404-8624
Fax: +1 202-767-2166
Email: cohn@itd.nrl.navy.mil

Virtual Environments for Intuitive Human-System Interaction

(RTO-TR-HFM-121-Part-II)

Executive Summary

At the present time the alliance is placing new performance requirements on military personnel which is driving a need for new approaches to training and equipment design. Innovative computer technologies have the potential to prepare militaries for their missions. Preparation for these challenges has to include both training military commanders, staffs and operators by means of appropriate media and developing equipment that has a well thought out integrated human-machine design that reduces personnel requirements, operator workload and reduces training time.

The new computer-generated media of Augmented, Mixed, and Virtual Environments (AMVE) can provide realistic training and natural human-system interaction (HSI) using complex realistic or abstract synthetic environments. AMVE technologies allow trainees and human operators to experience synthetic environment that are appropriate for the tasks to be performed. An extensive review of the national activities in member nations revealed that AMVE technologies have become much more useful for a variety of military application areas than they were a few years ago. VE applications have many demonstrated success stories in military education and training. They show that VE can be useful, and sometimes necessary to achieve the training objectives. But still AMVE cannot be considered an intuitive technology.

This report reviews a number of human use and effectiveness issues as they relate to AMVE. Prior to designing an AMVE application a detailed analysis has to be performed in order to identify the applicability of AMVE-technology to represent complex processes or tasks. This analysis has to address human factors issues on a number of different levels in order to make optimal use of the capabilities of the new technology. For the design of a new AMVE-system it is important to consider relevant human information processing resources and capabilities on a perceptual level. User feelings of presence or immersion also affect perceptions and ultimately the effectiveness of the system. Other factors that affect performance include workload, especially mental workload and simulator sickness. Methods and measures for determining and quantifying simulator sickness caused by AMVE-technology are described and discussed. Attaining situational awareness is another complex construct that has become vitally important for mission success. AMVE technologies can represent the complex nature of the battlefield and can be used to provide the training and experiences needed to allow trainees to rapidly acquire situational awareness. A final issue discussed is performance evaluation, the application of dependent measures and team measures to evaluate success.

Detailed case studies were presented and discussed in a workshop (RWS-136 on Virtual Media for Military Applications) in June, 2006. The main conclusions of the workshop are also included in this report.

Overall the RTG concluded that AMVE is applicable and has become practical in several areas. Regardless there are still many questions that remain unanswered and the topic continues to require further research. Military education and training have been identified as one of the main application areas. Recent advances in computer and display technologies strongly miniaturize available systems while increasing their performance and functionality. This opens new applications fields, especially embedding training and mission rehearsal capabilities.

Environnements virtuels d’interaction Homme-Système Intuitive

(RTO-TR-HFM-121-Part-II)

Synthèse

En ce moment, l’Alliance atlantique place de nouvelles exigences sur le personnel militaire, ce qui pousse à conduire de nouvelles approches sur l’entraînement et la conception des matériels. Les nouvelles technologies informatiques offrent aux militaires le potentiel pour se préparer à leurs nouvelles missions. La préparation de ces défis doit inclure à la fois les commandants militaires de l’entraînement, les états-majors et les opérateurs grâce à l’aide de médias appropriés, tout en développant des matériels dont la conception intégrée homme-machine a été bien pensée, et ce, pour réduire les exigences en personnel, la charge des opérateurs et le temps d’entraînement.

Le nouveau milieu, généré par ordinateur des environnements augmentés, mixtes et virtuels (AMVE), peut offrir un environnement de formation réaliste et une interaction naturelle homme-système (IHS) utilisant des environnements réalistes ou synthétiques abstraits. Les technologies AMVE permettent aux stagiaires et aux opérateurs humains d’expérimenter des environnements synthétiques correspondant aux tâches à réaliser. Un passage en revue extensif des activités nationales des nations membres a révélé que les technologies AMVE étaient devenues bien plus utiles pour différentes domaines d’application militaire qu’elles ne l’étaient il y a quelques années. Les applications de VE (environnement virtuel) disposent de nombreuses histoires à succès, démontrées, en matière d’éducation et entraînement militaire. Elles montrent que le VE peut être utile, voire parfois nécessaire, à l’achèvement d’objectifs d’entraînement. Mais, l’AMVE ne peut toujours pas être considérée comme une technologie intuitive.

Ce rapport passe en revue un certain nombre de problèmes d’utilisation humaine et d’efficacité pour autant qu’ils soient en rapport avec l’AMVE. Avant de concevoir une application AMVE, une analyse détaillée doit être faite pour identifier la possibilité de cette même technologie AMVE de représenter des processus ou des tâches complexes. Cette analyse doit se préoccuper de facteurs humains à certains niveaux différents pour utiliser au mieux les capacités de cette nouvelle technologie. Pour concevoir un nouveau système AMVE, il est important de prendre en compte les ressources pertinentes du traitement des informations humaines et des capacités au niveau de la perception. Les sentiments de présence ou d’immersion des utilisateurs affectent aussi la perception et, à terme, l’efficacité du système. D’autres facteurs, qui affectent la performance, incluent la surcharge de travail, en particulier le travail intellectuel et la maladie du simulateur. Les méthodes et mesures pour déterminer et quantifier la maladie du simulateur, causée par la technologie AMVE, sont décrites et traitées. Atteindre la prise de conscience de la situation est une autre construction complexe qui a acquis une importance vitale pour le succès de la mission. Les technologies AMVE peuvent représenter la nature complexe du champ de bataille, et on peut les utiliser pour fournir l’entraînement et les expériences nécessaires permettant aux stagiaires d’acquérir rapidement une compréhension de la situation. Finalement, on discute du problème de l’évaluation de la performance, de l’application de mesures dépendantes et des mesures de l’équipe pour évaluer le succès.

Des études détaillées de cas ont été présentées et discutées dans un atelier (RWS-136 Support virtuel pour les applications militaires) en juin 2006. Les principales conclusions de cet atelier sont aussi incluses dans ce rapport.

Le RTG (groupe de recherche) a conclu que l'AMVE est globalement applicable, et est devenu pratique dans plusieurs domaines. Sans tenir compte du fait que bien des questions restent encore sans réponses et que le sujet continue à nécessiter de nouvelles recherches. L'enseignement et l'entraînement militaires ont été identifiés comme étant l'un des principaux domaines d'application. Des avancées récentes dans le domaine des technologies informatiques et d'affichage permettent de miniaturiser encore les systèmes tout en augmentant leur performances et leur fonctionnalité. Ceci ouvre de nouveaux champs d'application, en fusionnant les capacités d'entraînement et de préparation de missions.

Introduction

The alliance is facing new challenges, including extended areas of operation, peace support missions, and combating-terrorism issues. To meet the associated military requirements, innovative concepts and technologies for an efficient and effective utilization of military forces with a limited manning level have to be developed.

Military operators usually have to interact with highly complex C4ISR systems and weapon systems under high physical, mental, and emotional workload. Therefore, the ergonomic design of human-system interaction is a critical issue. HFM-021/RSG-028 on "Human Factors in Virtual Environments" has identified Virtual Reality (VR) and Virtual Environment (VE) systems to be advantageous in facilitating a close, natural, and intuitive interaction by making better use of human perceptive, cognitive, and motor capabilities (RTO-TR-018). It was summarized, that VEs have become a useful technology for early phases of systems engineering like virtual product prototyping. Moreover, innovative approaches to integrate VEs into military mission support were identified.

But whereas computing, rendering, and display technologies have made a tremendous advance in recent years, the ergonomic design of the human-system interface has not. As a matter of fact, VE systems have to be operated by specially trained personnel and applications are often limited to passive presentations. User interfaces of VE systems are usually prototypic and are derived from common 2D graphical user interfaces. VE systems are often used only as an extension of existing concepts without exploiting their full interaction potential. Therefore, a significantly better integration into novel concepts for training, system design as well as command and control is required.

The operator-interactive part of a virtual environment must take task dynamics into account and should augment human perception, cognition, and decision making. It has to be systematically designed and evaluated on pragmatic, semantic, syntactic, and lexical levels.

Possible military applications of such advanced and intuitive VE systems were found as:

- Dynamic, task-driven user interfaces for C4ISR systems;
- Telepresence, teleoperation, and telemanipulation in reconnaissance, surveillance, and target acquisition;
- Realistic and distributed military simulation and training;
- Short-term mission preparation, including intended area of operation; and
- Mission support as wearable, augmenting technology for individual soldiers (including Mixed Reality and Augmented Reality).

The Task Group was initiated in order to investigate Augmented, Mixed and Virtual Environments (AMVE) as a mean of providing an advanced and intuitive human-system interaction for multiple military applications and report on the state-of-the-art and its potential.

During its three years duration the group started updating and extending the results of past NATO RTO Research Study Groups, especially of NATO HFM-021 on Human Factors in Virtual Reality. According to the prior work of that group, the Task Group has adopted the following definition of the term Virtual Reality:

> *"Virtual Reality is the experience of being in a synthetic environment and the perceiving and interacting through sensors and effectors, actively and passively, with it and the objects in it, as if they were real. Virtual Reality technology allows the user to perceive and experience sensory contact and interact dynamically with such contact in any or all modalities."*

The definition considers conceptual as well as technological aspects. It addresses the subjective, individual impressions of Virtual Reality (or Virtual Environments, respectively), which are caused by natural perceiving and experiencing computer-generated scenarios.

The first goal of the new Task Group was to update national activities in this field. Therefore, the main national activities were surveyed. The comprehensive results of the survey were published as first report (NATO TR-HFM-121-Part-I). Based on these findings, the group organized a NATO RTO-workshop on Virtual Media for Military Applications at the US Military Academy in West Point, NY, to bring together specialists from academia, developers, and military users. The goal of the workshop was to assess operational capabilities and potentials of AMVE. Promising (future) and successful (present) applications were presented. Keys for implementation of ready-to-use technology and knowledge gaps and thrusts were identified. A general roadmap for future development and application was discussed with the participants. The proceedings were independently published by NATO RTO as NATO MP-HFM-136. They specify the users' perspective and the operational needs for AMVE technology. A summary of the important findings is also included in this report.

It was apparent to the Task Group that there is a need for qualitative and quantitative measures to differentiate between AMVE-technologies. They define relevant criteria for applicability of special technology for a special purpose. Given the background of the Task Group under the Human Factors and Medicine Panel, these measures refer primarily to ergonomic measures, and not to technical measures (e.g., weight, transportability) or computer science (e.g., rendering).

The report is structured to mirror different aspects of AMVE technology. It refers to human factors issues in the design of AMVE systems, general relevant issues evolving from practical use and application, methods for evaluating performance and benefits of the new technology for special applications, and recommendations for the military user community.

From a technical perspective, it is important for the design of such systems to consider relevant human information processing resources and capabilities. This affects usability of the system as well as willingness and compliance of the user when working with the system strongly. There are several factors: general ergonomic factors (e.g., weight and fit of devices); physiological factors (e.g., visual, acoustic, haptic perception); psychological factors (e.g., scene realism, breaks in presence). Chapter 2 addresses these factors in details and specifies the according ergonomic requirements on system properties. It describes visual requirements, audio requirements, haptic feedback, and also exotic feedback modalities as olfactory feedback. In addition to the resulting fundamental issues, the concept of presence or immersion is tightly connected to AMVE. Presence may be loosely described as the feeling of "being there", in the virtual world instead of just experiencing stimuli from displays. The topic has caught a lot of attention in research. The according Chapter 3 presents an overview of research on presence. It introduces theories, definitions, multi-dimensionality of affecting factors, effects, measurement, and determinants.

The following three chapters refer to human factors evolving from practical use and application of AMVE systems, especially in training. They should be considered when determining usability and benefits or shortcomings of AMVE system usage. Generally speaking, workload, especially mental workload, is a factor resulting from the use of new technology affecting usability and effectiveness. There is a vast diversity of literature about research on workload and workload measurement techniques available. It is structured, described, and referenced in Chapter 4. Mental workload is multidimensional and not seriously challenged today. Yet, the application of different workload measurement techniques appears to be most reasonable. Psychophysiological measures are not recommended for applied problems until researchers can develop a formal, unifying theory that gives explanations. Instead, subjective ratings like NASA TLX or SWAT are a reasonable alternative. The following Chapter 5 deals with a negative side effect of AMVE technology, which is simulator sickness. Simulator sickness is a form of motion sickness that does not require motion. Motion sickness is multi-symptom due to abrupt, periodic, or unnatural accelerations.

Simulation sickness is primarily caused by these accelerative forces or by visual motion cues without actual movement. This chapter is the result of a comprehensive research in this area. It describes methods for determining and quantifying simulator sickness due to AMVE technology.

When deciding about practical applicability it is important to have means for performance evaluation at hand. Performance evaluation refers to application-dependent measures, and not to global concepts and issues as described in the prior chapters. The following set of three chapters addresses topics for performance evaluation of AMVE systems. A specific concern of training systems is situational awareness, which is the topic of Chapter 6. It is especially important with present military scenarios, characterized by great uncertainty of own and opponent forces. Understanding who are the combatants, civilians, and allied personnel, as well as knowing the rules of engagement for the given situation are all part of the soldiers' situational awareness. General performance measures are particularly relevant for developing methods and metrics to assess usability of AMVE for training purposes. Chapter 7 refers to team and collective performance. It describes different techniques and measures for collective team performance in connection with VR-systems. The final chapter on performance evaluation deals with a special method in simulation-based training: the after action review. The method is well-established for training sessions, but by applying AMVE-technology for training it is possible to use it more extensively. Chapter 8 describes the methodological background for it and presents capabilities of an after action review tool for mission analysis using VR technology. This way a mixture of intrinsic and extrinsic feedback can prepare individuals for their later task and enhances training effectiveness.

So far, the chapters gave recommendations about systems' design and use, as well as methods and technologies for performance evaluation. Another goal of this report was to give recommendations for military applications and future development. Because this was the motivation of the NATO HFM-136 workshop, Chapter 9 summarizes the main points of the workshop. It describes the technological development in recent years, the development in the main application areas, and future potential applications. Main keyword and topics are highlighted.

The final chapter of this report summarizes the main findings of the three years duration and several meetings of the group. It puts the development and the (past) promises into perspective. Moreover, the conclusion drafts a general roadmap for the future maturity of AMVE-technology in different military domains.

Perceptual Issues of Augmented and Virtual Environments

Helge Renkewitz, Thomas Alexander
Research Institute for Communication
Information Processing, and Ergonomics (FKIE)
Neuenahrer Strasse 20, 53343 Wachtberg
GERMANY

{renkewitz, alexander}@fgan.de

1.0 INTRODUCTION

For a sensible application of Augmented Reality (AR) and Virtual Environments (VE) it is necessary to include basic human information processing resources and characteristics. Because there is no fully functional model of human perceptual, cognitive, and motor behavior, this requires empirical analyses. Moreover, these analyses are often based on subjective ratings rather than objective measures. With regard to perception as the basic sensation of synthetic environments, each modality should be analyzed separately. There are special limits of human perception which limit the transfer of information of might even lead to unwanted negative effects or after-effects when not taken into consideration. One example for this is long exposition times and emotional inclusion of the user. They may even cause a user's isolation from the "real" daily life. In addition to a purely short-term, technological sight, it is necessary to evaluate the application of AR and VE in terms of its psychological and sociological impact.

Aspects of visual feedback are very important because of the dominance of the visual modality. The usability of the display is an important factor for the user's willingness and compliance to spend long times immersed in the virtual world. For example, HMDs need not to be too heavy, too large or too tightly fit. This category of factors groups the *General Ergonomic Factors*. The second category deals with *Physiological Factors* influencing vision. They subsume, e.g., graphics refresh rate, depth perception and lighting level influencing human performance with a VE display systems. One example is that more than 25 images per second in a dark environment cause the illusion of a continuous motion rather than single flickering images. However, the graphics refresh rates depends on the scene complexity expressed in number of polygons and shaded modality and not only on update rate of the display device itself. The third category of factors deals with *Psychological Factors* such as scene realism, scene errors (scale errors, translation errors, etc.) and the integration of feedback and command. It refers to the modification of the scene as a function of task-specific information. Markers or additional functionality can be added to the virtual world, which should help the user in performing several tasks. An example is an "intelligent agent" or tutor who serves as a figurative, anthropomorphic representation of the system status.

Acoustic feedback has a dual role. First, it is the medium for transmitting information. Second, it can be used to localize the source of the information. Ergonomic factors refer to the design of the hardware and its ease of use by humans. Physiological conditions refer to the sound frequency range which has to be within the range of audible sound (20 to 20.000 Hz) and sound intensity. If the intensity is too strong, it can produce discomfort or even, above 120 db, pain. Another factor is the sound/noise ratio. A more complex area is described by psychological factors. Sound perception and processing allows the mental reconstruction of a world that is volumetric and whose parts have specific conceptual components. A piano, for example, should not generate drum sound. Another example is a complex control panel, which concludes a large amount of visual feedback. An audio alarm can raise the user's attention to error conditions. Finally, sound or speech recognition can also be used as another, very natural input modality of the user.

Physical contact with the environment provides another important feedback. Some virtual tasks, especially manual manipulation, can only be performed accurate by adding tactile feedback to the environment.

But this is often difficult. The aim for the future is to provide touch and force feedback to the whole body. Today, haptic feedback stimulation is usually restricted to one hand only. Fortunately, many real tasks can be carried out like this. Therefore, this restriction does no degrade human performance.

Long immersion into a synthetic environment is likely to cause several severe effects. Simulation sickness, resulting into dizziness, nausea, and disorientation is thought to be caused by a sensorial conflict between visual feedback indicating motion and the kinesthetic cuing. The phenomenon is aggravated by poor image resolution.

Factors which have been identified as contributors to simulator sickness in virtual environment systems are shown in the following Table (Frank et al., 1983; Kennedy et al., 1989; Kolasinski, 1995; Pausch et al., 1992). These are divided into characteristics of the user, the system and the user's task. Few systematic studies have been carried out to determine the effects of the characteristics of virtual environment systems on the symptoms of simulator sickness. Hence much of the evidence for the effects of these factors comes from studies of visually-induced motion sickness and motion-induced sickness (i.e., sickness caused by actual vehicle motions), as well as the effects of exposures to simulators.

Table 2-1: Factors Contributing to Simulator Sickness in Virtual Environments (Kennedy et al. 1989)

User Characteristics	**System Characteristics**	**Task Characteristics**
Physical Characteristics Age Gender Ethnic origin Postural stability State of health *Experience* With virtual reality system With corresponding real-world Task *Perceptual Characteristics* Flicker fusion frequency Mental rotation ability Perceptual style	*Display* Contrast Flicker Luminance level Phosphor lag Pefresh rate Pesolution *System Lags* Time lag Update rate	*Movement through Virtual Environment* Control of movement Speed of movement *Visual Image* Field of view Scene content Vection Viewing region Visual flow *Interaction with Task* Duration Head movements Sitting vs. standing

2.0 USER CHARACTERISTICS

Physical Characteristics: Age has been shown to affect susceptibility to motion-induced motion sickness. Motion sickness susceptibility occurs most often for people between ages of 2 and 12 years. It tends to decrease rapidly from the age of 12 to 21 years and then more slowly through the remainder of life (Reason and Brand, 1975).

Females tend to be more susceptible to motion sickness than males. The differences might be due to anatomical differences or an effect of hormones (Griffin, 1990). In a study on the occurrence of sea-sickness on a ship, vomiting occurred among 14.1% of female passengers, but only 8.5 % of male passengers (Lawther and Griffin, 1986). As seasickness is another motion-induced sickness, gender effects are likely to exist for simulator sickness as well.

Ethnic origin may affect susceptibility to visually-induced motion sickness. Stern et al. (1993) have presented experimental evidence to show that Chinese women may be more susceptible than European-American or African-American women to visually-induced motion sickness. A rotating optokinetic drum was used to provoke motion sickness. The Chinese subjects showed significantly greater disturbances in gastric activity and reported significantly more severe motion sickness symptoms. It is unclear whether this effect is caused by cultural, environmental, or genetic factors.

Postural stability has been shown to be affected by exposure to virtual environments and simulators (Kennedy et al., 1993, 1995). Kolasinski (1995) has presented evidence to show that less stable individuals may be more susceptible to simulator sickness. Pre-simulator postural stability measurements were compared with post-simulator sickness data in Navy helicopter pilots. Postural stability was found to be associated with symptoms of nausea and disorientation, but not with ocular disturbances.

The state of health of an individual may affect susceptibility to simulator sickness. It has been recommended that individuals should not be exposed to virtual environments when suffering from health problems including flu, ear infection, hangover, sleep loss or when taking medications affecting visual or vestibular function (Frank et al., 1983; Kennedy et al., 1987, 1993; McCauley and Sharkey, 1992). Regan and Ramsey (1994) have shown that drugs such as hycosine hydrobromide can be effective in reducing symptoms of nausea (as well as stomach awareness and eyestrain) during immersion in VE.

Experience: Nausea and postural problems have been shown to be reduced with increased prior experience in simulators (Crowley, 1987) and immersive VEs (Regan, 1995). Frank et al. (1983) have suggested that although adaptation reduces symptoms during immersion, re-adaptation to the normal environment could lead to a greater incidence of post-immersion symptoms. Kennedy et al. (1989) have also suggested that adaptation cannot be advocated as the technological answer to the problem of sickness in simulators since adaptation is a form of learning involving acquisition of incorrect or maladaptive responses. This would create a larger risk of negative training transfer for individuals. For instance, pilots with more flight experience may be generally more prone to simulator sickness (Kennedy et al., 1987). This may be due to their greater experience of flight conditions, leading to greater sensitivity to discrepancies between actual and simulated flight. Another reason might be the smaller degree of control when acting as instructors in simulators (Pausch et al., 1992).

Perceptual Characteristics: Perceptual characteristics which have been suggested to affect susceptibility to simulator sickness include perceptual style, or field independence (Kennedy, 1975; Kolasinski, 1995), mental rotation ability (Parker and Harm, 1992), and level of concentration (Kolasinski, 1995).

3.0 SYSTEM CHARACTERISTICS

Characteristics of the Display: Luminance, contrast and resolution should be balanced with the task to be performed in order to achieve optimum performance (Pausch et al., 1992). Low spatial resolution can lead to problems of temporal aliasing, similarly to low frame rates (Edgar and Bex, 1995).

Flicker of the display has been cited as a main contributor to simulator sickness (Frank et al., 1983; Kolasinski, 1995; Pausch et al., 1992). It is also distracting and contributes to eye fatigue (Pausch et al., 1992). Perceptible flicker, i.e., the flicker fusion frequency threshold, is dependent on the refresh rate, luminance and field-of-view. As the level of luminance increases, the refresh rate must also increase to prevent flicker. Increasing the field-of-view also increases the probability of perceiving flicker because the peripheral visual system is more sensitive to flicker than the fovea. There is a wide range of sensitivities to flicker between individuals, and also a daily variation within individuals (Boff and Lincoln, 1988).

Other visual factors, which contribute to oculomotor symptoms reported during exposure to virtual environments, have been discussed extensively by Mon-Williams et al. (1993), Regan and Price (1993) and Rushton et al. (1994).

System Lags and Latency: Wioka (1992) has suggested that lags of less than 300 ms are required to maintain the illusion of immersion in a VE, because otherwise subjects start to dissociate their movements from the associated image motions (Wioka, 1992; Held and Durlach, 1991). It is unclear whether the authors attribute these effects to pure lags or the system update rates. However, lags of this magnitude, and update rates of the order of 3 frames per second, have both been shown to have large effects on performance and on subjects' movement strategies. The total system lag in the VE-system used in the experimental studies reported by Regan (1995) and Regan and Price (1994) was reported to be 300 ms (Regan and Price, 1993c).

There is an urgent need for further research to systematically investigate the effect of a range of system lags on the incidence of simulator sickness symptoms. The interaction between system lags of head movement velocity is likely to be important, since errors in the motion of displayed images are proportional to both total lag and head velocity.

Previous studies considering hand- and head-movements show that users are very sensitive to latency changes. Subjects were able to detect latency changes with a PSE of ~50 ms and a JND of ~8 – 15 ms, respectively (Ellis et al., 1999a; Ellis et al. 1999b). When examining random vs. paced head-movements PSEs of ~59 ms and JNDs of ~13.6 ms were determined (Adelstein et al., 2003). The same values are determined with changing visual condition (background, foreground) or realism of the VE (Mania et al., 2004; Ellis et al., 2004). Pausch (1992) cites data from Westra and Lintern (1985) to show that lags may affect subjective impressions of a simulator even stronger than they affect performance. Simulated helicopter landings were compared with visual lags of 117 ms and 217 ms. Only a small effect on objective performance measures occurred, but pilots believed that the lag had a larger effect than was indicated by the performance measures.

Richard et al. (1996) suggested that the frame rate (i.e., the maximum rate at which new virtual scenes are presented to the user) is an important source of perceptual distortions. Low frame rates make objects appear to move in saccades (discrete spatial jumps). Thus, the visual system has to bridge the gaps between perceived positions by using spatio-temporal filtering. The resulting sampled motion may also result in other artifacts such as motion reversals (Edgar and Bex, 1995). Low frame rates (particularly when combined with high image velocities) may cause the coherence of the image motion to be lost, and a number of perceptual phenomena may occur, including appearance of reversals in the perceived motion direction, motion appearing jerky, and multiple images trailing behind the target. This phenomenon is referred to as temporal aliasing. Edgar and Bex (1995) discuss methods for optimizing displays with low update rates to minimize this problem.

4.0 TASK CHARACTERISTICS

Movement through the Virtual Environment: The degree of control of the motion affects general motion-induced sicknesses and simulator sickness. The incidence of simulator sickness among air-crew has been reported to be lower in pilots (who are most likely to generate control inputs) than in co-pilots or other crew members (Pausch et al., 1992).

The speed of movement through a virtual environment determines global visual flow, i.e., the rate at which objects flow through the visual scene. The rate of visual flow influences vection and is related to simulator sickness (McCauley and Sharkey, 1992). Other motion conditions that have been observed to exacerbate sickness in simulators include tasks involving high rates of linear or rotational acceleration, unusual maneuvers such as flying backwards and freezing, or resetting the simulation during exposures (McCauley and Sharkey, 1992).

Regan and Price (1993c) have suggested that the method of movement through the virtual world affects the level of side-effects. Experiments to investigate side-effects in immersive VE have utilized a 3D mouse to generate movement (Regan, 1995; Regan and Price, 1993c, 1994; Cobb et al., 1995). This is likely to generate conflict between visual, vestibular and somatosensory senses of body movement. A more natural movement might be provided by coupling movement through a virtual environment to walking on a treadmill (Regan and Price, 1993c).

Visual Image: A wider field-of-view may enhance performance in a simulator, but also increase the risk of simulator sickness (Kennedy et al., 1989; Pausch et al., 1992). This happens although the effect of field of view is often confounded with other factors (Kennedy et al., 1989). Stern et al. (1990) have shown that restricting the width of the visual field to 15 degrees significantly reduces both. Circular vection and the symptoms of motion sickness induced by a rotating surround with vertical stripes (optokinetic drum). Fixation on a central point in the visual field also reduces the circular vection induced by rotating stripes observed with peripheral vision, and greatly reduces motion sickness symptoms (Stern et al., 1990). Circular vection increases with increasing stimulus velocity up to about 90 degrees per second (Boff and Lincoln, 1988). Further increases in stimulus velocity may inhibit the illusion. Vection is not dependent on acuity or luminance (down to scoptopic levels) (Liebowitz et al., 1979).

Linear vection can be induced visually by expanding pattern of texture points. Anderson and Braunstein (1985) showed that linear vection could be induced by a moving display of radial expanding dots with a visual angle as small as 7.5° in the central visual field. They suggested that the type of motion and the texture in the display may be as important as the field-of-view in inducing vection. The incidence of simulator sickness has been shown to be related to the rate of global visual flow, or the rate at which objects flow through the visual scene (McCauley and Sharkey, 1992). The direction of self-motion can be derived from the motion pattern of texture points in the visual field (Warren, 1976; Zacharias et al., 1985). The optical flow field appears to expand from a focal point, which indicates the direction of motion. For curved motion the expanding flow field tends to bend sideways, and the focal point is no longer defined. Grunwald et al. (1991) have shown how unwanted image shifts, which are due to lags in a flight simulator with a head-coupled head-mounted display, distort the visual flow field. In straight and level flight, the unwanted image motions which occur during head movements will cause the expanding visual pattern to appear to bend, creating the illusion of a curved flight path. The bending effect is proportional to the ratio of the magnitude of the image shifts and the apparent velocity along the line of sight. The apparent velocity depends on the velocity to height ratio. Hence the angular errors induced by the bending effect increase with decreased velocity and increased altitude.

Linear vection has been observed to influence postural adjustments made by subjects in the forward and rear direction. Lestienne et al. (1977) observed inclinations of subjects in the same direction as the movement of the visual scene movement, with a latency of 1 to 2.5 s, and an after-effect on the cessation of motion. The amplitude of the postural adjustments was proportional to the image velocity.

Interaction with the Task: Short exposure duration of less than 10 minutes to immersive virtual environments has already been shown to result in significant incidences of nausea, disorientation and ocular problems (Regan and Price, 1993c). Longer exposures to virtual environments can result in an increased incidence of sickness and require longer adaptation periods (McCauley and Sharkey, 1992). The severity of motion-induced sickness symptoms have been shown to increase with the duration of exposure to the provocation for duration up to at least 6 hours (Lawther and Griffin, 1986). Kennedy et al. (1993) reported that longer exposures to simulated flight increased the intensity and duration of postural disruption.

The extent of image position errors, and conflicts between visual and vestibular motion cues, will depend on the interaction between head motions and the motions of visual images on the display. Head movements in simulators have been reported to be very provocative (Lackner, 1990, reported by Pausch et al., 1992). However Regan and Price (1993c) found that over a ten minute period of immersion in a virtual environment, there was no significant effect of type of head movement on reported levels of simulator sickness. Sickness incidence was compared between two ten minute exposures to an immersive virtual environment. One exposure involved pronounced head movements and rapid interaction with the system. During the other exposure, subjects were able to control their head movements and their speed of interaction to suit them. There was some evidence that the pronounced head movements initially caused higher levels of symptoms, but that subjects adapted to the conditions by the end of the exposures. No measurements were made of head movements, so the effect of the instructions given to the subjects on the velocity and duration of head movements is unclear. The system lag was reported to be 300 ms, so even slow head movements may have been expected to result in significant spatio-temporal distortions. The authors suggest an urgent need for further research to systematically investigate the interaction between system lags and head movement velocity with the incidence of side-effects.

The levels of symptoms reported by seated subjects after immersion in a virtual environment have been reported to be slightly higher than the level of symptoms reported by standing subjects (Regan and Price, 1993c). However, the differences were not statistically significant after ten minute exposures.

The European Telecommunications Standards Institute has published several reports about Human Factors in many areas of computer science. In ETSI (2002) guidelines for the design and use of multimodal symbols is presented. It provides a study of the needs and requirements for the use of multimodal symbols in user interfaces, which can be also adapted to VE.

5.0 PERCEPTUAL REQUIREMENTS

5.1 Visual Requirements

Most environmental information is gained through the visual modality. The physiology of eye determines limitations and requirements for displaying information on a computer display. With current technology a faster presentation of information is possible than perception and processing of the information by the human. Therefore, Human-Computer-Interaction is mainly caused by the human operator and not the computer.

Basic visual perception starts with a projection of the image of the environment onto the retina. Special photoreceptors transform the visual stimuli into electronic stimuli. There are two different types of photoreceptors on the retina which are commonly referred to as “rods” and “cones”. Rods are sensitive to light, but saturate at high levels of illumination whereas cones are less sensitive, but can operate at higher luminance levels (Monk, 1984). Rods occur predominantly near the fovea, or focal point of the eye image and the cones are more predominant around the periphery. This results into a relatively small angle of view for clear and sharp images with a size of 1 or 2 degrees only. With growing angles, sharpness decreases rapidly. Consequently, information should be displayed within this small angle. Otherwise the eye has to moving continuously in order to catch a complete glimpse. For a complete overview additional cognitive resources are required to assimilate the single views into a complete mental page. In combination with the capacity of short term memory this allows only a small amount of information that can be displayed on a single screen.

The eye’s ability to distinguish color, luminance, contrast and brightness is another factor that has to be considered. The color of an object is determined by the frequency of the light that is reflected from it. The visible spectrum reaches from blue at 300 nm to red at 700nm. Different colors are obtained through combinations of wavelengths throughout this wavelength range. Color sensitivity is created by the existence of three different types of cones in the eye: blue, green, and red. Each type of cone responds to a certain, not exact, range of wavelengths. By combining wavelengths, the human eye can distinguish more than 8,000 different colors (Monk, 1984). Approximately 8% of the male population and less than 1% of the female population suffer from color blindness to some degree. Color blindness is the inability to distinguish certain colors, notably reds and greens. This fact is also important to remember when designing visual displays for a larger user group.

Luminance is a measure of the amount of light reflected from a surface. It is determined by the amount of light that shines on an object and the reflectance of the surface of the object. Its unit of measure is Candela per square Metre (cd/m^2). Research has determined that there is a range of optimal luminance levels and that low illumination can be a hindrance to an otherwise good HCI.

Contrast is defined as the difference between the luminance of an object and its background divided by the luminance of the background (Downton, 1991). It is a measure of an eye's ability to distinguish foreground from background easily. A bright background with black writing has a low luminance for the writing and a high luminance for the background. This screen therefore, has a negative contrast. The higher the absolute value of the contrast the easier it is to distinguish objects.

Brightness is usually thought of as a subjective property of light. It depends on many factors. The main one is comparative illumination. A cloudy day may seem quite dull. The same day would be quite bright if you were just emerging from a dark room. Brightness contrast can cause several common optical illusions as well.

5.2 Special Visual Issues

There are several other issues which have to be considered when designing visual output. They are based on characteristics and deficits of human visual perception.

5.2.1 Eye Dominance

The majority of people have a distinct preference for one eye over the other. This is typically, quickly, and easily found through sighting tests (Peli, 1990). This eye dominance has shown only a limited performance advantage in military targeting tasks (Verona, 1980). Yet, the dominate eye will be less susceptible to suppression in binocular rivalry and this likelihood of suppression will further decrease over time.

An estimated 60% of the population is right eye dominant. Subsequently, it is evident that eye dominance does not correspond with users being left or right handed as only 10% of the population is left handed.

5.2.2 Pupil Adaptation

For controlling the amount of light entering the eye, the pupil will constrict (reducing the amount of light) or dilate (letting more light in). When the illumination is suddenly increased, the pupil will overcompensate by constricting and then dilating slowly to match the light level. After reducing the illumination the pupil cycles through several dilations and constrictions. Complete constriction may take less than one minute, but complete dilation may take over 20 minutes (Alpern and Campbell, 1963). This is caused partially by the fact that the cones (responsible for color perception) recover more quickly than rods (which are responsible for night vision), but have lower sensitivity. The size of the pupil will decrease once a target gets closer than I meter away (Alpern and Campbell, 1963). This is very likely due to the increase luminance caused by the light reflected off the target.

5.2.3 Visual Field

The visual field (the area the eye can perceive) is roughly 60 degrees above and below the center and slightly over 90 degrees to the outside (and 60 degrees the inside for each eye, where it is partially blocked by the nose). The lateral visual field slowly declines with age. At the age of 20 it has a size of nearly 180 degrees horizontally. At the age of 80 it is reduced to 135 degrees. Women have slightly larger visual fields then men, primarily due to differences of the nasal side (Burg, 1968).

5.2.4 Accommodation

Accommodation is the focusing of the lens of the eye through muscle movement. As humans get older, their ability (speed and accuracy) to accommodate decreases (Soderberg et al., 1993). For instance, the time to accommodate between infinity to 10” for a 28 year-old takes .8 seconds while a 41 year-old will take about 2 seconds (Kruger, 1980). The ability to rapidly accommodate appears to decline at the age of 30 and those over 50 will suffer the most. Younger humans (under the age of 20) will accommodate faster regardless of target size. However, the ability to accommodate may begin to decline as early as age 10. Accommodation for binocular viewing is both faster and more accurate than monocular viewing for all age groups (Fukuda et al., 1990). The Resting Point of Accommodation (RPA) describes the accommodation state the eye assumes when at rest. It migrates inward over time. In addition, the response time to obtain both the RPA and far point focus increase over time (Roscoe, 1985). Given these changes a VVS (**V**irtual **V**iew **S**ystem) with adjustable focus is likely to lead to improved product usability.

5.2.5 Sensitivity to Flicker

Sensitivity to flicker is highest when the eyes are light adapted. Thus users may notice flicker in the display until their eyes dark adapt. The periphery of the eye is also more sensitive to flicker and motion detection, and the closer an object is to the eye, the more likely that flicker can be detected (Kelly, 1969).

5.2.6 Vision Deficiencies

There are a wide variety of visual deficiencies in the visual system that may occur in to members of the general population. If untreated, these may lead to discomfort when using visual displays. An example of the most common of these problems will be briefly discussed in the following.

In his review of the “Private Eye” viewing device, Peli (1990) reported a large portion of the discomfort associated with the display was due to pre-existing visual conditions. This was confirmed by Rosner and Belkin (1989) who recommend a complete eye exam and correction for existing visual problems be undertaken prior to using a display system. These problems will become more prevalent with older users. Visual acuity and performance decline with age. People in their 20’s tend to have 20/20 vision on average; younger subject may have 20/15 vision. With progressing age visual acuity decreases to 20/30 by age 75 (Owsley et al, 1983).

It is estimated that 3% to 4% of the general population suffer from strabismus, which describes the inability to focus both eyes to the same single point. This condition usually develops before the age of eight and is hereditary in many cases. Patients with early, untreated strabismus will also likely develop amblyopia (lazy eye phenomenon). This is a condition in which one eye will drift while the other remains focused on an object. Both lead impaired depth perception. It is estimated, that approximately 2% of the general population suffer from it (Peli, 1990).

Phoria is the tendency for a covered eye to deviate from the fixation point of the open eye. While these deviations can be very larger even after only several hours of occlusion, normal vision will return after only 1 minute (Peli, 1990). Phoria can cause the temporary elimination or reduction of stereoscopic depth perception even after both eyes are uncovered. Additional research on adults has shown that even after eight days of one-eye occlusion subjects were able to regain normal vision hours after both eyes were uncovered. Measurable, though slight phoria was found to exist after using the “Private Eye” monocular viewing device (Peli, 1990). Changes in phoria are most likely to occur in individuals who already suffer from uncorrected visual problems (Saladin, 1988). Half of patients with near- or far-sightedness suffer from additional hyperphoria, a tendency for the eyes to drift upward. This also affects depth perception.

For the development of normal binocular vision, each eye must function well throughout the early development years during childhood. This period of development is most sensitive to disruption up to age of five years and remains critically until the age of nine years when the visual system matures (Peli, 1990). While constant use of a visual display by a person under the age of six years could lead to visual problems, it is doubtful that most of the common VR-displays can be worn comfortably by such young users. Nor is it likely that they could use such a display long enough. In addition, common AR-displays are often designed as see-through device. It is doubtful that they will attend to the monocular stimulus for a sufficient amount of time to cause permanent damage.

5.3 Audio Requirements

Although it is no question that visual is the primary modality for transferring information from a computer, practically each personal computer has a sound card today. Audio is becoming a common way of presenting additional information. Many help packages for software have an audio as well as visual component. Having a basic understanding of human hearing, capabilities and limitations also helps the designer in setting-up audio VR-components.

Hearing basically involves the same problems as seeing: Perception of environmental stimuli, translating them into nerve impulses, and combining meaning to them (Sutcliffe, 1989). At a physical level, audio perception is based on sound waves. They travel as longitudinal waves through air or other media. Sound is characterized by frequency and amplitude. Frequency determines the pitch of the sound and amplitude determines its volume. Frequency is measured in cycles per second or hertz, with 1 cycle per second equaling 1 hertz. Young children can hear in the range of about 20 Hz to over 15,000 Hz. This range decreases with age. Audible speech is between 260 and 5600 Hz – but even with a limited range between 300 and 3000 Hz communication (telephone transmission) is still possible (Sutcliffe, 1989). Speech, as well as most everyday sounds, is a very complex mixture of frequencies.

The volume or intensity of a sound is expressed in decibels (dB). This is a logarithmic expression for the ratio between the amplitude of the primary sound to the background sound and gives a measurement of the ability to hear what is intended. A whisper is 20 dB. Normal speech registers between 50 and 70 dB. Hearing loss can result from sounds exceeding 140 dB (Downton, 1991). Below 20 dB sounds can be heard, but they are not distinguishable. The ear cannot determine frequency changes below this level.

More important for acoustic perception than physical characteristics of sound is the human ability to interpret sound. The auditory centre of the cortex appears to be able to distinguish three different types of sound: background unimportant sounds (noise), background sounds that have significance (child’s cry, dog’s bark, etc.) and speech (Sutcliffe, 1989). Language is full of mispronounced words, unfinished sentences, missing words, interruptions, etc., but the brain still has to be able to interpret it. This seems to be done by comparison to past experience and analyzed as a stream. The same sounds can therefore be “heard” differently depending on the context. Speech is continuous. When analyzed, it doesn’t appear as disjointed syllables or phonemes, but as a continuous stream that must be interpreted at a rate of between 160 and 220 words per minute (Sutcliffe, 1989).

5.3.1 Sound Perception

There are several auditory localization cues to help locate the position of a sound source in space. The first is the interaural time difference. This means the time delay between sounds arriving at the left and right ears. The second one is head shadow. It defines the time for a sound to go through or around the head before reaching an ear. The third one is pinna response. It is the effect that the external ear, or pinna, has on sound. The forth one refers to the shoulder echo. It describes the reflection of the sound in the range of 1 – 3 kHz by the upper torso of the human body.

The fifth localization cue is caused by movement of the head. It helps to determine a location of a sound source. Another one is the occurrence of early echo response in the first 50 – 100 ms of a sounds life. Further reverberations are caused by reflections from surfaces around. The final cue is the visual modality, which helps us to quickly locate and confirm the location and direction of a sound.

5.3.2 Sound Processing

VR immersive quality can be enhanced through the use of properly cued, realistic sounds. For the design of a VR system synthetic sounds have to be generated like those in the real world. Sound processing includes encoding of directional localization cues on several audio channels, transmission or storage of sound in a certain format and the playback of sound.

5.3.2.1 Different Types of Sounds

Mono sound:

- Recorded with one microphone; signals are the same for both ears.
- Sound only at a single point (“0”-dimensional), no perception of sound position.

Stereo sound:

- Recorded with two microphones several feet apart and separated by empty space; signals from each microphone enter each single ear respectively.
- Perceived commonly by means of stereo headphones or speakers; typical multimedia configuration of personal computers.
- Gives a better sense of the sound’s position as recorded by the microphones, but only varies across one axis (1-dimensional), and the sound sources appear to be at a position inside the listener’s head.

Binaural Sound:

- Recorded in a manner more closely to the human acoustic system: by microphones embedded in a dummy head.
- Sounds more realistic (2-dimensional), and creates sound perception external to the listener’s head.
- Binaural sound was the most common approach to specialization; the use of headphones takes advantage of the lack of crosstalk and a fixed position between sound source (the speaker driver) and the ear.

3D Sound:

- Often termed as spatial sound, is sound processed to give the listener the impression of a sound source within a three-dimensional environment.
- New technology under developing, best choice for VR systems.
- The definition of VR requires the person to be submerged into the artificial world by sound as well as sight. Simple stereo sound and reverb is not convincing enough, particularly for sounds

coming from the left, right, front, behind, over or under the person – 360 degrees both azimuth and elevation. Hence, 3D-sound was developed.

5.3.2.2 3D Sound Synthesis

3D Sound synthesis is a signal processing system reconstructs the localization of each sound source and the room effect, starting from individual sound signals and parameters describing the sound scene (position, orientation, directivity of each source and acoustic characterization of the room or space).

Sound rendering is a technique that creates a sound world by attaching a characteristic sound to each object in the scene. This pipelined process consists of four stages:

1) Generation of each object's characteristic sound (recorded, synthesized, modal analysis-collisions).
2) Sound instantiation and attachment to moving objects within the scene.
3) Calculation of the necessary convolutions to describe the sound source interaction within the acoustic environment.
4) Convolutions are applied to the attached instantiated sound sources.

Its similarity to ray-tracing and its unique approach to handling reverberation are noteworthy aspects, but it handles the simplicity of an animated world that is not necessarily real-time.

Modeling the human acoustic system with head-related transfer function (HRTF) is another approach. The HRTF is a linear function that is based on the sound source's position and takes into account many of the cues humans use to localize sounds. Here, the process works as follows:

- Record sounds with tiny probe microphones in the ears of a real person.
- Compare the recorded sound with the original sounds to compute the person's HRTF.
- Use HRTF to develop pairs of finite impulse response (FIR) filters for specific sound positions.
- When a sound is placed at a certain position in virtual space, the set of FIR filters that correspond to the position is applied to the incoming sound, yielding spatial sound.

The computations are so demanding that they currently require special hardware for real-time performance.

3D sound imaging approximates binaural spatial audio through the interaction of a 3D environment simulation. First the line-of-sight information between the virtual user and the sound sources is computed. Subsequently, the sounds emitted by these sources will be processed based on their location, using some software DSP algorithms or simple audio effects modules with delay, filter and pan and reverb capabilities. The final stereo sound sample will then be played into a headphone set through a typical user-end sample player, according to the user's position. This approach is suitable for simple VE systems where a sense of space is desired rather than an absolute ability to locate sound sources.

The utilization of speaker locations works with strategically placed speakers to form a cube of any size to simulate spatial sound. Two speakers are located in each corner of the cube, one up high and one down low. Pitch and volume of the sampled sounds distributed through the speakers appropriately give the perception of a sound source's spatial location. This method has less accuracy than sound yielded by convolving sound, but yields an effective speedup of processing, allowing a much less expensive real-time spatial sound.

5.3.2.3 Advantages and Problems

Spatial sound facilitates the exploitation of spatial auditory cues in order to segregate sounds emanating from different directions. It increases the coherence of auditory cues with those conveyed by cognition and other perceptual modalities. This way of sound processing is a key factor for improving the legibility and naturalness of a virtual scene because it enriches the immersive experience and creates more “sensual” interfaces. A 3D audio display can enhance multi-channel communication systems, because it separates messages from one another, thereby making it easier for the operator to focus on selected messages only.

However, today the costs for high-end acoustic rendering are still the biggest barrier to the widespread use of spatial audio. Especially exact environmental modeling for different auditory cues is extraordinarily expensive. Common problems in spatial sound generation that tend to reduce immersion are front-to-back reversals, intracranial heard sounds, and HRTF.

Spatial audio systems designed for the use with headphones may result in certain limitations such as inconvenience of wearing some sort of headset. With speakers, the spatial audio system must have knowledge of the listener’s position and orientation with respect to the speakers. And as auditory localization is still not fully understood, developers cannot make effective price/performance decisions in the design of spatial audio systems.

5.4 Haptic Feedback

Haptic perception relates to the perception of touch and motion. There are four kinds of sensory organs in the hairless skin of the human hand that mediate the sense of touch. These are the Meissner’s Corpuscles, Pacinian Corpuscles, Markel’s Disks, and Ruffini Endings. As shown in Table 2-2, the rate of adaptation of these receptors to a stimulus, location within the skin, mean receptive areas, spatial resolution, response frequency rate, and the frequency for maximum sensitivity are, at least partially, understood. The delay time of these receptors ranges from about 50 to 500 msec.

Table 2-2: Functional Features of Cutaneous Mechanoreceptors

Feature	Meissner Corpuscles	Pacinian Corpuscles	Merkel's Disks	Ruffini Endings
Rate of adaptation	Rapid	Rapid	Slow	Slow
Location	Superficial dermis	Dermis and subcutaneous	Basal epidermis	Dermis and subcutaneous
Mean receptive area	13 mm^2	101 mm^2	11 mm^2	59 mm^2
Spatial resolution	Poor	Very poor	Good	Fair
Sensory units	43%	13%	25%	19%
Response frequency range	10 – 200 Hz	70 – 1000 Hz	0.4 – 100 Hz	0.4 – 100 Hz
Min. threshold frequency	40 Hz	200 – 250 Hz	50 Hz	50 Hz
Sensitive to temperature	No	Yes	Yes	> 100 Hz
Spatial summation	Yes	No	No	Unknown
Temporal summation	Yes	No	No	Yes
Physical parameter sensed	Skin curvature, velocity, local shape, flutter, slip	Vibration, slip, acceleration	Skin curvature, local shape, pressure	Skin stretch, local force

It is important to notice that the thresholds of different receptors overlap. It is believed that the perceptual qualities of touch are determined by the combined inputs from different types of receptors. The receptors work in conjunction to create an operating range for the perception of vibration that extends from at least 0.04 to greater than 500 Hz (Bolanowski et al., 1988). In general, the thresholds for tactile sensations are reduced with increases in duration. Skin surface temperature can also affect the sensitivity of sensing tactile sensations.

These details provide some initial guidance for the design and evaluation of tactile display devices in such areas as stimulus size, duration and signal frequency. For example, Kontarinis and Howe (1995) note that the receptive areas and frequency response rates indicate that a single vibratory stimulus for a fingertip can be used to present vibration information for frequencies above 70 Hz, whereas an array-type display might be needed for the presentation of lower frequency vibrations.

Additional information is available when looking at a higher level that the receptors just discussed, that is, at the receptivity of the skin itself. The spatial resolution of the finger pad is about 0.15 mm, whereas the two-point limit is about 1 to 3 mm. Detection thresholds for features on a smooth glass plate have been cited as 2 mm high for a single dot, 0.06 mm high for a grating, and 0.85 mm for straight lines. Researchers have also looked at the ability to detect orientation. The threshold for detecting the direction of a straight line has been measured at 16.8 mm. When orientation is based on the position of two separate dots, the threshold was 8.7 mm when the dots were presented sequentially, and 13.1 mm when presented simultaneously. Reynier and Hayward (1993) discuss these findings and the results of additional work in this area. Data on the temporal acuity of the tactile sense is also reported by the authors, who note that two tactile stimuli (of 1 msec) must be separated by at least 5.5 msec in order to be perceived as separate. In general, increases in tactile stimulus duration can lower detection thresholds.

When we touch an object, typically both the tactile and kinesthetic are relevant to the experience (Heller, 1991). The object exerts a certain pressure on our hands which gives a sense of the weight and texture of the object. It also conveys a certain temperature to our hands and as we move our hands above the object, our kinesthetic sense gives information about the size of the object. Consequently, there are three basic forms distinguishable: The vibro-tactile, the temperature, and the kinesthetic sense.

The skin is sensitive to numerous forms of energy: Pressure, vibration, electric current, cold and warmth. In relation to display technology, by far the majority of the active tactile display is based on vibration. There are two major principles to generate vibration: Electrodes attached to the skin and mechanical vibration. Although both techniques are quite different, psycho-physical experiments show that the characteristics of the skin are the same for both. The human threshold for detection of vibration at about 28 dB (relative to 1 mm peak) for frequencies in the range 0.4 – 3 Hz, this decreases for frequencies in the range of 3 to about 250 Hz (at the rate of -5 dB/octave for the range 3 – 30 Hz, and at a rate of – 12 dB/octave for the range 30 – 250 Hz), for higher frequencies the threshold then increases (Shimoga, 1993b).

The perception of warmth and cold is another sensation modality. The human skin includes separate receptors for warmth and cold, hence different qualities of temperature can be coded primarily by the specific receptors activated. However, this specificity of neural activation is limited. Cold receptors respond only to low temperatures, but also to very high temperatures (above 45°C). Consequently, a very hot stimulus will activate both warm and cold receptors, which in turn evoke a hot sensation.

The literature also provides information on the just-noticeable-difference (JND) for changes of temperatures. Researchers Yarnitsky and Ochoa (1991) conducted experiments that looked at the JND of temperature change on the palm at the base of the thumb. They found that two different measurement methods gave different results, and the difference between results increased as the rate of temperature change increased. Using the more traditional measurement approach based on a method of levels, and starting at a baseline temperature of 32°C, the rate of temperature change (1.5, 4.2, and 6.7°C/sec) had no detectable effect on the JND for warming temperatures (~0.47°) or cooling temperatures (~0.2°). Subject reaction time was independent of the method used, and also independent of the rate of temperature change, although the reaction time for increases in warming (~0.7°) was significantly longer than the reaction time for increases in cooling (~0.5°). In reviewing work in this area, Zerkus et al. (1995) report on findings that the average human can feel a temperature change as little as 0.1°C over most of the body, though at the fingertip a sensitivity of 1°C is typical. He also states that the human comfort zone lies in the region of 13 to 46°C. LaMotte (1978) reports that the threshold of pain varies from 36 to 47°C depending on the locus on the body, stimulus duration, and base temperature.

Most of the research on kinesthetic perception has been focused on the perceptions of exerted force, limb position and limb movement. The kinesthetic system also uses the signals about force, position, and movement to derive information about other mechanical properties of objects in the environment, such as stiffness and viscosity (Jones, 1997). Understanding the perceptual resolution of the kinesthetic system for such object properties is very important to the design of haptic interfaces. Here is an overview of the results of studies on psychophysical scaling and JNDs for several parameters.

The subjective level of force increases with time (Stevens, 1970; Cain, 1971; Cain, 1973). The JND for force is about 7 % (Jones, 1989; Pang, 1991; Tan, 1995). The JND for stiffness (the change in force divided by the change in distance) is much higher. It is difficult to present a general value for the JND of stiffness, since the different studies revealed considerably different JNDs. The JNDs reported vary between 19 % and 99 % (Jones, 1990; Roland, 1977). The JND values for viscosity (a change in force divided by a change in velocity, expressed in Ns/m) depend on the reference values. For small values, the JNDs are high: 83 % at 2 Ns/m to 48 % at 16 Ns/m (Jones, 1993). For higher values, the JND is lower. Reported values range from 9.5 to 34 % (Jones, 1993; Jones, 1997; Beauregard, 1995; Beauregard, 1997). Finally, the reported JNDs for mass (defined as the ratio of applied force to achieved acceleration) are relative uniform across studies: 10 % is found for weights of 50 g, and a smaller JND for weights above 100 g (Ross, 1982; Brodie, 1984; Brodie, 1988; Ross, 1987; Darwood, 1991; Hellström, 2000). For very heavy weights, the JND decreases to 4 % (Carlson, 1977).

5.5 Olfactory Feedback

The olfactory system has been researched extensively and for different purposes. The entertainment industry has also experimented with synthetic smell production, in the form of accompanying smells to enhance the experience of films (Lefcowitz, 2001, Somerson, 2001). In the Aroma Rama and the Smell-o-vision systems, smells were released in cinema theatres in certain scenes of the film. In the John Waters film “Polyester” in 1981, the audiences were given “scratch and sniff” cards and asked to release smell at certain places during the film. These experimental systems were mainly novelties and not very successful, with reactions from the audiences reaching from allergic reactions to nausea.

Those systems were all manually controlled, and the scents were all pre-produced. With respect to the inclusion of smell in the user interface, it only becomes interesting when the production of smell can be computer controlled and can be produced based on a computerized descriptions of particular smells. Then it will be possible to include olfactory displays in computer systems. For smell to gain acceptance among audiences there are many more factors that need to be in place, such as natural smelling odors, non-allergenic smells, etc.

The main idea of how an olfactory display would work is that the user has a peripheral device for smell production. This device is connected to the computer, and controlled by the computer. Using codified descriptions of smell, the computer can signal the release of a particular smell. A specific smell is generated by mixing a set of primary odors, most likely in the form of oil-based fragrances (Bonsor, 2001; Cook, 1999).

6.0 REFERENCES

Adelstein, B.D., Lee, T.G. and Ellis, S.R. (2003). Head Tracking Latency in Virtual Environments: Psychophysics and a Model. In: Proceedings of the Human Factors and Ergonomics Society 47 Annual Meeting.

Alpern, M. and Campbell, F.W. (1963). The behavior of the pupil during dark adaptation. Journal of physiology, 65, 5-7.

Andersen, G.J. (1986). Perception of self-motion. Psychological Bulletin, 99, 52-65.

Andersen, G.J. and Braunstein, M.L. (1985). Induced self-motion in central vision. Journal of Experimental Psychology, Human Perception and Performance, 11, 122-132.

Balzarotti, G., Fiori, L. and Midollini, B. (1994). SIRPH: Steerable infrared picture on helmet mounted display. In: Proceedings AGARD Mission Systems Panel Symposium on Low-Level and Nap-of-the-Earth (N.O.E.) Night Operations, Rome, 25-27 October.

Barnes, G.R. (1980). Vestibular control of oculomotor and postural mechanisms. Clinical Physical and Physiological Measures. 1, 3-40.

Beauregard, G.L. and Srinivasan, M.A. (1997). Sensorimotor interactions in the haptic perception of virtual objects (RLE Technical Report No. 607), Cambridge, MA: Massachusetts Institute of Electronics".

Beauregard, G.L., Tan, M.A. and Durlach, N.I. (1995). The manual resolution of viscosity and mass, Proceedings of the ASME Dynamic Systems and Control Division, vol. 57(2), pp. 657-662.

Benson, A.J. (1987). Aetiological factors in simulator sickness. In: Motion Cues in Flight Simulation and Simulator Induced Sickness. AGARD Conference Proceedings No. CP433.

Benson, A.J. and Barnes, G.R. (1978). Vision during angular oscillation: the dynamic interaction of visual and vestibular mechanisms. Aviation, Space and Environmental Medicine, 49, 340-345.

Boff, K.R. and Lincoln, J.E. (1988). Engineering Data Compendium: Human Perception and Performance, AAMRL, Wright Patterson AFB, Ohio.

Bolanowski, S.J., Gescheider, G.A., Verrillo, R.T. and Checkosky, C.M. (1988). Four Channels Mediate the Mechanical Aspects of Touch. Journal of the Acoustical Society of America, 84(5), pp. 1680-1694.

Bonsor, K. How Internet Odors Will Work, Marshall Brain's HowStuffWorks, http://www.howstuffworks.com/internet-odor.htm?printable=1, (2001, 25 September).

Brodie, E.E. and Ross, H.E. (1984). Sensorimotor mechanisms in weight discrimination, Perception & Psychophysics, 36(5), pp. 477-481.

Brodie, E.E. (1988). Sex and hand-preference effects in the simultaneous and consecutive discrimination of lifted weight, Perception & Psychophysics, 43(4), 326-330.

Bryson, S. and Fisher, S.S. (1990). Defining, modelling and measuring system lag in virtual environments, Stereoscopic Displays and Applications, SPIE Vol. 1256, 98-109.

Burg, A. (1968). Lateral visual field as related to age and sex, Journal of applied Psychology, 52, 10-15.

Cain, W.S. (1973). Nature of perceived effort and fatigue: roles of strength and blood flow in muscle contractions, Journal of Motor Behavior, 5, pp. 33-47.

Cain, W.S. and Stevens, J.C. (1971). Effort in sustained and phasic handgrip contractions, American Journal of Psychology, 84, pp. 52-65.

Carlson, N.K., Drury, C.G. and Webber, J.A. (1977). Discriminability of large weights, Ergonomics, 20(1), pp. 87-90.

Cheung, B.S.K., Howard, I.P. and Money, K.E. (1991). Visually-induced sickness in normal and bilaterally labyrinthine-defective subjects. Aviation, Space and Environmental Medicine, 62, 527-531.

Cobb, S., Nichols, S. and Wilson, J.R. (1995) Health and safety implications of virtual reality: in search of an experimental methodology. In: Proc. FIVEí95 (Framework for Immersive Virtual Environments), University of London, December.

Cook, M.J., Angus, C., Campbell, C. and Cranner, C. (1999). Speech interfaces in real-time control systems, People in control. London: Institution of electrical engineers, pp. 428-433.

Crowley, J.S. (1987). Simulator sickness: a problem for army aviation. Aviation, Space and Environmental Medicine, 58, 355-357.

Darwood, J.J., Repperger, D.W. and Goodyear, C.D. (1991). Mass discrimination under Gz acceleration, Aviation, Space, and Environmental Medicine, 62(4), pp. 319-324.

Daunton, N.G., Fox, R.A. and Crampton, G.H. (1984). Susceptibility of cat and squirrel monkey to motion sickness induced by visual stimulation: correlation with susceptibility to vestibular stimulation. In: Motion Sickness: Mechanisms, Prediction, Prevention and Treatment, AGARD Conference Proceedings no. 372.

Downton, A. (1991). Engineering the Human-Computer Interface, McGraw-Hill, London, 1991.

Edgar, G.K. and Bex, P.J. (1995). Vision and Displays. In: Carr, K. and England, R., Eds. Simulated and Virtual Realities – Elements of Perception. London: Taylor and Francis.

Ellis, S.R. (1995). Virtual environments and virtual instruments. In: Carr, K. and England, R., Eds. Simulated and Virtual Realities – Elements of Perception. London: Taylor and Francis.

Ellis, S.R., Mania, K., Adelstein, B.D. and Hill, M.I. (2004). Generalizability of latency detection in a variety of virtual environments. In: Proc. of the 48th Annual Meeting Human Factors and Ergonomics Society.

Ellis, S.R., Young, M.J., Adelstein, B.D. and Ehrlich, S.M. (1999). Discrimination of changes in latency during head movement. In: Proceedings, Computer Human Interfaces, pp. 1129-1133.

Ellis, S.R., Young, M.J., Ehrlich, S.M. and Adelstein, B.D. (1999). Discrimination of changes of rendering latency during voluntary hand movements. In: Proceedings, 43rd annual meeting HFES, pp. 1182-1186.

ETSI EG 202 048, V1.1.1, (2002). Human Factors (HF); Guidelines on the multimodality of icons, symbols and pictograms.

Frank, L.H., Casali, J.G. and Wierville, W.W. (1988). Effects of visual display and motion system delays on operator performance and uneasiness in a driving simulator. Human Factors, 30, 201-217.

Frank, L.H., Kennedy, R.S., Kellog, R.S. and McCauley, M.E. (1983). Simulator sickness: a reaction to a transformed perceptual world 1. scope of the problem, In: Proceedings of the Second Symposium of Aviation Psychology, Ohio State University, Columbus OH, 25-28 April.

Fukuda, T., Kanada, K. and Saito, S. (1990). An Ergonomic Evaluation of Lens accommodation related to visual circumstances, Ergonomics, 33,(6), 811-83 1.

Griffin, M.J. (1990). Handbook of Human Vibration. London: Academic Press.

Grunwald, A.J., Kohn, S. and Merhav, S.J. (1991). Visual field information in nap-of-the-earth flight by teleoperated head-mounted displays. Large-screen-projection, Avionic and Head-mounted displays, SPIE Vol. 1456, 132-153.

Held, R. and Durlach, N. (1991). Telepresence, time delay and adaptation. In: Ellis, S.R. Ed. Pictorial communication in virtual and real environments. New York: Taylor and Francis.

Heller, M.A. (1991). Introduction, In: M.A. Heller and W. Schiff (Eds.), The psychology of touch, Hillsdale, NJ: Lawrence Erlbaum.

Hellström, A. (2000). Sensation weighting in comparison and discrimination of heaviness, Journal of Experimental Psychology: Human Perception and Performance, 26(1), pp. 6-17.

Hettinger, L.J. and Riccio, G.E. (1992). Visually induced motion sickness in virtual environments. Presence, 1, 306-310.

Howard, I.P. (1986). The perception of posture, self-motion and the visual vertical. In: Boff, K.R., Kaufman, L. and Thomas, J.P. Eds. Handbook of Human Perception and Performance, Vol. 1, Sensory Processes and Perception. New York: Wiley.

Jones, L., Hunter, I. and Lafontaine, S. (1997). Viscosity discrimination: a comparison of an adaptive two-alternative forced-choice and an adjustment procedure, Perception, 26, pp. 1571-1578.

Jones, L.A. and Hunter, I.W. (1990). A perceptual analysis of stiffness. Experimental Brain Research, 79, pp. 150-156.

Jones, L.A. and Hunter, I.W. (1993). A perceptual analysis of viscosity, Experimental Brain Research, 94, pp. 343-351.

Jones, L.A. (1989). Matching forces: constant errors and differential thresholds, Perception, 18, pp. 681-687.

Kelly, D.H. (1969). Flickering patterns a lateral inhibition. Journal of the Optical Society of America, 59, 1361-1370.

Kennedy, R.S. (1975). Motion sickness questionnaire and field independence scores as predictors of success in naval aviation training. Aviation, Space and Environmental Medicine, 46, 1349-1352.

Kennedy, R.S., Fowlkes, J.E. and Lilienthal, M.G. (1993). Postural and performance changes following exposures to flight simulators. Aviation, Space and Environmental Medicine, 64, 912-920.

Kennedy, R.S., Fowlkes, J.E., Berbaum, K.S. and Lilienthal, M.G. (1992). Use of a motion sickness questionnaire for prediction of simulator sickness. Aviation, Space and Environmental Medicine, 63, 588-593.

Kennedy, R.S., Lane, N.E., Berbaum, K.S. and Lilienthal, M.G. (1989). Simulator sickness in US Navy flight simulators. Aviation, Space and Environmental Medicine, 60, 10-16.

Kennedy, R.S., Lanham, D.S., Drexler, J.M., Massey, C.J. and Lilienthal, M.G. (1995). Cybersickness in several flight simulators and VR devices: a comparison of incidences, symptom profiles, measurement techniques and suggestions for research. In: Proc. FIVEí95 (Framework for Immersive Virtual Environments), University of London, December.

Kennedy, R.S., Lilienthal, M.G., Baltzley, D.R., Allgood, G.O. and Gower, D.W. (1987). Consistency across measures of simulator sickness: implications for a biocybernetic safety reporting device. In: Proc. SAFE Association 25th Annual Symposium, Las Vegas, 16-19 November, A89-1052-01-54, 95-100.

Kolasinski, E.M. (1995). Simulator sickness in virtual environments. Technical Report 1027, U.S. Army Research Institute for the Behavioural and Social Sciences, Alexandria, VA.

Kontarinis, D.A. and Howe, R.D. (1995). Tactile Display of Vibratory Information in Teleoperation and Virtual Environments. To appear in Presence.

Krendel, E.S. and McRuer, D.T. (1965). Human pilot dynamics in compensatory systems. AFFDL-TR-65-15. Air Force Flight Dynamics Lab., Wright-Patterson Air Force Base, Ohio.

Kruger, H. (1980). Ophthalmologic aspects of work with display workstations. Proceeding, 5 of the International Workshop, Milan, March.

Lackner, J. (1990). Human orientation, adaptation and movement control. In: Motion Sickness, Visual Displays and Armoured Vehicle Design. Ballistic Research Laboratory, APG MD, Washington DC, April.

LaMotte, R.H. (1978). Intensive and Temporal Determinants of Thermal Pain. In: D.R. Kenshalo (Ed.), Sensory Functions of the Skin of Humans. Plenum Press: NY.

Lane, N.E. and Kennedy, R.S. (1988). A new method for quantifying simulator sickness: development and application of the simulator sickness questionnaire (SSQ). Technical Report No. EOTR 88-7, Essex Corporation.

Lassig, P.H. and Molski, J. (1995). Simulation of the eye-head system combined with head-mounted display – contributing to the problem of self-motion and cybersickness. In: Proc. FIVEí95 (Framework for Immersive Virtual Environments), University of London, December.

Lawther, A. and Griffin, M.J. (1986). The motion of a ship at sea and the consequent motion sickness amongst passengers. Ergonomics, 29, 535-552.

Lefcowitz, E. (2001). Retrofuture – Smell-o-vision!, [Web document], available: http://retrofuture.com/smell-o-vision.html, 2001.

Lestienne, F., Soechting, J. and Berthoz, A. (1977). Postural re-adjustments induced by linear motion of visual scenes. Experimental Brain Research, 28, 363-384.

Lewis, C.H. (1984). The development of an adaptive image stabilisation system for head-mounted displays. In: Proc. United Kingdom Informal Group Meeting on Human Response to Vibration, Heriot-Watt University, Edinburgh.

Lewis, C.H. and Griffin, M.J. (1996). Temporal aspects of head-coupled system performance in the helicopter cockpit. ISVR Contract Report No 96/17, University of Southampton.

Lewis, C.H., Higgins, A.F., Wells, M.J. and Griffin, M.J. (1987). Stabilisation of head-coupled systems for rotational head motion. Final report on contract no. T33615-82-C-0513, MacAuley-Brown inc.

Liebowitz, H.W., Rodemer, C.S. and Dichgans, J. (1979). The independence of dynamic spatial orientation from luminance and refractive error. Perception and Psychophysics, 25, 75-79.

Liu, A., Tharp, G., French, L., Lai, S. and Stark, L. (1993). Some of what one needs to know about using head-mounted displays to improve teleoperator performance. IEEE Transactions on Robotics and Automation, 9, 638-648.

Liu, A., Tharp, G., Hirose, M. and Stark, L. (1991). Visual factors affecting human operator performance with a head-mounted display. SAE Technical Paper 911389.

Mania, K., Adelstein, B., Ellis, S.R. and Hill, M. (2004). Perceptual Sensitivity to Head Tracking Latency in Virtual Environments with Varying Degrees of Scene Complexity. In: Proc. ACm Siggraph Symposium on Applied Perception in Graphics and Visualization, 39-47, ACM Press.

McCauley, M.E. and Sharkey, T.J. (1992). Cybersickness: perception of self-motion in virtual environments. Presence, 1, 311-318.

McRuer, D.T. (1973). Development of pilot-in-the loop analysis. Journal of Aircraft, 10, 514-518.

Meyer, K., Applewhite, H.L. and Biocca, F.A. (1992). A survey of position trackers. Presence, 1, 173-200.

Monk, A. and Gaines, B.R. (1984). Fundamentals of Human-Computer Interaction, Academic Press, London, 1984.

Mon-Williams, M., Wann, J.P. and Rushton, S. (1993). Binocular vision in a virtual world: visual deficits following the wearing of a head-mounted display. Ophthalmic and Physiological Optics, 13, 387-391.

Oman, C.M. (1982). A heuristic mathematical model for the dynamics of sensory conflict and motion sickness. Acta Otolayngologica, Suppl. 392, 21-44.

Owsley, C., Sekuler, R. and Siemsen, D. (1983). Contrast sensitivity throughout adulthood. Vision Research, 23, 689-699.

Pang, X.D., Tan, H.Z. and Durlach, N.I. (1991). Manual discrimination of force using active finger motion, Perception & Psychophysics, 49, pp. 531-540.

Parker, D.E. and Harm, D.L. (1992). Mental rotation: a key to mitigation of motion sickness in the virtual environment? Presence, 1, 329-333.

Pausch, R., Crea, T. and Conway, M. (1992). A literature survey for virtual environments: military flight simulator visual systems and simulator sickness. Presence, 1, 344-363.

Peli, E. (1990). Visual Issues in the use of a head-mounted monocular display, Optical Engineering, 29, No. 8, 883-893.

Rabin, J. and Wiley, R. (1995). Dynamic visual performance comparison between helmet-mounted CRTs and LCDs. Journal of the Society for Information Display, 3, 97-100.

Reason, J.T. and Brand, J.J. (1975). Motion sickness. London: Academic Press.

Regan, E.C. (1995). An investigation into nausea and other side-effects of head-coupled immersive virtual reality. Virtual Reality, 1(1), 17-32.

Regan, E.C. and Price, K.R. (1993a). Subjective views on an immersion virtual reality system and its peripherals. Working Paper 16/93, Army Personnel Research Establishment, Farnborough.

Regan, E.C. and Price, K.R. (1993b). Some side-effects of immersion virtual reality: an investigation into the relationship between inter-puillary distance and ocular related problems. Report 93R023, Army Personnel Research Establishment, Farnborough.

Regan, E.C. and Price, K.R. (1993c). Some side-effects of immersion virtual reality: the effects of increasing head movements, of rapid interaction, and of seated subjects. Report 93R022, Army Personnel Research Establishment, Farnborough.

Regan, E.C. and Price, K.R. (1994). The frequency of occurrence and severity of side-effects of immersion virtual reality. Aviation, Space and Environmental Medicine, 65, 527-530.

Regan, E.C. and Ramsey, A.D. (1994). Some side-effects of immersion virtual reality: an investigation into the effectiveness of hycosine hydrobromide in reducing nausea induced during immersion. Report 94R025, Army Personnel Research Establishment, Farnborough.

Reynier, F. and Hayward, V. (1993). Summary of the Kinesthetic and Tactile Function of the Human Upper Extremities. McGill Research Center for Intelligent Machines, McGill University, Montreal, Canada.

Richard, P., Birbent, G., Coiffet, P., Burdea, G., Gomez, D. and Langrana, N. (1996). Effect of frame rate and force feedback on virtual object manipulation. Presence, 5, 95-108.

Roland, P.E. and Ladegaard-Pedersen, H. (1977). A quantitative analysis of sensations of tension and of kinaesthesia in man, Brain, 100, pp. 671-692.

Roscoe, S.N. (1985). Bigness is in the eye of the beholder. Human Factors, 27(6), 615-636.

Rosner M., and Belkin, M. (1989). Video display units and visual function. Survey of Ophthalmology, 33(6), 515-522.

Ross, H.E. and Brodie, E.E. (1987). Weber fractions for weight and mass as a function of stimulus intensity, Quarterly Journal of Experimental Psychology 39A, pp. 77-88.

Ross, H.E. and Reschke, M.F. (1982). Mass estimation and discrimination during brief periods of zero gravity, Perception & Psychophysics, 31, pp. 429-436.

Rushton, S., Mon-Williams, M. and Wann, J.P. (1994). Binocular vision in a bi-ocular world: new generation of head-mounted displays avoid causing visual deficit. Displays, 15, 255-260.

Saladin, J.J. (1988). Interpretation of divergent oculomotor imbalance though control system analysis. American Journal of Ophthalmology Physiology, 65, 439-447.

Shimoga, K.B. (1993b). A Survey of Perceptual Feedback Issues in Dextrous Telemanipulation: Part II. Finger Touch Feedback. In: Proc. IEEE Annual Virtual Reality International Symposium, Seattle, Washington, September 18-22, pp. 271-279.

So, R.H.Y. (1994). An investigation of the effects of lags on motion sickness with a head-coupled visual display. In: Proc. United Kingdom Informal Group Meeting on Human Response to Vibration, Institute of Naval Medicine, Alverstoke.

So, R.H.Y. (1995). Optimising the response of head-coupled systems to dynamic head movements. PhD thesis, Institute of Sound and Vibration Research, University of Southampton.

So, R.H.Y. and Griffin, M.J. (1992). Compensating lags in head-coupled displays using head position. Journal of Aircraft, 29, 1064-1068.

So, R.H.Y. and Griffin, M.J. (1995a). Effects of lags on human-operator transfer functions with head-coupled systems. Aviation, Space and Environmental Medicine, 66, 550-556.

So, R.H.Y. and Griffin, M.J. (1995b). Modelling the effects of display lags and reticle lags on head tracking performance. ISVR Contract Report No. CR 95/25, University of Southampton.

So, R.H.Y. and Griffin, M.J. (1995c). Head-coupled virtual environment with display lag. In: Carr, K. and England, R., Eds. Simulated and Virtual Realities – Elements of Perception. London: Taylor and Francis.

So, R.H.Y. and Griffin, M.J. (1996). Use of phase-lead filters to compensate lags in head-coupled visual displays. IEEE Transactions on Systems, Man and Cybernetics, in press.

Soderberg, I., Calissendorff, B., Elofsson, S., Knave, B. and Nyman, K.G. (1993). Investigation of visual strain experience by microscope operators at an electronics plant. Applied Ergonomics, 14, (4), 297-305.

Somerson, P. (2001, March 2001). Click and lick, ZDNet, available: http://www.zdnet.com/smartbusinessmag/stories/all/0,6605,2677285,00.html, 2001.

Stern, R.M., Hu, S., Anderson, R.B., Leibowitz, H.W. and Koch, K.L (1990). The effects of fixation and restricted visual field on vection-induced motion sickness. Aviation, Space and Environmental Medicine, 61, 712-715.

Stern, R.M., Hu, S., LeBlanc, R. and Koch, K.L (1993). Chinese hyper-susceptibility to vection-induced motion sickness. Aviation, Space and Environmental Medicine, 64, 827-830.

Stevens, J.C. and Cain, W.S. (1970). Effort in isometric muscular contractions related to force level and duration, Perception & Psychophysics, 8, pp. 240-244.

Sutcliffe, A. (1989). Human-Computer Interface Design, Springer-Verlag, New York, 1989.

Tan, H.Z., Durlach, N.I., Beauregard, G.L. and Srinivasan, M.A. (1995). Manual discrimination of compliance using active pinch grasp: the roles of force and work cues, Perception & Psychophysics, 57, pp. 495-510.

Tharp, G., Liu, A., French, L., Lai, S. and Stark, L. (1992). Timing considerations of head-mounted display performance. Human Vision, Visual Processing and Digital Display III, SPIE Vol. 1666, 570-576.

Verona, R.W. (1980). Head aiming/tracking accuracy in a helicopter environment, AGARD Conference Proceedings, 255 Operational Helicopter Aviation Medicine, NATO.

Warren, R. (1976). The perception of ergomotion. Journal of Experimental Psychology: Human Perception and Performance, 2, 448-456.

Wells, M.J. and Haas, M.W. (1990). Head movements during simulated air-to-air engagements. Helmet Mounted Displays II, SPIE Vol. 1290, 246-257.

Wells, M.J. and Venturino, M. (1990). Performance and head movements using a head-mounted display with different fields of view. Optical Engineering, 29, 870-877.

Wioka, M.W. (1995). Lag in multi-processor virtual reality. Presence, 4, 50-63.

Woodruff, R.R., Hubbard, D.C. and Shaw, A. (1986). Comparison of helmet-mounted visual displays for flight simulation. Displays, 7, 179-185.

Yarnitsky, D. and Ochoa, J.L. (1991). Warm and Cold Specific Somatosensory Systems. Brain, 114, pp. 1819-1826.

Young, L.R. (1970). On visual vestibular interaction. In: Proc. 5th Symposium on the role of the Vestibular Organs in Space Exploration, NASA SP-314, 205-210.

Young, L.R. and Oman, C.M. (1969). A model for vestibular adaptation to horizontal rotation. Aerospace Medicine, 40, 1076-1080.

Young, L.R., Dichgans, J., Murphy, R. and Brandt, T. (1973). Interaction of optokinetic and vestibular stimuli in motion perception. Acta Otolaryngologica, 76, 24-31.

Zacharias, G.L. and Young, L.R. (1981). Influence of combined visual and vestibular cues on human perception and control of horizontal rotation. Experimental Brain Research, 41, 159-171.

Zacharias, G.L., Caglayan, A.K. and Sinacori, J.B. (1985). A visual cueing model for terrain following applications. Journal of Guidance, Control and Dynamics, 8, 201-207.

Zerkus, M., Becker, B., Ward, J. and. Halvorsen, L. (1994). Thermal Feedback in Virtual Reality and Telerobotic Systems. Technical Report N95-15987. NASA Scientific and Technical Information Office, NASA.

Presence

Erik den Dekker
TNO Defence, Security and Safety
Business Unit Human Factors
P.O. Box 23
3769ZG Soesterberg
THE NETHERLANDS

Nico Delleman
TNO Defence, Security and Safety
Business Unit Human Factors
P.O. Box 23
3769ZG Soesterberg
THE NETHERLANDS

Université Paris Descartes
UFR Biomédicale, EA4070 Ergonomie – Comportement et Interactions
Laboratoire d'Anthropologie Appliquée
45 rue des Saints-Pères
75270 Paris Cedex 06
FRANCE

1.0 INTRODUCTION

If one wants to turn virtual reality (VR) technology into an effective tool, it is vital to obtain a basic understanding of the key elements for success. A concept that has received a lot of attention from VR researchers is presence. Presence may be loosely described as the feeling of 'being there' in the virtual world instead of just experiencing stimuli from displays [3] (for a clear description of the difference between presence and immersion, refer to Slater [10]; immersion is defined as the extent to which the actual system delivers a surrounding environment, one which shuts out sensations from the 'real world', which accommodates many sensory modalities, has rich representational capability, and so on; the term system immersion may be used for this, as opposed to an immersive response such as presence). Letting a user feel presence in a virtual world might be a goal in itself, but it has often been suggested that presence is a necessary element to reach higher goals. For instance, some researchers claimed that by developing an understanding of the factors that drive presence, their interrelationships, and how they relate to human performance and after-effects, a set of design principles could be specified that should lead to enhanced human performance in VR systems [15]. This chapter presents an overview of research on presence, i.e., theories, definitions, and multi-dimensionality (Section 2), effects (Section 3), measurement (Section 4), and determinants (Section 5). Section 6 contains some concluding remarks.

2.0 THEORIES, DEFINITIONS, AND MULTI-DIMENSIONALITY

Several theories have been proposed on the nature of presence in VR. Schuemie et al. give an extensive review of different conceptualizations, definitions and theories with regard to presence [8]. Here, the most influential ones will be discussed. An important theory on presence is based on a definition of Lombard and Ditton, who define presence as the 'perceptual illusion of non-mediation', i.e., the extent to which a person fails to perceive or acknowledge the existence of a medium during a technologically mediated experience [5]. Although a user knows that an experience is mediated and can distinguish between mediated and direct stimuli, at some level, the illusion of non-mediation can be perceived. Another, somewhat similar, commonly used definition is that a user feels present when he or she is engaged to the point of 'suspending disbelief' in what he or she is experiencing [9]. Again another viewpoint emphasizes the exclusiveness of presence by arguing that an individual can only feel present in one environment at a time. Presence may oscillate between the real, virtual and internal (or imaginable) world. In this case, the level of presence experienced in a VR simulation depends on the amount of time being present in the virtual world. Witmer and Singer have related presence in part to the allocation of attentional resources [19]. It is argued that by focusing more attention to an environment a user will get more involved and as a result get a higher sense of presence. In a practical approach to the exclusiveness theory, Slater considered presence as a perceptual mechanism for selection between alternative hypotheses based on Gestalt psychology [11]. When engaged in a VR simulation a user may be receiving competing signals from several different environments. Moment by moment, a selection mechanism organizes the streams of sensory signals into an environmental gestalt. Sensory data relevant to other environmental gestalts are relegated to the background. Given these competing signals, at any moment action is chosen on the basis of selection between alternative hypotheses, alternative interpretations. The presence selection mechanism is an answer to a fundamental question: "Where am I?" Slater claims that the environment hypothesis is continually reverified or else a break in presence (BIP) occurs. A BIP is the moment of switch between responding to signals with source in the virtual environment to those with source in the physical or internal environment. Finally, several researchers made an effort to explain presence as part of bigger theories that describe how human beings make sense of the world around them [8]. They acknowledge the role of a human being as interpreter, making a mental model that estimates reality. It is theorized that humans conceptualize their environment in terms of the actions that can be taken on it. In predicting the outcome of actions, humans can suppress certain contributions, thereby creating a self-consistent representation despite conflicting features. Actions that are successfully supported by the environment lead to presence. Actions are supported when they are lawful, i.e., similar to the real world in which our perceptual system evolved.

There is a growing awareness that presence is a multi-dimensional concept, i.e., there are several different types of presence [3, 4, 5, 16]. In this sense, presence might be compared to a concept like emotion, of which existence of different dimensions is well established (e.g., instead of having one scale to express the 'amount of emotion' of a person, a distinction is made between anger, happiness, surprise, etc.). So, efforts have been put in trying to discriminate between the different dimensions of presence that exist. Several dimensions have been proposed, but are likely to be overlapping or non-orthogonal to some extent. First of all, a distinction must be made between physical presence and social presence as mentioned by IJsselsteijn et al. [3] and IJsselsteijn and Riva [4] basing their work on the findings of Lombard and Ditton [5]. Physical presence refers to the sense of being physically located in mediated space, whereas social presence refers to the feeling of being together, i.e., of social interaction with a virtual or remotely located communication partner. A clear distinction between these two is that the ability to communicate is essential for social presence, but not for physical presence. Co-presence can be defined as being together in a shared space, combining characteristics of physical and social presence [3, 4]. Schubert et al. [7] and Regenbrecht and Schubert [6] showed that items assessing subjective experiences of physical presence can be split into three different components:

1) Spatial presence, i.e., the human mind understands the relationship between the body and the environment in terms of distance, direction, size, etc.
2) Involvement, i.e., the attention distribution between internal, virtual and real world.
3) Realism, i.e., whether the virtual world seems as "real" as the real world.

The multi-dimensionality of presence is outlined in Figure 3-1.

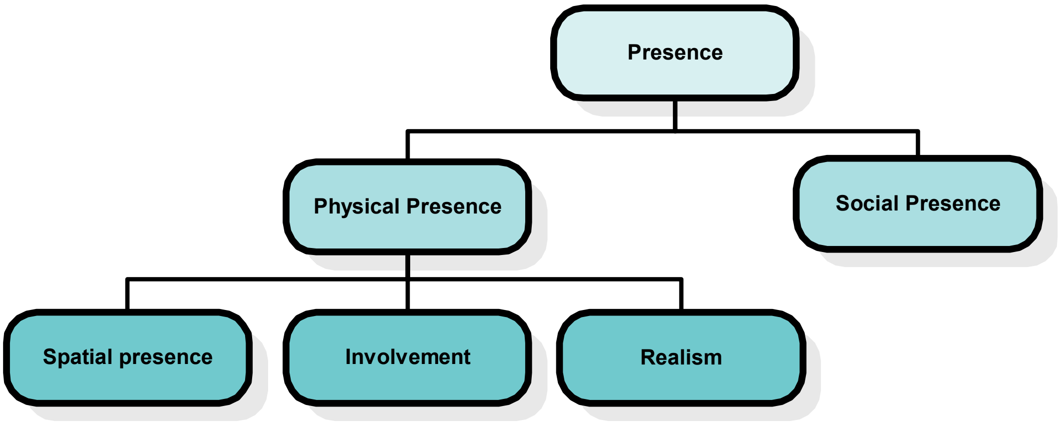


Figure 3-1: Dimensions of Presence.

3.0 EFFECTS OF PRESENCE

For many VR experiences, full presence is not necessarily required for the application to be useful [9]. For instance, when an application's goal is to explore a database, a high level of presence may not be necessary, possible, or even desirable. The degree to which presence is desirable for a particular experience varies based on the goals of the experience. If the experience's success is based on how engrossed the participant becomes, presence plays a key role in the experience's fulfilment.

It may be argued that for most entertainment applications, creating a compelling sense of presence is a desirable goal [3]. However, for other application areas this may be less straightforward – e.g., with certain tasks, changes that may diminish presence may in fact enhance performance. Regenbrecht and Schubert argue that training relies on a transfer of learning from the virtual to the real world, and think that it is facilitated by a high realness (a component of presence) of the virtual world [6]. Witmer and Singer suggest that many of the same factors that increase presence are also instrumental in promoting learning and performance [19]. Thus, they believe virtual environments to be used for entertainment or for training should be designed to induce high levels of presence for most tasks.

Schuemie et al. in their review describe the following four potential effects of presence [8]:

1) The subjective sensation of presence; this can be a goal in itself for certain applications.
2) Task performance; although there is a strong belief that there is a positive correlation between presence and task performance, it remains controversial whether presence actually contributes to better task performance [8, 15]. It is very likely that this relationship depends on the characteristics of the task that the user has to perform. In the case where the task was to train for a real world situation such a positive correlation is deemed likely [5].
3) Physical responses and emotions; these are supported by more empirical evidence. The problem remains however that only weak evidence for certain relationships have been found and the question of causality has often not yet been addressed.
4) Simulator sickness; the relationship between presence and simulator sickness has been investigated, with contradicting results presented in different papers [8, 15, 19].

Schuemie et al. [8] and IJsselsteijn et al. [3] conclude that much uncertainty remains as to the usefulness of presence. Sherman and Craig suggest that presence assists in creating a sense of faith in the user that the world is consistent beyond what is immediately presented, e.g., that the world includes events that are presumed to have occurred and places that are presumed to exist [9].

4.0 MEASUREMENT OF PRESENCE

A good measurement technique of presence assists developers to enhance VR systems that have the intention to invoke a sense of presence in its users. It may also aid human factors specialists in investigating relationships of presence and other measurable entities, like for instance task performance. Moreover, by addressing the issue of how to measure presence, greater understanding of the factors that drive this phenomenon may result [3, 15].

A measurement technique for presence should satisfy the basic rules that apply to all measuring techniques. It should be repeatable, reliable, robust, and sensibly correlated with measurable characteristics of a VR system. Due to its psychological aspects, most measurement techniques of presence are derived from measuring methodologies used in psychological research [15]. For instance, rating scales or equivalence classes (e.g., “On a 1 – 10 scale, rate the amount of presence this VR system produces for you”), method of paired comparisons (e.g., “Which of these two VR systems produces the greater amount of presence for you?”), or cross-modality matching (e.g., “Make this light as bright as the strength of the presence you experienced in this VR system”).

One of the biggest problems with current measurement techniques for presence is that a single, accepted paradigm for the assessment of presence is lacking. As a consequence, a variety of presence measures has been proposed [3]. These presence measures are based on different assumptions on the nature of presence. Without consensus on one such paradigm, it is difficult to compare presence among research groups to assess the relative efficacy of different VR systems.

An important development in measuring techniques for presence is the monitoring of BIPs [11, 12]. Because of its definition there is a straightforward relation between BIPs and presence. Slater has developed a stochastic model that estimates the tendency that a user is in a presence state during an experience based on the reported BIPs [11]. Results of experiments using this method are promising.

Despite the lack of consensus, the reported measuring techniques of presence can be classified into two distinct categories: subjective measurement and objective measurement. It seems likely that both subjective and objective tools will be required for the comprehensive measurement and understanding of the sense of presence [15].

4.1 Subjective Measurement

All subjective measurement methods are characterized by the fact that the user him/herself has to report on the amount of presence that is experienced. Some researchers argued that ‘subjective report is the essential basic measurement’ of presence [3]. Stanney et al. say that subjective assessment of VR systems is thought to be particularly useful in early evaluation and exploration phases [15].

There are several techniques that require a user him/herself to report on the amount of presence. The majority of all studies measure presence through post-test questionnaires and rating scales, which have the advantage that they do not disrupt the experience and are easy to administer [3]. Several questionnaires have been published, each of them created from a different point of view, and with different applicability. The most influential questionnaires, referred to as SUS6, PQ, IPQ and ITC-SOPI, are all reviewed by Schuemie et al. [8].

Although usage of post-test questionnaires is the most common means of presence measurement, there are disadvantages associated with this method. For instance, Slater [11] states that presence in VR systems reported via questionnaires is usually high, especially relative to the paucity of the virtual world compared to the real world. He refers to an experiment which reported that presence in a virtual office was the same as for a real office [18]. Furthermore, ambiguous and erroneous interpretations of questionnaire questions will not all be exorcised by pilot experiments [17]. Questions require great care. Oral debriefing of users resolves many ambiguities.

Another limitation of discrete post-test subjective ratings is that they do not provide any measure of temporal variations in presence [3]. Such variations are likely to occur through changes in the stimulus (i.e., both in VR system hardware components and the virtual world) or the participant (e.g., prolonged exposure may have an impact on presence through perceptual-motor adaptation, boredom, fatigue, etc.) and would be lost in an overall post-test presence rating. In addition, discrete subjective measures are potentially subject to inaccurate recall and anchoring effects. To overcome these limitations, the method of continuous presence assessment has been applied. This method requires users to make an on-line judgment of presence using a hand-held slider. When applied to non-interactive stereoscopic media, it was found that presence ratings were subject to considerable temporal variation depending on the extent of sensory information present in the stimulus material.

Measuring BIPs is a special form of continuous presence measurement. It is special in the sense that it doesn’t require a user to directly report presence, but instead a user reports the transition from being present in the virtual world to being present in the real or mental world. Having to report such a transition doesn’t break the experience of presence itself since it has already been broken [11].

IJsselsteijn et al. [3] comment on a potential criticism that may be raised against continuous presence assessment, i.e., that users are in fact required to divide their attention between both the display and the on-line rating they are asked to provide, thereby possibly disrupting the presence experience. However, observers are unlikely to report on a belief of being in the displayed scene, since they are usually well aware of actually being in a laboratory. Rather, they report on a sensation of being there that approximates what they would experience if they were really there. This does not necessarily conflict with providing a continuous rating – especially given the fact that the measurement device requires very little attention or effort to operate. The need to consciously reflect on the amount of presence experienced in order to give a reliable rating is of course intrinsic to both continuous and discrete subjective measures of presence.

Another interesting psychophysical method for measuring presence to be mentioned is the method of paired comparisons, which, in the context of VR systems, has sometimes been referred to as the ‘virtual reality Turing test [3]. Here, a user is asked to distinguish between a virtual and a real scene. It is suggested that the probability that a human perceives that he or she is physically present in a given virtual environment can serve as a subjective measure of presence. Since users are unlikely to confuse the real environment with the virtual one, a number of perceptual limitations are imposed on the user during the test (e.g., both the real and virtual environments have to be viewed with a limited visual field, with reduced contrast or colour, not using sound, etc.). Taking the amount of degradation of the real scene that is necessary to make it indistinguishable from the virtual one has been suggested as a measure of presence. A potential criticism that can be raised against this methodology is that it becomes a test of participants’ ability to discriminate between two degraded images rather than a measure of presence.

While subjective measurement methods are effective means of assessing presence, it is important to note that such methods should be used judiciously due to inconsistencies across different raters or rating situations. Different users may apply scales differently and thus a means of calibrating individual differences may be required to control for bias [1, 3]. Relatedly, it has been shown that subjective ratings of presence can be biased by previous experience, e.g., rating a different attribute in a previous session [3]. On top of that subjective methods require users to have a fair understanding of what is meant by the term ‘presence’ and subjective measures can only provide conscious, voluntary, and supported responses from users while it has been shown that sometimes a user’s behavior contradicts their own assessment [3, 9, 11].

4.2 Objective Measurement

The problems that are inherent to subjective measurement of presence have led to the search for objective measurement techniques. Most objective measurement methods are based on involuntary response monitoring. The nature of the involuntary response can be postural (e.g., ducking for an incoming object or a startling reaction), social (e.g., facial expressions, eye contact or gestures) or physiological (e.g., cardiovascular behavior, skin conductance or neurophysiological patterns) [3]. Both postural and social responses are dependent of a particular situation that a user experiences and are hence unsuitable as a measurement method in all VR systems. Physiological responses however might prove to be more suitable in a general sense [3, 9, 11]. Slater points out that a promising way forward may be a combination of physiological measures and BIPs [11]. If there proves to be a common physiological BIP response that is invariant and which can be isolated under many different conditions, we're in business, so to speak.

Another objective measurement method, that is consistent with the view that the allocation of attentional resources plays an important role in presence, is known as secondary reaction time measure [3]. The fundamental assumption is that as more effort is dedicated to the primary task, fewer resources remain for the secondary task. Typical results show that, when attending to the primary task, people make more errors in responding to the secondary task or take longer to respond to a cue (often a tone or a light). It may thus be hypothesized that as presence increases, more attention will be allocated to the virtual world, which would mean an increase in secondary reaction times and errors. This hypothesis has yet to be empirically investigated.

Generally speaking, the objective measure that is being used as a presence measure should be tailored to the media experience the user is intended to have [3]. A drawback to objective measures is that they tend to be "all-or-none" and, even when they are not binary, their quantitative relationship to presence may be unclear [15].

5.0 DETERMINANTS OF PRESENCE

During the past years there has not really been agreement among researchers about the concept that the term presence encompasses [15]. However, there has been speculation and empirical research concerning the determinants of presence. Researchers have tried to find evidence for things that will enhance to feeling of presence, but since the evidence has been based on varies theories and measurement methods, it is not easy to compare and categorize them. It is important to designers of VR systems to have an understanding of the relative weighting of determinants, but it is found impossible to rank the importance of a given factor for presence without specifying the situation and task under consideration. For example, if a task requires a close view of an object and fine hand-eye coordination, then stereopsis is likely to be very important for the sense of presence. However, if the task is one of driving a car along a winding road without going out of the lane, the provision of stereopsis will be irrelevant for presence. This conclusion, however, does not eliminate the need for research on the various determinants.

Moreover, most determinants have not been described in terms of the effects they have on the identified dimensions of presence, as described in the section on theories, definitions, and multi-dimensionality. For example, researchers have coined virtual world realism, action realism and presentation realism as determinants of presence, while recent insights tell us that realism is actually a dimension of presence. There have been proposed various categorizations of determinants, see for instance Schuemie et al. [8]. Here the determinants will be described along the components of a VR system.

The computer system, which is a key component of a VR system, is assumed to be working without errors. Any process interruption or failure at either computer hardware level, operating system level or at the level of the VR simulation may diminish presence [3]. In practice however, this is easier said than done; smooth, error free operation is not always easy to accomplish.

The human body, which is central to the other side of the model of VR systems, is assumed to be the body of a healthy individual with normal abilities. Any abnormalities in bodily functions or mental abilities may have strong effects on the presence that can be elicited in an individual.

5.1 Sensors and Displays

Sensors and displays have often been identified to influence presence, which is logical because they ultimately determine what kind of information can be exchanged between the virtual world and the human mind. The following aspects of sensors and displays will be discussed: coverage, richness, device characteristics, obtrusiveness and consistency.

5.1.1 Coverage and Richness

Both the coverage (the amount of different modalities covered) and richness (the amount of coverage in a single modality) of usefully displayed sensory information presented in a consistent manner to the appropriate senses of a user are often mentioned as important factors to enable presence [2, 3, 5, 9, 11, 15, 16]. It is believed that feeding richer information to the senses and adding more displays to encompass more senses result in higher presence. For example, media that provide both aural and visual stimuli are said to produce a greater sense of presence than audio-only or video-only media. It is not entirely clear which senses contribute most to presence. In general, our visual and aural senses dominate our perception and have been most often identified with presence [5]. A similar argument can be held for the coverage and richness of the sensors in a VR system, as long as the signals coming from the sensors are processed in some meaningful way and presented to the user. Adding tracking of the head and hands has indeed been identified to increase presence [3, 6, 17].

5.1.2 Device Characteristics

There has been quite some research on the effects of specific characteristics of displays and sensors on presence [3, 9, 17]. Most displays rely on a process where discrete digital signals from the computer system are converted to some continuous analogue signal, and most sensors work just the other way around. In both cases, it is necessary that the discrete signals have a sufficiently high temporal resolution (update rate), signal resolution and spatial resolution to avoid breaking the illusion of non-mediation. Each human sense has a particular range of acceptable resolutions that can be perceived meaningfully. The desired resolution is the point at which the brain perceives the discrete sensory inputs as continuous input.

Many characteristics specific to visual displays encourage a sense of presence, including image quality, image size and viewing distance (which together determine the proportion of a user's visual field occupied by an image), motion and colour, monocular and binocular depth cues, and the use of a variety of camera techniques [3, 5, 15]. The characteristics of auditory displays that are most frequently recognized to be important for presence are sound quality and spatial hearing.

5.1.3 Obtrusiveness

For an illusion of non-mediation to be effective, the medium should not be obvious or obtrusive, i.e., it should not draw attention to itself and remind the user that he or she is having a mediated experience [6]. This implies that user grounded sensors and displays should have good ergonomical features (e.g., low weight) [3, 15], and ideally there would be no cables attached to a user [17].

Glitches, distortions or malfunctions in the hardware components make the mediated nature of the experience obvious and interfere with presence [3, 5]. It is suggested that noise, broadly defined as "information that is irrelevant to the intended communication regardless of the sensory channel through which it is transmitted" discourages presence.

5.1.4 Consistency

Another determinant of presence is the consistency between information in the different modalities. Information received through the various channels should describe the same objective world. Failure to meet this criterion emphasizes the artificial and thus the mediated nature of an experience [5, 15]. For instance, Usoh et al. identified investigator location incongruity (looking at the experimenter's voice location and seeing no one) caused many BIPs [17].

5.2 Virtual World

Sherman and Craig mention that, given a compelling presentation, presence can be caused by the content of the virtual world alone [9]. For instance, immersion is not necessary when reading a novel, nor is it desired. Although this is a valid observation, it is likely only to refer to involvement rather than the other dimensions of presence. In this section the influence of situation and task, virtual world realism, transfer of object permanence and occurrence and representation of human beings will be discussed.

5.2.1 Situation and Task

Amongst the most influential determinants of presence are the situation in the virtual world (sometimes also referred to as setting or story) that has been chosen by the designers of a VR system and the task that must be accomplished (sometimes also referred to as activity or goal) [3, 5]. As Heeter points out, the type of task may make it more or less difficult to establish presence [2]. Some situations are easier to represent by a VR system than others and they may depend on hardware or software components that are not easily created or combined with other components. A significant aspect or feature of a situation that is not properly supported by the VR system, but that a user expects to be there might cause more damage to the sense of presence than it does well. Perceptually significant anomalies in the virtual world may cause BIPs. Slater states that the study of presence is concerned with what minimal set of significant features is essential for maintaining the perception of which situation the participant is experiencing [11]. It seems that some minimal set of cues is needed to establish presence in a place and that the mind fills in the gaps. Therefore, choosing a right world and situation for a certain activity often comes down to searching a situation in which as many aspects as possible can be left out from the simulation.

5.2.2 Virtual World Realism

Realism can be applied to almost every aspect of a virtual world. Realism refers to aspects that users are familiar with of the real world that are modelled in the virtual world. For instance, realistic object properties, like shape, colour, weight, texture, and so on. Or, realistic behavior (laws) like persistency, consistency, object permanency or the law of gravity which are basic concepts in our real world. On a higher level realism can even refer to structural, social, economical or political aspects that are known from our human world. For instance, social elements, such as the acknowledgement of the users through the reactions of others, virtual or real, will be important for establishing a sense of social presence [3]. The ability to affect the world directly with meta-commands may be regarded as reducing the presence of the experience, so designers generally do not allow a user to perform such an operation [9].

Storylines, characters, and acting in some media content is more realistic than in others [5]. In a dramatic film or an interactive video game, if the story makes sense and doesn't depend only on coincidence, if the characters act in consistent and understandable ways, if the actors skilfully and convincingly create their personae, the experience is more likely to "ring true" for its users. Although it has not been empirically tested, this suggests that such realistic experiences are also more likely to evoke a sense of presence. To the extent that the content "rings false" the consumer is reminded of the mediated and artificial nature of the experience and the sense of presence should be destroyed. This concept of realism has been referred to as social realism, a component of perceived realism, verisimilitude, plausibility, and authenticity and believability. While social realism is usually applied to traditional media content, a virtual world can also contain more or less social realism: a world with a green sky, flying trains, and misshapen animals that speak Dutch would surely seem more surreal than real, and therefore would be less likely to evoke presence.

5.2.3 Transfer of Object Permanence

Like young children who don't completely learn the concept of object permanence in the real world until the age of 18 months, people entering a virtual world for the first time may have difficulty 'believing' in the permanence of objects there [9]. The addition of multiple senses corroborating the existence of an object in the world increases the believability of that object and, by extension, of the world itself. Developers can take advantage of sensory carryover to increase the impression of realness of the world. This works both to increase realism of particular objects and of the world as a whole. The more realistic an individual object seems, the more the user will expect it to behave naturally. One way to enhance realism is to make the sonic aspect of an object follow that object through the virtual space – even when it is no longer in view of the user. The result is that the participant 'realizes' that the object has a property of permanency. Likewise, the realism of the world as a whole can be improved when some of the objects are sensorially enhanced. So, when a user encounters one object that seems very real to them, the 'reality' of the other objects in the world will probably also increase. This can be very useful in helping participants overcome an initial barrier to suspending disbelief. Thus, participants' expectation of realness becomes higher, and they begin to trust the world as it is represented without testing it.

Since the haptic sense is very difficult to fool, haptic feedback that corroborates other sensory input can be particularly effective [9]. In one experiment a tracker was mounted on a physical dinner plate to produce a (passive) haptic display. The tracker on the real plate was linked to a plate representation in the virtual world. When a user is given the virtual plate and finds that it possesses all the properties of the real counterpart, he or she is more apt to extend his or her conception of realness to the rest of the virtual world. Transfer of object permanence can increase the user's suspension of disbelief to the degree that they won't attempt to walk through a wall and, thus, never discover that the walls are not as fully represented as the plate. Another experiment that exploits transfer of object permanence using passive haptics used styrofoam, foam core, and plywood located coincidentally with visual representations of walls and surfaces in the virtual world to physically mimic a portion of a virtual world. Thus, as the users see a virtual surface, and they reach out and touch it, they can feel some physical material in the proper location. When it comes to believe something is "real", the haptic sense is quite powerful. By coming into physical contact with an object, its existence is verified.

5.2.4 Occurrence and Representation of Human Beings

Another feature that may encourage a sense of presence is the number of people the user can encounter while using the VR system. Heeter suggested that "people want connection with other people more than any other experience [2]. Placing more than one person in a virtual world may be an easy way to induce a sense of presence regardless of the other perceptual features of the world". People that are represented by agents need to exhibit a range of autonomous behaviours to be believable [3].

Presence correlates highly with the degree of association with the own virtual body, the avatar [3, 17]. Avatar realism is worth a lot of work and investment, since user identification with the virtual body is such a strong factor in presence. Also, clothing identification has been identified as being surprisingly important in some studies [17].

5.3 Action

Most researchers have either implicitly assumed or explicitly suggested that a major or even the primary cause of presence is the ability to perform actions in the virtual world [5]. Often the term interaction is used in this context, which refers to an action and its associated response. Sherman and Craig state that when the virtual world responds to our actions, we become more involved with that world, increasing our sense of presence [9].

An action requires active usage of the user's body. The real proprioceptive sensations of an action together with a presentation of any changes to the virtual world, strongly affect presence [13, 14]. Regenbrecht and Schubert suggest that interactivity should first and foremost influence spatial presence because it directly determines the meshings formed between body and the virtual world [6]. It should affect involvement only as far as it draws additional attention to the virtual world.

In this section action types, number and range, control of action, action realism and illusionary action will be discussed.

5.3.1 Action Types

Several types of actions, i.e., manipulation, travel and communication, have been identified to influence presence [15]. A sense of presence develops from the mental representation of movement of the own body (or body parts) as a possible action in the virtual world, or from the meshing of bodily actions with objects or agents in the virtual world [6].

The ability to travel, in its different forms (from flying to virtual walking to real walking about in a significant space) has widely been acknowledged to enhance presence [6, 16, 17]. Usoh et al. have shown results, suggesting that presence is higher for virtual walkers than for flyers, and higher for real walkers than for virtual walkers [17]. However, the difference between groups diminishes when oculomotor discomfort is taken into account. If one wants increased presence or a visceral estimate of spatial extents of human-scale spaces, real walking is best, and virtual walking seems clearly better than flying. Regenbrecht and Schubert write that travel as a basic possibility to interact with the virtual world has an increasing effect on spatial presence [6]. Furthermore, it increases judgments of realness. These results imply that the possibility to move oneself freely through a virtual space increases the sense of being in this space and acting in it, as well as the sense that this space is real. There was no effect of travel on involvement.

There seems to be an obvious difference between communication and the other two action types identified [4]. Communication is central to social presence, but unnecessary to establish a sense of physical presence. Indeed, a medium can provide a high degree of physical presence without having the capacity for transmitting reciprocal communicative signals at all. Use of communication with respect to virtual worlds has mostly been limited to use of voice. IJsselsteijn et al. predict that when technology increasingly conveys non-verbal communicative cues, such as gaze direction or posture, social presence will increase [3]. To perceive a technology as a social entity instead of an artificial medium, the user needs to be able to interact with it. The number of previous user inputs that are acknowledged in a response is especially important. In a different context, a computer which appears to have no memory of recent events in an interaction should be less likely to evoke the illusion of presence [5].

5.3.2 Number and Range

The number (and type) of characteristics of a VR experience that can be modified by the user also help determine the degree to which it can be called interactive. It is suggested that a highly responsive virtual world, one in which many user actions provoke even unnatural responses (e.g., entering a room produces verbal or musical greetings or rain) could evoke a greater sense of presence than less responsive environments [5] ("the more possibilities there are of interacting, the more cognitive meshings are possible, and presence increases" [6]).

Another determinant is the range or amount of change possible in each characteristic of the experience [5]. Interactivity, and perhaps therefore presence, is enhanced by expanding the degree to which users can control each attribute of the mediated experience. For example, in a highly interactive VR system the user can look in any direction, move over large distances in each direction, proceed at any pace and in any sequence desired, pick up, feel, and move many different objects each with different textures, and change the type and volume level of ambient sounds. In a different context, the larger the vocabulary of a computer speech recognition system (i.e., the more words it recognizes and to which it responds appropriately) the more interactive the experience is.

5.3.3 Control of Action

An important aspect of interaction is the amount of control that a user has over it [3]. Stanney et al. suggest that the more control a user has over their actions in a virtual world, the higher the ensuing sense of presence [15]. They report on a study that found that presence was higher for users who were in control of their own actions in the virtual world as compared to passive observers. Driving a virtual car created higher presence than merely being a passenger in it [6]. This suggests that if users are provided with a high level of user-initiated control, presence may be increased.

5.3.4 Lag/Latency

Another important factor that affects interactivity is the lag that is introduced by the accumulated processing time required by the different components of a VR system [5]. The lag determines the speed with which a VR system is able to respond to user inputs. An ideal interactive medium responds in “real time” to user input; the response or lag time is not noticeable. Noticeable lags, temporal distortions and response latency have widely been acknowledged to affect interactivity and hence presence [2, 3, 5, 8, 9, 15, 17].

5.3.5 Action Realism

The degree of correspondence between the type of user input and the type of medium response is another variable that determines how interactivity affects presence [5]. The “mapping” between these two can vary from being arbitrary (e.g., pressing a sequence of keys on a keyboard to adjust a visual display) to natural (e.g., turning one’s head in a virtual reality system to see the corresponding part of the environment). The mapping between the user’s actions and the perceptible spatiotemporal effects of those actions need to be consistent. For example, using head tracking, a turn of the user’s head should result in a corresponding real-time update of the visual and auditory displays [3]. Lombard and Ditton write that it is a widely accepted working hypothesis that using our familiar sensorimotor skills to manipulate virtual objects directly by means of whole-hand input devices contributes to our sense of presence much more than writing programs, twisting knobs, or pushing a mouse to accomplish the same task [5].

Weghorst and Bellinghurst (cited by Stanney et al. [15]) manipulated the design of VR systems through the degree of abstractness of objects, as well as the use of a ground plane and other spatial landmarks. They found that designs that eased the interaction were most predictive of the sense of presence. This suggests that if interaction can be streamlined, interactive fidelity or presence may be enhanced.

5.3.6 Illusionary Action

An interesting experiment by Regenbrecht and Schubert shows that presence develops from the mental representation of possible bodily interactions, and not from the objective possibility to interact per se [6]. It follows that, under some circumstances, objective possibilities for interactions should not enhance presence, for example, when they are not seen or ignored, or when actions and consequences cannot be causally linked. From assuming that it is not objective possibilities to interact but perceived possibilities to interact that determine presence, it follows that the mere illusion of an interaction should enhance presence, even when objectively no interaction takes place. Results suggest that the anticipation of possible interactions increased spatial presence. Neither involvement nor judgment of realness was influenced.

5.4 Presentation

Although designers of VR systems do not always make decisions consciously about the way in which the world is presented to a user, there have been identified some issues that can influence presence. In this section visual and aural presentation realism and use of conventions will be discussed.

5.4.1 Visual Presentation Realism

Several researchers have emphasized the relationship between presentation realism and presence [15, 17]. For instance, visual presentation realism has been mentioned to enhance the sense of presence (although the level of action control turned out to be more important). It has even been suggested that any VR experience must be extremely realistic for it to be able to give a sense of presence. Anything that demonstrates that you are not in the real world will result in a BIP.

On the other hand, various papers suggest that presence is possible in experiences where visual realism is not a factor. In such cases, other factors, such as the interaction with the game or the story, keep users engaged. For instance, books, which have very limited visual realism, effectively engage our minds and imagination [15]. Additionally, it has been shown that cartoonishness of a rendered virtual world doesn't prevent users to become present. Attempts to render a world in a photo-realistic way can even make presence difficult, because any flaw in the realism will spoil the effect [9]. However, anomalies in an environment are not equal in their significance: some will induce a BIP, and others won't. For example, in the depiction of a virtual human, an anomaly in overall body shape is likely to be far less significant than the shape and movements around the eyes and mouth [11].

Despite these seemingly contradictory points-of-view, this dispute about the necessity of visual fidelity to experience presence seems to fit well with the multi-dimensional view of presence, where realism is separated from involvement and spatial presence.

5.4.2 Aural Presentation Realism

Similar argumentation as with visual presentation realism may be held for aural presentation realism. Aural realism of virtual spaces requires replicating the spatial characteristics of sounds like the changing intensity of a race car engine as it approaches a listener and screeches past (Doppler effect), or the tapping of footsteps as they echo in a dark, empty corridor; or the chatter of a conversation off in the corner of a room [5]. It is argued that sound greatly enhances the participant's ability to become present in the world. Sound is compelling. From ambient sounds that give cues to the size, nature and mood of the setting to sounds associated with particular objects or characters nearby, aural presentation is said to be the key to user understanding and enjoyment [9]. The volume (loudness) of audio stimuli may also have an impact on presence, with particularly low and perhaps particularly high levels less effective than moderate ("realistic") levels. It has been suggested that the proper use of ambient sounds and music can evoke an atmosphere or sense of place, thereby heightening the overall feeling of presence in the virtual world [5].

Computer-based technologies increasingly present information to users with voices (either recorded human voices or computer-generated ones). The use of voice is a potent social cue and has been shown to elicit perceptions that one computer is made up of multiple distinct entities and to evoke gender stereotypes [5]. It seems likely that voices that sound more human (with higher audio realism) enhance the illusion of social presence.

5.4.3 Use of Conventions

One way in which users are reminded of the true nature of their experience is through the use of conventions that users have come to associate specifically with mediated presentations and experience [5]. In movies and television when the passage of time is represented by spinning hands on a clock, when the transition to a dream or flashback is represented with a distorted dissolve between images, when dramatic or emotional background music telegraphs the end of a segment, when credits and other text messages are superimposed over story action, when identification logos appear in the corner of the screen, when an unseen narrator describes events, and when plots and dialogue follow predictable formulae, the media user is reminded that rather than having a non-mediated experience he or she is watching something created and artificial. This realization is likely to interfere with a sense of presence.

5.5 Human Mind

There are many aspects of the human mind that are likely to play a role in presence, which is not surprising because presence itself is a subjective experience caused by processes of the human mind [3]. Characteristics that are thought to influence presence include the user's perceptual, cognitive and motor abilities (e.g., stereoscopic acuity, susceptibility to motion sickness, concentration), prior experience with and expectations towards the mediated experiences, and willingness to suspend disbelief. Allocating sufficient attentional resources to the mediated environment has also been proposed as an important component of presence. Relevant individual characteristics will possibly vary with age and sex of the user. It is likely that various mental health conditions, like depression, anxiety, or psychotic disorders, are also likely to affect an individual's sense of presence, since they are known to have a clear effect on how people experience the world around them.

5.5.1 Individual Characteristics / Personality Type

Sensory dominance is an important variable that affects depth of presence [13]. There are three different types of sensory dominance: Visual, auditory or haptic learners. The depth of presence is related to the users' innate sensory dominance and the type of feedback given by the VR system. For applications focused on visual feedback, the visually oriented person will experience more presence than the auditory oriented person. In a virtual world without sounds, aurally dominant users felt less present. For kinaesthetically oriented users, use of an avatar in the virtual world resulted in more presence.

Lombard and Ditton give a list of characteristics of an individual that may influence presence, including age, gender, sensory dominance, cognitive style, degree to which a user "screens" complex stimuli, level of sensation seeking, need to overcome loneliness, introversion/extroversion, locus of control, dominance/ submissiveness [5].

5.5.2 Willingness to Suspend Disbelief

An identical VR system with the same virtual world might generate a sense of presence in one user and not in another, or might generate presence in the same user on one occasion, but not on another one [5]. Although almost no research has been conducted on the issue, it seems clear that characteristics of users are important determinants of presence. One variable that is likely to be especially important in this regard is the user's willingness to suspend disbelief.

A person participating in a VR experience has chosen to engage in the activity and knows that it is a mediated experience [5]. He or she can encourage or discourage a sense of presence by strengthening or weakening this awareness. If we want to increase a sense of presence for ourselves we try to "get into" the experience, we overlook inconsistencies and signs that it is artificial, we suspend our disbelief that the experience could be non-mediated. When we want to decrease presence, as when we watch frightening or disturbing media content, we remind ourselves that "this isn't really happening, it's only a movie, game, etc." The willingness to suspend disbelief probably varies both across individuals (e.g., some people are so naturally curious about how a medium works that they simply can not suspend disbelief and enjoy the experience) and within the same individuals across time (e.g., it may be more difficult to suspend disbelief and escape to a mediated world when one is preoccupied by problems at home or at work).

5.5.3 Attention

How sharply users focus their attention on the virtual world partially determines the extent to which they will become involved in that environment and how much presence they will experience. Witmer and Singer suggest that as users focus more attention on the stimuli from the virtual world, they should become more involved in the experience, leading to increased presence [19]. Attention and thus involvement depend on the significance or meaning that the individual attaches to stimuli, activities, or events [15]. The world has to be personally meaningful. If the participant does not find the topic or style in which the content is conveyed absorbing, there is little hope of engagement [3, 9]. This argument for the importance of attention in sense of presence is similar to the concept that the experience of presence is based on attention to the continuity, connectedness, and coherence of the stimulus flow [15]. Distractions that draw the user's attention from the virtual world to the real world are likely to diminish the user's sense of presence [3].

5.5.4 Mood

The experience of presence is so much affected by the state of mind (mood) of the user that anything that affects the user's subconscious mind may affect how the experience is perceived [9]. For example, the venue, or setting in which the VR system resides, can have a great impact on how an event is experienced, because the venue is part of the whole experience. The venue puts the participant in a certain state of mind when entering the virtual world, although an individual may experience the same application in the same venue differently on different days. A primary effect of ambient sounds is generally used to set the mood of an experience, which can have the effect of making the experience more compelling, increasing presence.

5.5.5 Knowledge, Prior Experience and Expectations

It should be easier for users unfamiliar with the nature of a VR system and how it functions to experience presence [5]. An engineer can not help but notice flaws in a virtual world or the image in a VR display because he or she either knows or wants to know what is responsible for the flaw. This knowledge reminds him that the experience is mediated. The situation is analogous to a magician who knows how a trick is performed and is therefore unimpressed with the illusion.

Closely related to this is the effect of experience with a medium [5]. The first time a person uses a system capable of generating a sense of presence, he or she is unfamiliar with the system, how it is used, and the nature of the experience. This unfamiliarity likely discourages a sense of presence, but as the user becomes more expert at using and manipulating the experience and more comfortable with it in general, this effect should fade. Continued experience may then either increase presence ("having 'been there before' helps you believe you are there again"[2]) or decrease it as the well-known habituation effect causes an initially impressive and novel sense of presence to fade as users become more experienced.

Extended exposure may increase presence because it enhances other factors thought to be related to presence, including the amount of practice with tasks, the familiarity with the virtual world, and the level of sensorial adaptation to the intersensory and sensorimotor discordances presented by the displays, but exposure and presence could also be negatively related [15]. Thus, it is uncertain whether long-duration exposure will enhance presence by engendering familiarity or reduce presence due to adverse side effects.

It matters what the users expect of the environment, and what they have in mind in terms of anticipations, goals, and experiences. These variables influence how they mentally construct the environment in terms of possible actions in it, and therefore these variables influence the sense of presence. Pausch et al. (cited by Regenbrecht and Schubert [6]), having observed the reactions of thousands of guests at a Disney installation, state that they can improve the experience by telling a pre-immersion background story and by giving the guest a concrete goal to perform in the virtual environment."

6.0 CONCLUDING REMARKS

It has taken some time for researchers to reach consensus about the concept that presence encompasses. One reason for this is that there are so many ways to look at it, for instance from a philosophical, psychological, neurobiological or technical perspective. Another reason is that presence is inherently subjective, and comparing experiences has proven to be very difficult. However, there seems to be convergence in the views on presence that are presented in the stream of scientific publications, both on theoretical as well as empirical issues. An important insight is the multi-dimensionality of presence. Presence can be differentiated into social presence and physical presence, of which the latter can be differentiated further into spatial presence, involvement and realism. The next step forward would be insight into the relative importance of all determinants of presence distinguished.

7.0 REFERENCES

[1] Ellis, S. (1996). Presence of Mind: A Reaction to Sheridan's Musings on Telepresence, Presence: Teleoperators & Virtual Environments, 5(2), 247-259.

[2] Heeter, C. (1992). Being There: The Subjective Experience of Presence, Presence: Teleoperators & Virtual Environments, 1(2), 262-271.

[3] IJsselsteijn, W.A., de Ridder, H., Freeman, J. and Avons, S.E. (2000). Presence: Concept, Determinants and Measurement, Proceedings of SPIE: Human Vision and Electronic Imaging, 3959, 520-529.

[4] IJsselsteijn, W.A. and Riva, G. (2003). Being There: The Experience of Presence in Mediated Environments, in Being there, concepts, effects and measurements of user presence in synthetic environments, Riva, G., Davide, F., IJsselsteijn, W., Eds., 3-16.

[5] Lombard, M. and Ditton, T. (1997). At the Heart of It All: The Concept of Presence, Journal of Computer-Mediated Communication (on-line serial) 3(2). Available: http://jcmc.indiana.edu/vol3/issue2/lombard.html.

[6] Regenbrecht, H. and Schubert, T. (2002). Real and Illusory Interactions Enhance Presence in Virtual Environments, Presence: Teleoperators & Virtual Environments, 11(4), 425-434.

[7] Schubert, T., Friedmann, F. and Regenbrecht, H. (2001). The Experience of Presence: Factor Analytic Insights, Presence: Teleoperators & Virtual Environments, 10(3), 266-281.

[8] Schuemie, M.J., van der Straaten, P., Krijn, M. and van der Mast, C.A.P.G. (2001). Research on Presence in VR: a Survey, Journal of Cyberpsychology and Behavior, 4(2), 183-202.

[9] Sherman, W.R. and Craig, A.B. (2003). Understanding Virtual Reality.

[10] Slater, M. (1999). Measuring Presence: A Response to the Witmer and Singer Presence Questionnaire, Presence: Teleoperators & Virtual Environments, 8(5), 560-565.

[11] Slater, M. (2002). Presence and the Sixth Sense, Presence: Teleoperators & Virtual Environments, 11(4), 435-440.

[12] Slater, M. and Steed, A. (2000). A Virtual Presence Counter, Presence: Teleoperators & Virtual Environments, 9(5), 413-434.

[13] Slater, M., Usoh, M. and Steed, A. (1994). Depth of Presence in Virtual Environments, Presence: Teleoperators & Virtual Environments, 3(2), 130-144.

[14] Slater, M. and Wilbur, S. (1997). A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments, Presence: Teleoperators & Virtual Environments, 6(6), 603-616.

[15] Stanney, K.M., Salvendy, G., Deisigner, J., DiZio, P., Ellis, S., Ellison, E., Fogleman, G., Gallimore, J., Hettinger, L., Kennedy, R., Lackner, J., Lawson, B., Maida, J., Mead, A., Mon-Williams, M., Newman, D., Piantanida, T., Reeves, L., Riedel, O., Singer, M., Stoffregen, T., Wann, J., Welch, R., Wilson, J. and Witmer, B. (1998). Aftereffects and Sense of Presence in Virtual Environments: Formulation of a Research and Development Agenda, International Journal of Human-Computer Interaction, 10, 135-187.

[16] Steuer, J. (1992). Defining Virtual Reality: Dimensions Determining Telepresence, Journal of Communication, 42(2), 73-93.

[17] Usoh, M., Arthur, K., Whitton, M.C., Bastos, R., Steed, A., Slater, M. and Brooks, F.P. (1999). Walking > Walking-in-Place > Flying, in Virtual Environments, Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), 359-364.

[18] Usoh, M, Catena, E., Arman, S. and Slater, M. (2000). Using presence questionnaires in reality. Presence: Teleoperators & Virtual Environments, 9(5), 497-503.

[19] Witmer, B.G. and Singer, E. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire, Presence: Teleoperators & Virtual Environments, 7(3), 225-240.

A Review of the Mental Workload Literature

Brad Cain
Defence Research and Development Canada Toronto
Human System Integration Section
1133 Sheppard Avenue West
Toronto, Ontario M3M 3B9
CANADA
Telephone: 1-416-635-2195

E-mail: brad.cain@drdc-rddc.gc.ca

1.0 INTRODUCTION

The intent of this paper is to provide the reader with an overview of the mental workload literature. It will focus on other state of the art surveys with reference to some specific reports of the practical application of mental workload measurement. The surveys will be limited to those in English. Manzey[1] reportedly provides a review of psychophysiological methods in German; a comparable, recent review in English was not found although a NATO RTO report (Wilson 2004, pp. 64-65 and Chapter 8) provides some guidance in this respect. Appendix 1 lists a search for references to workload measurement techniques using the GOOGLE[2] search engine. The intent is to give the reader an appreciation of where work has been focused, or at least as reported on the Internet.

1.1 Definitions of Workload

Despite interest in the topic for the past 40 years, there is no clearly defined, universally accepted definition of workload. Huey and Wickens (1993, p. 54) note that the term "workload" was not common before the 1970's and that the operational definitions of workload from various fields continue to disagree about its sources, mechanisms, consequences, and measurement." Aspects of workload seem to fall within three broad categories: the amount of work and number of things to do; time and the particular aspect of time one is concerned with; and, the subjective psychological experiences of the human operator (Lysaght, Hill et al. 1989).

Workload is thought of as a mental construct, a latent variable, or perhaps an "intervening variable" (Gopher and Donchin 1986, p. 41-4), reflecting the interaction of mental demands imposed on operators by tasks they attend to. The capabilities and effort of the operators in the context of specific situations all moderate the workload experienced by the operator. Workload is thought to be multidimensional and multifaceted. Workload results from the aggregation of many different demands and so is difficult to define uniquely. Casali and Wierwille (1984) note that as workload cannot be directly observed, it must be inferred from observation of overt behaviour or measurement of psychological and physiological processes. Gopher and Donchin (1986, p. 41-2) feel that no single, representative measure of workload exists or is likely to be of general use, although they do not provide guidance on how many workload measures they feel are necessary or sufficient.

There are few formal definitions of workload. Most definitions are of the form:

---

[1] Manzey, D. (1998) Psychophysiologie mentaler beanspruchun. In F.Rösler (Ed.) Ergebnisse und Anwendungen der Psychopsychologie, Serie 1. Biologische Psychologie. Enzyklopãdie der Psychologie. Göttingen, Germany: Hogrefe.

[2] http://www.google.com

1) "Mental workload refers to the portion of operator information processing capacity or resources that is actually required to meet system demands." (Eggemeier, Wilson et al. 1991, p. 207)
2) "... mental workload may be viewed as the difference between the capacities of the information processing system that are required for task performance to satisfy performance expectations and the capacity available at any given time." (Gopher and Donchin 1986, p. 41-3)
3) "… the mental effort that the human operator devotes to control or supervision relative to his capacity to expend mental effort … workload is never greater than unity." (Curry, Jex et al. 1979)
4) "... the cost of performing a task in terms of a reduction in the capacity to perform additional tasks that use the same processing resource." (Kramer, Sirevaag et al. 1987)
5) "... the relative capacity to respond, the emphasis is on predicting what the operator will be able to accomplish in the future." (Lysaght, Hill et al. 1989, p. 27)

Gopher and Braune (1984) suggest that the workload construct was conceived to explain the inability of human operators to cope with the requirements of a task, and that workload measures are an attempt to characterize performance of a task relative to the operator's capability. They note that there is little knowledge to link the measurement of workload by any one paradigm to others and the lack of a formal theory of workload has lead to a proliferation of disparate methods with little chance of reconciliation. Gopher and Braune's early findings seemed to argue that workload reflects demands on a single, undifferentiated pool of resources (Gopher and Braune 1984, p. 530), where all tasks interact similarly and concurrent task demands were principally additive with a constant "overhead". This strict perspective is no longer held. It is now thought that the human information processor is appropriately represented as comprising multiple resources that are engaged differently according to the characteristics of the task demands (Jex 1988; Wickens and Hollands 1999). Although task demands and operator capabilities may be multidimensional, it is unclear whether the **conscious perception of workload** should be represented this way or as a single, scalar quantity.

Mental workload can be influenced by numerous factors that make a definitive measurement difficult. Jex (1988) implies that mental workload derives from the operator's meta-controller activities: the cognitive "device" that directs attention, copes with interacting goals, selects strategies, adjusts to task complexity, sets performance tolerances, etc. This supports the intuitive notion that workload can be represented as a function, and the utility of univariate workload measures as globally sensitive estimates of workload, while acknowledging that tasks of differing characteristics interfere differently. Alternatively, Wierwille (1988, p. 318) suggests that an operator faced with a task is fully engaged until the task is done, then is idle or engages in another task. It is not clear how this can be reconciled with multitask performance demonstrating interference effects without resorting to some manner of time sharing among concurrent tasks. Wierwille's position seems to preclude interleaving idle and active intervals during task execution.

Workload is frequently described by terms such as *mental strain* ("... the concept of mental effort ...") and *emotional strain* ("... the excess mental effort that comes form anxiety evoking cognitive aspects of the task ..."). Boucsein and Backs (1999, p. 8) outline what is perhaps an alternative formulation for representing workload or strain, as a Three Arousal Model, more tightly coupling emotions and stress to workload. Gaillard (1993) maintains that workload and stress, while related, lack proper and distinct definitions. Both stress and workload involve environmental demands and the ability of the operator to cope with those demands, but these two concepts come from different theoretical backgrounds. Gaillard separates workload from emotion, with both under the control of a higher, mental mechanism, similar to a metacontroller. If this workload is a manifestation of the investment of effort by the metacontroller, then affective factors play a complementary role to information processing in the perception of workload, best represented as a two dimensional model of cognitive energy mobilization. Information-processing models would be incomplete according to this perspective.

Colle and Reid (1999) state that "... the concept of mental workload is an applied construct and ... does not have a one-to-one relationship with attentional capacity or resources in information processing theories." Colle and Reid focus on the amount of mental work that can be accomplished within a period of time; "... mental workload is considered to be the average rate of mental work ...". They describe a procedure for defining workload or demand equivalence of tasks using double trade-off evaluations, but they do not present a philosophy to identify the appropriate time interval for a task in such an assessment. They present the results of three experiments as support for their proposal to develop a globally sensitive secondary task measure battery. Huey and Wickens (1993, pp. 57-68) provide a good overview of many of the external, task factors contributing to workload.

In summary, a commonly accepted, formal definition of workload does not exist. Workload can be characterized as a mental construct that reflects the mental strain resulting from performing a task under specific environmental and operational conditions, coupled with the capability of the operator to respond to those demands. Operational definitions will likely continue to be proposed and tested, but unless an imperative need arises for a universal definition, each field and perhaps each investigator will continue with their "culturally preferred" definition of workload.

1.2 Reasons for Measuring Workload

The principal reason for measuring workload is to quantify the mental cost of performing tasks in order to predict operator and system performance. As such, it is an interim measure and one that should provide insight into where increased task demands may lead to unacceptable performance. Wickens (1992, p. 390) asserts "... performance is not all that matters in the design of a good system. It is just as important to consider what demand a task imposes on the operator's limited resources. Demand may or may not correspond with performance." Mulder, Mulder et al. (1999, p. 140) note "the main reason to measure physiological activity during and after mental work is to assess the costs involved in performing mental tasks and to measure the duration of the imposed effects upon the task performer." Presumably, these purposes are only interim objectives in applied or laboratory settings; the ultimate objective is assumed to be improved working conditions, intuitive workstation design, or more effective procedures.

There may also be legal reasons to measure workload. Workload measurement during the assessment of new user interfaces may be a requirement in order to attain certification for use; for example, certification of new aircraft cockpit designs[3]. The certification process may specify the method of workload measurement selected and, hopefully, there are some rational, validated criteria justifying its use as a surrogate for in-service performance.

In the comparison of system designs, procedures, or manning requirements, workload measurement can be used to assess the desirability of a system if performance measures fail to differentiate among the choices. Implicit in this approach is the belief that as task difficulty (workload) increases: performance usually decreases; response times and errors increase; control variability increases; fewer tasks are completed per unit time; task performance strategies change (Huey and Wickens 1993); and, there is less residual capacity to deal with other issues. There is, however, strong evidence to show this is not necessarily the case for monitoring or vigilance applications. In monitoring applications, workload may be considered low despite the difficulty of

---

[3] FAA (1993) Airworthiness Standards. FAR Parts 23 and 25, Appendix D. http://ecfr.gpoaccess.gov

maintaining attention. The dissociation between workload and performance is generally acknowledged although not well understood. Nevertheless, part of the system designer's objective is to optimize system performance and workload is considered one of the factors to be considered in the optimization process (Mitchell 2000).

1.3 Criteria for Workload Measurement Methods

If workload is being measured in an experimental setting, the measurement options are generally wider than those for an operational setting. Many of the workload measurement techniques can be used successfully to differentiate among empirical conditions, and perhaps even produce interval or ratio measures.

Some concerns about the practical application of workload measures based on laboratory studies are: the lack of ecological validity or context complexity; the lack of subject acceptance, commitment or expertise; the lack of assessment of the effect of strategy shifts both on performance, scheduling, and on the workload measurements themselves. To address these and other concerns, O'Donnell and Eggemeier (1986) proposed several criteria to guide the selection or development of mental workload measurement techniques:

1) The method must be reliably sensitive to changes in task difficulty or resource demand and discriminate between significant variations in workload.
2) The method should be diagnostic, indicating the source of workload variation and quantify contribution by the type or resource demand.
3) The method should not be intrusive or interfere with performance of the operator's tasks, becoming a significant source of workload itself.
4) The method should be acceptable to the subjects, having face validity without being onerous.
5) The method should require minimal equipment that might impair the subject's performance.

Other criteria have since been added to this list:

6) The method should be timely and sufficiently rapid to apply to capture transient workload changes.
7) The method should be reliable, showing repeatability with small variance compared with main effects.
8) The method should be selectively sensitive to differences in capacity demand and not to changes unrelated to mental workload (such as emotional stress; others such as Gaillard (1993) or Gaillard and Kramer (1999) might debate this restriction).
9) The method should be insensitive to other task demands, such as physical activity beyond the conduct of the tasks. (Casali and Wierwille 1984, p. 1034)

The measurement technique selected should also meet formal, axiomatic constraints. Colle and Reid (1997) note that only after workload measurement techniques adopt this approach, such that subjective and performance (and physiological) measures can be directly compared, can we fully understand the concept of workload. Others had related concerns that would feed into formal measurement methods, characterizing not only the tasks, but also the operator-abilities and costs incurred performing tasks (Derrick 1988). Objective measurement techniques, while attending to the technical requirements of measuring the physical quantity involved, do not address how these measurements can be transformed into a workload measure. The work of Colle and colleagues on subjective scale development has potential in objective performance and physiological measures as well the subjective scales.

Some measurement techniques attempt to address individual differences, either through developing weights to proportion scale ratings to an overall workload calculation that are operator specific (as in SWAT and NASA TLX) or through base-lining, such as in the physiological measures. Many researchers feel that the concept of individual differences is central to workload measurement. As such, measurement techniques should be designed to capture those differences and reflect them in the values obtained from a sound theoretical framework.

As mentioned, the choice of workload measurement technique for an operational application is more constrained than in a laboratory setting, or an empirical setting. Measurement of workload by several, unobtrusive techniques has been proposed to take some pre-emptive, mitigating action to maintain performance, such as automated aiding, so real time responses are necessary. The measurement technique has to have minimal interference with operator activity, either mental or physical.

2.0 NOTABLE REVIEWS OF WORKLOAD MEASUREMENT TECHNIQUES

The most organized discussion of workload assessment and its measurement was found in Chapters 41 and 42 of the “Handbook of Perception and Human Performance” (Boff, Kaufman et al. 1986). In Chapter 41, Gopher and Donchin (1986) present both a historical and a state of the art review of workload, its definition and early models up until the mid 1980s. Gopher and Donchin give an introduction to the various classes of workload measurement, noting advantages, disadvantages, and controversies about their use. In Chapter 42, O’Donnell and Eggemeier (1986) expand on Gopher and Donchin’s overview, discussing numerous specific measurement techniques for each workload measurement class. These are useful references for learning the fundamentals of workload measurement as well as identifying potential measurement techniques.

Moray (1979), Roscoe, Ellis et al. (1978) provide collections of papers in separate monographs that also review the state of knowledge up to the 1980s. The papers cover a broad range of topics from the development of specific measurement techniques to more general reviews of the philosophy and problems associated with workload measurement. Moray’s (1979) NATO workshop proceedings are among the early attempts to collate and organize “… the enormous amount of knowledge about workload and many models for it …” into a coherent model that would be theoretically sound and practicable. It involved psychologists, engineers, modellers, and ergonomists, resulting in almost 30 papers. While it did not create a functional workload model, it did serve to focus and inform the various communities on various aspects of workload measurement. Some of the work in Moray (1979) is still relevant and useful, although the science has advanced somewhat in the past 25 years; nevertheless, this monograph also makes a good starting point on various aspects of workload measurement. Roscoe, Ellis et al. (1978) provide an early review of the state-of-the-art, referencing works from the 1950s that predated current workload theories, associating workload both with mental effort (without defining effort formally) and with the extent to which an operator is engaged by his duties. They acknowledge task demands, operator capabilities, and contextual temporal demands as being components of workload.

An interesting proposal from the experimental psychology group of Moray’s NATO symposium was that not only is workload multidimensional, it should be considered as a vector rather than the more typical scalar quantity. Further, this vector representation will be task specific (or perhaps specific classes of tasks), although they readily admit they do not know the dimensions of such a vector. An analogous problem is that of speed and velocity. Speed is the magnitude of the velocity vector, but in some situations, direction counts such as in the fuel requirements for aircraft flying in winds of different headings, covering similar distances in similar durations: to the airline the result is important to their profitability, but to the passenger, the result isn’t noticeable. Some measurement techniques may be sensitive to specific components of the workload vector only and insensitive to others. Alternatively, some measures may be sensitive to several dimensions, but they may not be able to differentiate the contributions, resulting in an apparent dissociation of causes and effects.

Wierwille and colleagues reported a focused series of experiments stressing different aspects of mental demand on workload methodology in aircraft simulators. Each study confirmed that measurement techniques are differentially sensitive to the load manipulations. In the study of psychomotor loading (Wierwille and Connor 1983), only 5 of the 20 workload measurement techniques were judged appropriate for the piloting task and of those 5, only 3 had a monotonic relationship with the load manipulation. In the communications (Casali and Wierwille 1983) and cognitive (mediation) studies (Wierwille, Rahimi et al. 1985), 7 of 16 measures showed significant sensitivity. In the perceptual study, (Casali and Wierwille 1984) 7 of 14 measures were sensitive to increasing perceptual loads. The specific techniques that were found sensitive are noted later in this document under "5.0 Recommending workload measures".

Wickens (1992) provides a brief overview of the general classes of techniques, noting some of the advantages and disadvantages of each class. de Waard and Farmer provide more recent reviews of workload measurement methods. de Waard (1996) provides an extensive as well as critical assessment of both general categories and specific measurement techniques that provides a more current perspective on the views of O'Donnell and Eggemeier. de Waard's thesis is available on-line[4] and is a valuable reference both for those starting in the field of workload as well as for researchers looking for potential workload measurement methods. Farmer and Brownson (2003) provide a concise, current review of workload measurement methods and offer professional recommendations on techniques suitable for use in human-in-the-loop simulations. Farmer and Brownson focus on commonly used methods, commenting on suitability for the Eurocontrol Integra program involving air safety and air traffic management.

Castor (2003) and Wilson (2004) provide some of the most current reviews of workload measurement techniques, although distribution of these reports may be restricted and difficult to obtain. Each of these reports assesses various techniques on a number of levels, providing guidance on the maturity, sensitivity, reliability, and usefulness of each method. Castor (2003) also provides an assessment process to help select which of the various measures may be best applied based on the phenomenon under study.

These notable reviews suggest that workload is on the minds of many practitioners, however, the researchers cited are rather few in number. The works of these researchers re-appear across the literature, and the current review is no exception. Observations in the current review are based on the literature rather than conclusions arising from experiments conducted. In the following pages, different methods from each of the three categories are reviewed and reported with references for the interested reader to form their own hypotheses that can be tested. The reviews of Gopher and Donchin (1986) and O'Donnell and Eggemeier (1986) are recommended as concise historical and technical overviews of the field, as well as that of de Waard (1996) for his assessment of specific methods.

3.0 WORKLOAD MEASUREMENT TECHNIQUE CATEGORIES

Workload measurement techniques are typically organized into three broad categories: self-assessment or subjective Rating scales; Performance measures (including subdivisions of primary and secondary task measures); and, Psychophysiological measures (Eggemeier, Wilson et al. 1991, p. 207). It has already been noted that different measures are sensitive to different aspects of workload and not all workload measures are

---

[4] http://www.home.zonnet.nl/waard2/mwlch1.htm 9Jan04.

assessing the same thing. Part of this confusion arises from the lack of an accepted definition of workload and the tendency to use the term workload to mean either the demands imposed on the user, the effort the user exerts to satisfy those demands, and the consequences of attempting to meet those demands (Huey and Wickens 1993, p. 55).

Questionnaires and interview techniques, while informative, are not considered here as methods of measuring workload as they are more verbal descriptions of what the operator is experiencing. Questionnaires are complex to design properly to avoid unwanted biases, awkward to validate, and difficult to generalize (although the latter is also true of primary task measures). Because of this, the decision was made to exclude them from the current review and focus on quantitative workload measures that can be validation empirically.

3.1 Rating Scales

It seems appropriate that mental workload be measured by subjective means, as it is a psychological construct. Jex (1988) states "In the absence of any single objective measure of the diffuse metacontroller's activity, the fundamental measure, against which all objective measures must be calibrated, is the individual's subjective workload evaluation in each task." Casali and Wierwille (1984, p. 1046) note that their findings and other's indicate "… that properly designed rating scales with associated instructions … are particularly sensitive measurement instruments, especially with highly-trained populations …". Gopher and Donchin (1986, p. 41-2), however, assert "… an operator is often an unreliable and invalid measuring instrument." Nevertheless, subjective measures such as rating scales have a long history for measuring feelings of workload, effort, mood, fatigue, etc. Subjective methods attempt to quantify the personal interpretations and judgements of their experienced demand.

Most subjective workload measures imply (if they do not explicitly state) that it is mental workload that is being measured and the effects of physical work associated with gross motor muscles are not considered. The NASA TLX technique discussed below does have a category for Physical Demand that could capture the demands associated with physical labour although the wording describing this dimension seems more directed towards fine motor skills. Other subjective and physiological scales exist to measure physical labour such as the Borg's[5] subjective scale of relative effort or oxygen consumption (VO_2) as a measure of metabolic rate. It seams reasonable that experiments could be devised to test the independence of various subjective mental workload measures from physical exertion.

The repeatability and validity of such quantitative subjective techniques are sometimes uncertain and data manipulations are often questioned as being inappropriate. While the ordinal nature of the ratings is seldom questioned, use of interval or ratio arithmetic operations on the data has been the topic of much debate, with no definitive outcome in sight (Annett 2002). The issue is one of both philosophy and pragmatism. In many cases, there is no evidence that the data are anything but ordinal and as such, not amenable to arithmetic manipulation or parametric statistics, yet these are the everyday tools one is taught to use in science and engineering in order to support the conclusions and decisions made. On the other hand, subjective workload rating data may very well be interval – there is insufficient evidence to support or contradict this position. Other tools exist that could be used to explore this problem, such as non-parametric statistics, yet they have not found favour in practice. The typical approach seems to be to ignore possible violations of mathematical axioms in favour of convenience, accepting the risk that conclusions may not be justified given the data used.

---

[5] Borg, G. (1977) Simple rating methods for estimations of perceived exertion. Proceedings of the First International Symposium on Physical Work and Effort. Wenner-Gren Center, Stockholm.

To address this issue, non-parametric statistical analysis approaches are recommended over parametric statistics in such cases where the data should be considered ordinal. If this is undesirable, then parametric analyses should be validated through additional non-parametric analyses. This approach requires additional forethought on test designs to ensure appropriate plans for the questions to be answered since the analysis procedures are typically not as powerful. A mathematical assessment of the hazards of using parametric analyses with subjective methods that use ratings such as 5 or 7 point Likert scales would be useful guidance or even case-by-case assessment of conclusions based on parametric analysis compared to a corresponding non-parametric analysis. Alternative methods such as fuzzy logic may skirt the issue, providing a calculus to combine subjective data to form conclusions, although validation is still a difficult issue to resolve.

Meshkati and Lowewinthal (1988, p. 257) note many researchers feel that "… unless the subjective measurement technique is properly structured, they may serve only as gross indicators of stress level and have little diagnostic value, i.e., they may not indicate the source or type of workload involved." Meshkati, Hancock et al. (1995, p. 756) go further and note that some researchers believe that subjective feeling of difficulty is essentially dependent on the time stress involved in performing the task *for time-stressed tasks only*. This does not seem to address the source of workload in self-paced situations where time stress is low, but the number of tasks or task complexity is high, resulting in a sense of being overwhelmed by the reasoning or planning process required to complete the task.

While subjective measures have high face validity, their interpretation and ability to predict performance is uncertain. Vidulich (1988), Yeh and Wickens (1988) describe numerous instances where a dissociation between subjective and performance measures has been found. Vidulich concludes that subjective measures tend to be more sensitive to processing load in working memory while having low sensitivity to response execution demands. The hypothesis is that subjective workload is only sensitive to manipulations that are well represented consciously, so that varying demands in skill based tasks (or alternatively subject disinterest and inattentiveness) will not change in subjective ratings substantially. This suggests that subjective measures are highly suited to assessments of modern technologies that aid judgement and decision making, but are less suited to assessing physical or mechanical aids for repetitive or highly learned tasks. Others (Brookhuis and de Waard 2002) embrace dissociation as a natural constraint on all measures and focus on the picture presented by an assortment of measures that are differentially sensitive to the array of factors that contribute to workload.

Wierwille (1988, p. 320) suggests analytical techniques that average workload over time are inappropriate, despite indications that this is what subjective measures seem to be capturing. Instead, Wierwille argues momentary workload values represent the appropriate measure, particularly for design analyses. This suggests that the usual subjective measures on their own are insufficient to adequately characterize workload and some additional means of capturing instantaneous or momentary work overload is necessary. This seems to support more analytical, workload estimation approaches for applied assessments.

Self assessments involve rating demands on numerical or graphical scales, typically anchored either at one or two extrema per scale. Some subjective techniques use scales that are categorical, with definitions at every level, such as the Modified Cooper-Harper scale. Other techniques use an open-ended rating with a "standard" reference task as an anchor and subjects rate other tasks relative to the reference task. Hart and Wickens (1990) subdivide rating scale methods into uni-dimensional ratings such as Overall Workload (Vidulich and Tsang 1987), hierarchical ratings such as Modified Cooper-Harper or Bedford scales (Wierwille and Casali 1983), and multidimensional ratings such as SWAT (Reid and Nygren 1988) and NASA TLX (Hart and Staveland 1988).

Unidimensional and hierarchical measures mentioned above have good face validity, are easy to understand and use, and generally have good user acceptance. While useful in their own right, they also may serve as standards to assess more complex, multidimensional measures discussed below. What they lack is a calculus to combine ratings for predicting workload in different situations involving similar tasks. While not all multi-dimensional workload scales have a predictive mode, several do. The rest of this review of subjective measures will concentrate on measures that have, or could have, predictive capability through constructive modelling.

The VACP method (Visual, Auditory, Cognitive, Psychomotor: Aldrich and McCracken 1984; McCracken and Aldrich 1984; Aldrich, Szabo et al. 1989) is an early attempt at a simple, diagnostic workload measurement tool that, when coupled with task network modelling, can be used for predictive workload assessment. Subjects assess task demands according to a standardized, categorical list in each of the four dimensions. Each dimension's levels were assessed to create interval rankings that could be used in predictive, constructive simulations. A simple summation process was proposed with a rather arbitrary limit placed on the maximum value any dimension could attain before operator overload would occur. Despite objections to the lack of validation and inappropriate aggregation of ordinal data, it has proven useful in system design to identify where systems might over-burden the users. The VACP method remains a frequently used metric and can be used as demand-estimates in more complex workload prediction models.

W/Index (Workload Index: North and Riley 1989) was developed to formalize Wickens' concept of contention among multiple resources in workload calculations. VACP ratings of individual task demands have been suggested as input data to W/Index. While possibly an improvement on the VACP method for predicting workload, W/Index did not have an overload criterion. Further, neither VACP, W/Index nor the analysis tools that used them had mechanisms to reallocate or delay tasks to form coping strategies as human operators do, resulting in unrealistic predictions of workload.

Boles and Adair (2001) attempt to assess the cause of workload through the Multiple Resource Questionnaire (MRQ) and suggest that this method correlates well with other subjective methods. Whether MRQ is a better method, or a practical workload measurement technique, remains to be seen, but linking MRQ with hierarchical measures such as the Modified Cooper Harper (MCH) scale may provide analysts with the power they require: an easy to use, validated workload measurement that has diagnostic support.

Of all the subjective techniques, SWAT seems to be the most common technique reported in the literature (see Appendix 1). Multidimensional or multi-scale assessment techniques often have aggregation procedures to produce an overall workload rating. SWAT seems to be alone in proposing an aggregation procedure that has a sound metrological basis, conjoint scaling, although NASA TLX also makes some claim to aggregation legitimacy with a paired-comparison weighting scheme. Other workload estimation techniques, such as VACP (Aldrich, Szabo et al. 1989) and W/Index (North and Riley 1989) make cruder assumptions about aggregation, although perhaps, an approximation appropriate for the level of precision incurred with subjective assessment techniques.

Reid and Nygren (1988) describe the SWAT technique and the theoretical assumptions underlying it. They note that "... performance measures cannot, of themselves, describe workload ..." because operators may vary effort to maintain a constant performance level, but the perceived workload involved will vary commensurately with effort. SWAT rates experiences on three dimensions (Time Load, Mental Effort, and Psychological Stress), each with three integer or categorical levels. SWAT addresses the principal complaint about subjective measures: the guarantee that ratings are interval or ratio scaled and not just ordinal rankings. This is accomplished through conjoint measurement and scaling of the subjective ratings provided by the subjects (that may be only ordinal before scaling) through a comparison of the relative importance of all levels over all three dimensions (27 distinct comparisons of each level and dimension with every other combination).

Hart and Staveland (1988) describe the NASA Task Load Index (TLX) workload measurement technique and present the empirical validation supporting it. They assert that "... subjective ratings may come closest to tapping the essence of mental workload and provide the most generally valid and sensitive indicator." They also note that subjective techniques are subject to operator biases and preconceptions, and can only be based on what is remembered or abstracted from an experience. They suggest, however, that subjects are unlikely to remember experiences sufficiently well, making absolute judgements of workload or even relative measures for different tasks meaningless. This is a bit at odds with some of the literature that shows assessments are reasonably stable when elicited by detailed after-action review, but ties in with Colle and Reid's (1998) concept that context plays a significant role in workload measurement.

There are two concerns with the formal NASA TLX method. One is the process of scoring "Own Performance"; the other is the scaling procedure based on the paired comparisons. The first is a concern of how the scales are presented. Each dimension is presented as a graduated, horizontal line, anchored at each end. Hart and Staveland (1988) validated the dimensions through an extensive series of experiments and analyses, developing the anchors such that the rating on each scale corresponds linearly to their contributions to overall workload. In 5 of the 6 scales, the ratings range from Low on the left to High on the right; this is reversed for Own Performance, which goes from Excellent to Poor. While this makes sense if one considers each dimension as a contributor to workload, the Own Performance scale can cause confusion among subjects. Subjects may tend not to think of the scales as contributors to workload so much as independent measures for the trial. Periodically reversing the direction of successive scales is advocated by experts in scale development to reduce the tendency of subjects filling them in with little consideration of their meaning. This was not the purpose of the scale reversal in the NASA TLX, where the rating scales were arranged such that each dimension contributed positively to the overall workload score. It is a trivial process to reverse the Own Performance scale presentation, but the implications of this change on the validation of the method have not been assessed. Castor (2003, p. 36) notes that the Own Performance dimension does not need to be part of the workload assessment portion of the NASA TLX algorithm as multidimensional scaling suggests Own Performance is quite separate from the other five dimensions that seem to cluster on a single factor.

The second concern is how NASA TLX aggregates ratings by summing weighted ratings from each scale. The weights are determined by a paired comparison of the relative importance of each scale to workload. In this process, the lowest ranked scale can receive a weighting of 0; in other words, it may not contribute to the computed composite workload measure. Hart and Staveland (1988) imply that the workload contribution weights associated for each of the scales are context dependent and so, such an occurrence is appropriate. If it is indeed the case that only relative workload estimates within the same tasks are feasible, this is not a problem. If, however, one wishes to compare different tasks or experiences, then allowing the weights to vary between experiences significantly complicates the comparison. It seems sensible that different tasks will produce different scale ratings, and that different subjects may perceive the importance of each scale differently, subjectively combining each scale to result in personal assessment of the workload. It does not follow that subjects will change their perception of the contribution of each scale to workload and it seems more sensible that the weights would be largely invariant with task type for an individual. No studies were found that examine how subjects' assessments of scale weights change with context, although Hart and Staveland (1988) suggest that the within-subject weights are largely stable across studies, which suggests that studies of the stability of weights versus individual differences would be feasible.

Colle and Reid (1998) assessed SWAT and NASA TLX scales in a series of experiments, finding SWAT somewhat more sensitive to the various experimental manipulations than was NASA TLX. The pattern of results was similar between the two methods and the differences were too small to state that one technique was better than the other. This is at odds with some other studies that found NASA TLX more sensitive, particularly at low workload levels (Hart and Staveland 1988; Nygren 1991; Hill, Iavecchia et al. 1992). Byers, Bittner et al. (1988) and Hill, Iavecchia et al. (1992) found NASA TLX somewhat more sensitive than SWAT to the experimental manipulations, however, Hart and Wickens (1990, p267) also report that of several subjective measures, NASA TLX proved to correlate best with performance measures while displaying low intersubject variability and good user acceptance. Conversely, Whitaker, Hohne et al. (1997) found SWAT more sensitive than NASA TLX. Both NASA TLX and SWAT are usually reported to be more sensitive than the unidimensional scales such as Overall Workload (OW) and Modified Cooper-Harper (MCH) although the early work by Wierwille and colleagues found the MCH method more sensitive. The small number of levels within each SWAT dimension has been criticized as not providing adequate sensitivity; more dimensions would make the tedious card-sorting process impracticable. While NASA TLX offers greater precision within each subscale, it is difficult to say that it offers greater diagnosticity or greater accuracy, if such a concept is applicable here. Obviously, there is room for improvement, debate on measurement procedures and the interpretation of results.

NASA TLX seems to have higher user acceptability than SWAT because of the shorter scale development phase. The card-sorting procedure of the SWAT method seems to be a significant factor in lower acceptance by users; subjects find even the paired comparison procedure of NASA TLX an imposition, despite taking only a few minutes to complete. Further, it is unclear whether the extra burden of performing scaling procedures for aggregation are of much value (Hendy, Hamilton et al. 1993), It is likely that simple averaging of all scales (if even that is mathematically permissible) would tend to overestimate the true workload, since some of the scales may overlap in their assessment. Some practitioners have found the SWAT scale development somewhat onerous and several have suggested alternatives that they claim are as sensitive if not more sensitive, although lacking the mathematical rigour of the original SWAT. Luximon and Goonetilleke (2001) explored 5 variants on the SWAT method, finding that the simpler approaches were more sensitive and less time consuming to perform, although they lacked the mathematical rigour of the standard SWAT. Biers and Masline (1987; Biers 1995) have explored alternative formulations that, in the limited studies conducted, performed as well as SWAT, but with less work.

Should SWAT, or other ratings based on conjoint scaling, prove too onerous for practical use, it may nevertheless serve as a scale-development standard; that is, the more onerous method may provide an error estimate associated with approximate methods. For instance, the 6-scale NASA TLX procedure, with 10 or 20-points per scale (potentially infinite), would be prohibitively time consuming for practical conjoint methods. Subjects might be induced to willingly suffer a large, tedious, card sorting process (or a subset thereof) if there was high likelihood of developing a universal scale at the end, but if the process proves to be individually specific, there may be little advantage to this approach. The similarity between SWAT and NASA TLX empirical results would seem to suggest that there is hope for interval or ratio data resulting from subjective ratings, which has been a problem for predictive models of workload (Reid and Nygren 1988)[6].

---

[6] Other SWAT references of interest:

Reid, G. B., C. A. Shingledecker, et al. (1981). *Application of conjoint measurement to workload scale development*. Proceedings of the Human Factors Society – 25th Annual Meeting, Rochester, New York.

Reid, G. B., C. A. Shingledecker, et al. (1981). *Development of multidimensional subjective measures of workload*. Conference on Cybernetics and Society sponsored by IEEE Systems, Man and Cybernetics Society, Atlanta, Georgia.

Reid, G. B., F. T. Eggemeier, et al. (1982). *An individual differences approach to SWAT scale development.* Proceedings of the Human Factors Society – 26th Annual Meeting.

Choosing between the methods is largely a matter of preference: if an interval scale is desired, then SWAT is preferred; if ease of use is desired, while likely maintaining a close correlation to SWAT, then NASA TLX seems a viable option.

While the NASA TLX method can be performed with paper and pencil, computerized versions have also been developed. The original DOS™ version may still be available from NASA or from the HSIAC website[7], however, a more recent implementation for Windows™ has been developed[8] to support experimental investigations at DRDC Toronto. This DRDC version can use the original NASA TLX format or a variant that include a univariate Overall Workload rating (not validated) as well as reversing the Own Performance scale. Experimental designs can be specified and the results exported as comma-separated text files that can be imported into spreadsheet programs for further data analysis.

While many subjective methods may be useful for assessing the workload associated with a task, job or experience, most are not useful for predicting the workload associated with combinations of tasks or activities. It seems reasonable that such a predictive version could be developed for some of the methods described above. Prediction using the VACP and W/Index methods were early such attempts, however, there is little theoretical or practical evidence that the approaches used were valid.

The DRAWS (Defence Research Agency Workload Scale) measurement technique (Farmer, Belyavin et al. 1995; Farmer, Jordan et al. 1995) asks subjects to rate their perception of the Input, Central, and Output task demands. DRAWS can be used for assessing single tasks, or for assessing experiences with multiple, concurrent tasks. There is a fourth category called time pressure that is also rated. The scale is nominally from zero (no load) to 100 (fully loaded), although subjects are permitted to record values greater than 100. The DRAWS ratings are thought to represent the time pressure associated with each stage of Input, Central and Output processing of the task. No information concerning validation of the scale was found, however, the POP (Prediction of Operator Performance) model was developed as a predictive form of DRAWS. POP integrates subjective DRAWS ratings of individual task demands when those tasks are performed in together by the operator. POP uses a Markov process to model the interference for contending tasks, and aggregates the individual task DRAWS ratings into an overall workload or time pressure for the operator. The POP model has been validated against a selection of laboratory studies with good predictive ability of the multitask DRAWS ratings; it has not been validated against field studies. This suggests that DRAWS ratings and the POP model make a suitable measurement and prediction scheme, although further validation would be appropriate.

The IP model (Hendy and Farrell 1997) is one of a few methods that claims to relate workload to observable aspects of task execution, in this case time pressure, reflecting individual capabilities and task demands as a single, measurable quantity (processing time required divided by the environmentally imposed time available). The IP model does not measure workload itself, although it postulates that workload equates to time pressure and error rates are relationships that depend on time pressure alone. Although the IP model can be considered a uni-dimensional scale, it considers other factors usually associated with workload (such as strategy selection and individual differences) and uses these factors to moderate time pressure, either by increasing the execution time or by decreasing the amount of time available. Validation of the IP model has been limited to studies on a simplified air traffic control task, although its assessment on tasks similar to the POP validation tasks is planned.

---

[7] DOS version of NASA TLX: http://iac.dtic.mil/hsiac/products.htm

[8] Windows version of NASA TLX: For information contact Brad Cain, brad.cain@drdc-rddc.gc.ca.

The POP and IP models have similar fundamental assumptions, and are being integrated into a computational model for predictive analysis, taking advantage of the strengths of each[9]; this product, POPIP, is being developed within the Integrated Performance Modelling Environment (IPME), although its validity has to be established. If this integration proves successful, the combination of DRAWS and POPIP should provide a useful means of measuring, modelling and predicting workload.

3.2 Performance Measures

Performance measures of workload can be classified into two major types: primary task measures and secondary task measures. In most investigations, performance of the primary task will always be of interest as its generalization to in-service performance is central to the study. In secondary task methods, performance of the secondary task itself may have no practical importance and serves only to load or measure the load of the operator. As Lysaght, Hill et al. (1989, p. 67) point out, "A statement about operator performance is meaningless unless system performance is also acceptable. Accordingly, there is a need to measure both." Thus, there is a necessary precondition that system performance be acceptable; operator workload is not a sufficient measure for assessments.

In order to have primary task measures that are reliable, tests must have appropriate context, relevance, representation, and time-on-task training. Despite the relevance of the primary task to operational activities, it is often not possible to assess the cost of performing the primary task by performance measures alone because of changes in "strategic reallocation of mental capacity". Wilson (2004) notes, "Because of the protective (compensatory) effect of increased effort, it is clear that measuring performance is not sufficient to assess the state of the operator. The level of performance does not provide information about the costs involved in the adaptive response to stress. Under conditions where there is no discernible breakdown of performance under stress, physiological and subjective measures of operator functional state mainly reflect the amount of mental effort (strain) required to maintain task performance." Thus, while primary task measures may be considered a necessary measure of workload, they should not be considered sufficient on their own. This is supported by the apparent dissociation of performance and demand noted in the previous section.

Although operational performance measures are easy to justify, they often lack scientific rigour, making interpretation of the results difficult. Uncontrolled and perhaps unknown factors may dominate results rather than the intended manipulations in the trials. Conversely, laboratory tasks provide more experimental control, but lack the ecological validity of operational task measures. A combination of experimental and operational assessment is often the best approach. Advances in simulator technologies are creating experiences with a greater sense of presence such that simulators should become increasing accurate estimates of operational performance before real world measurements are undertaken. Nevertheless, task performance measures are key for postulating predictive models based on other operator-state factors that can be evaluated in constructive simulations with many replications; virtual simulations are typically restricted to a few replications and so can consider commensurately fewer conditions.

Primary task measures attempt to assess the operator's performance on the task of interest directly, and this is useful where the demands exceed the operator's capacity such that performance degrades from baseline or ideal levels. Speed, accuracy, reaction or response times, and error rates are often used to assess primary task performance. Primary tasks measures are thought to be "... globally-sensitive and provide an index of variations in load across a variety of operator information processing functions" (Eggemeier, Wilson et al. 1991, p. 209).

---

[9] Dr. A.J. Belyavin, QinetiQ, Plc., Farnborough, Hants, UK. Personal communication.

Wickens (1992, p. 392) notes that primary task measures may not be sufficient or adequate:

1) If the variability in the task demands are insufficient to result in observable primary task performance changes (no information on remaining capacity can be inferred);
2) If alternative primary tasks use different modalities (it may be the reserve capacity available for other secondary-tasks that is important to determine);
3) If the primary task demands cause incidental effects such as fatigue or stress that become important performance shaping factors in longer exposures (short evaluations may not show a difference between contending designs); and
4) If other factors, such as strategy, affect performance and workload differently (giving rise to a perception of dissociation between the two).

Hart and Wickens (1990) note that while primary task measures are important in workload assessment, they are more a measure of what the system can achieve rather than an estimate of the cost of operator achievement and that a dissociation between workload and primary task performance is frequently observed (Yeh and Wickens 1988).

Secondary task measures provide an index of the remaining operator capacity while performing primary tasks, and are more diagnostic than primary task measures alone. The characteristics of the secondary task are used to infer the interaction between the primary and secondary tasks and this approach is frequently used when the operator can adapt to demand manipulations such that primary-task performance is apparently unaffected. The secondary-task paradigm can be further classified into Auxiliary Task and Loading Task methodologies, but the intent of both is to increase operator load to the point where changes in effort or strategy are no-longer able to compensate for changes in the demand manipulation.

In auxiliary task methods (the more common of the secondary task approaches), operators are instructed to maintain consistent performance on the primary task regardless of the difficulty of the overall task. The variation of performance on the secondary auxiliary-task is measured as an indicator of the operator's reserve capacity, serving as a surrogate workload measurement under the various loading conditions.

The loading task approach deliberately causes degradation of the primary task, requiring consistent performance on the secondary task. This shifts the primary task performance into a region where it is sensitive to the demand manipulation. The performance decrement of the primary task is measured as the loading task difficulty is increased.

Whether a non-intrusive (auxiliary) or an intrusive (loading) secondary task approach is adopted, there is still a multitasking issue that should be considered. Williges and Wierwille (1979, p. 558) note that subjects probably expend more mental effort during dual task performance than the sum of the effort required to perform each task alone, even if the tasks do not interfere. This overhead is attributed to the management and scheduling of the two tasks by some form of metacontroller within the cognitive system. This hypothesis is difficult to determine without a formal psychological model of the workload process. A major premise of these dual or multitask environments is that the workload is inherently different from a single task condition, regardless of the single task level of difficulty. Fracker and Wickens (1989) argue against that generality, noting in some cases, dual-task performance cannot be distinguished from a more complex single task. They do not consider the possibility that a single, complex task might be treated as a collection of task elements, and processing of these elements in unison requires coordination. The POPIP model (noted elsewhere in this text), if successful, may present an opportunity to test these hypotheses formally in a manner that can be validated empirically.

Colle and Reid (1999) note that secondary task (and presumably primary task) performance measures, while of conceptual and theoretical interest, are less practical than subjective measures, particularly in operational assessments. They note that globally sensitive secondary task measures need to be developed, inferring that secondary tasks can be of use in more diagnostic applications. Unfortunately, secondary tasks based on rigid, laboratory tasks are often too constrained or contrived. This can lead to measurements that do not adequately reflect the operator's ability to dynamically shift tasks to accomplish them in a timely manner. An aspect of performance measures that makes them awkward is their inherent lack of generalization, although sometimes extrapolation from one domain to another seems plausible. Different performance measures are often required for specific primary tasks in different applications, making standardization across domains difficult (Meshkati and Lowewinthal 1988).

Selection of secondary tasks must not be done "will-he, nil-he"; there is a need to match the secondary tasks to the primary tasks such that the operator is loaded appropriately and context-specific sensitivity is captured (Wickens 1977). The choice of secondary task can a profound effect on performance and, hence, on the outcome of the experiment. Damos (1991) suggests that a number of considerations for selecting secondary tasks and gives further references to other reviews. In particular, Damos (1991, p. 114) notes that practice on each task alone is not sufficient to ensure optimal dual task performance; the tasks must be also practiced within the context of the dual task experience, an aspect related to the metacontroller's role described by Jex (1988). Important features of the dual task training are the appropriateness and timeliness of feedback to guide the subject's selection of strategies.

Meshkati, Hancock et al. (1995, p. 754) note that some researchers are concerned that secondary task measures might produce an undesired change of strategy, distorting performance on the primary task by affecting strategies. Since performance on either the primary task or the secondary task must be held constant over all manipulation levels for this technique to be useful, an embedded secondary task (one that is a normal component of the operator's responsibilities, albeit of a lower priority than the primary task) is more appropriate, gaining both operator acceptance and ecological validity. Using artificial secondary task methods in operational or even training assessments presents problems because the intrusiveness of the secondary task into the primary task may present safety hazards or adversely affect the training objectives (Meshkati and Lowewinthal 1988). Presumably, undesirable intrusion is less of an issue when the secondary task is a natural, embedded task that is part of the operator's normal routine. This leads to an ecologically valid, secondary task measure that often gains greater operator acceptance.

Many of the workload reviews examined discuss specific secondary tasks, however, none included a table of tasks that described the key secondary task characteristics and identified families of primary tasks with which they would be appropriate to pair. Secondary task selection criteria that were identified in these reviews include:

1) Non-interference with primary task (consume similar resources, but not interact with the primary task);
2) Easily learned; and
3) Self paced (easily interrupted or delayed).

Typical variables for secondary task measures are:

1) Reaction time;
2) Time estimation variance;
3) Accuracy and response time (to mental arithmetic or memory search);

4) Signal detection rates;
5) Tracking performance (such as RMS error or control reversals);
6) Number of concurrent tasks in an interval; and
7) Percentage of time occupied.

Secondary tasks include, but are no means limited to:

1) Rhythmic tapping;
2) Random number generation;
3) Probe reaction time (Sternberg memory search task, Bakan task);
4) Verbal shadowing;
5) Spatial reasoning;
6) Time estimation and time production (retrospective estimate of elapsed time);
7) Critical instability tracking task; and
8) Compensatory or pursuit tracking tasks.

There is a major caveat with current dual or multitask workload measurement paradigms: most give little or no thought to formal, measurement theory when evaluating secondary task workload measurement (Colle and Reid 1997). While other fields of psychology and human performance have embraced formal methods, the performance oriented mental workload community seems to be remiss. The definition of mental workload equivalency curves to characterize the demands of tasks appears to be lacking in the performance measurement approach to workload assessment.

3.3 Psychophysiological Measures

"The goal of psychophysiological applications in the assessment of mental workload is to develop measures with well known properties that can be applied in specific situations. This goal has come about from the complex nature of the mental workload construct and the acceptance that there is no golden yardstick of mental workload" (Neuman 2002, p. 601). Much of the psychophysiological literature focuses on determining the aspects of workload that particular methods are sensitive to. While sensitivity and relevance are obviously important factors in selecting any method, they seem more so with physiological measures because these measures tend to be general, systemic indicators of stress. Lysaght, Hill et al. (1989, p. 137) note that "... the use of an inappropriate technique may be misleading; it may be a good technique in some instances, but (may be) the wrong tool for the question at hand."

The principal attractions of psychophysiological measures are continual and objective measurement of operator state. Psychophysiology attempts to interpret the psychological processes through their effect on the body state, rather than through task performance or perceptual ratings. If successful, this approach would have a number of advantageous applications, however, as Wickens (1992, p. 199) notes, psychophysiological measures are "... one conceptual step removed from the inference that the system designers would like to make." This requirement to infer workload is an issue both for researchers seeking to assess workload as well as for designers of automated-support systems that attempt to assess operator state and provide assistance accordingly. Meshkati, Hancock et al. (1995, p. 757) suggest that "… physiological methods do not measure the imposed load, but rather they give information concerning how the individuals themselves respond to the load and, in particular, whether they are able to cope with it." Psychophysiological measurements may be particularly useful when subjective methods or performance measures become insensitive due to covert changes in operator strategies, or the applied level of effort lead to an apparent dissociation among subjective and performance measures.

A requirement of most psychophysiological measures is for reference data that establishes the operator's unstressed background state. Such background states are subject to many factors and may change markedly over time so an operational baseline state is often used. The operational baseline state is measured when the subject is not under any specific stress, but it will reflect systemic stresses incurred from being "in-theatre" or reflect the day-to-day changes in a subject's life. Thus, as with subjective and performance measures, contextual issues should be considered in the evaluation of the results. Further, the baseline values as well as the operational values may vary considerably between individuals, so psychophysiological-measurement systems often need to be tailored to each individual rather than using group norms, making interpretation more involved. Another practical consideration is the availability of a suitable test environment that is related in part to context. Physiological measures are of little use early in systems design since it is unlikely there will be a physical mock up or simulator to provide the appropriate stimuli (Mitchell 2000).

In the past, physiological measures often entailed cumbersome, invasive equipment, unsuitable for most applied settings. This has changed dramatically in the past decade as advances in technology have made the equipment much more portable and capable. There is still a significant degree of invasiveness with some techniques that users may object to, making in-service use awkward. There should be a clear advantage demonstrated to operational personnel expected to use these procedures if there is any serious hope that operational users will endure the invasive measurements.

While most of the negative issues associated with psychophysiological measures from the user's perspective are technological (and hence susceptible to improvements in hardware and methods), "… the lack of a strong conceptual link from the physiological measures to performance …" is its greatest weakness from the analyst's perspective (Kramer 1991). Wilson and Eggemeier (1991, p. 351) reiterate the need for a better understanding of the links among the various physiological responses and workload; this would no doubt be well served by formal measurement theory. Wilson and O'Donnell (1988) lament the failures of early attempts to find the essential link between objective, non-intrusive, physiological measures and mental workload. They note that attempts to use individual physiological measures, while being sensitive to specific demands, were unlikely to be generally sensitive because of the multi-facetted nature of workload. Instead, they suggest a battery of psychophysiological measures would be necessary, along with a suitable interpretation scheme. Wilson and O'Donnell critically review several of the more common psychophysiological measures and describe their Neurophysiological Workload Test Battery (NWTB), although it is not known if this is consistent with Colle and Reid's formal measurement theories. Kramer (1991), Wilson and Eggemeier (1991) present critical reviews, although both these assessments are over a decade old.

NATO RTO has recently published a review of operator functional state assessment that includes an overview of psychophysiological techniques (Wilson 2004, Chp4). The report provides a brief description of a number of techniques, assesses the advantages and disadvantages of each technique, and documents the requirements for use. Numerous references are provided for each technique, and this document seems a good starting point for entry into psychophysiological-measurement techniques. Fahrenberg and Wientjes (1999, pp. 111-112) note a number of references that detail measurement techniques for various psychophysiological approaches as well as references to dealing with measurement problems and artefacts. The Society for Psychophysiological Research maintains a web site[10] that contains several guidelines for specific psychophysiological measurement techniques.

---

[10] http://unix.wlu.edu/~spr/

Corwin, Sandry-Garza et al. (1989) studied a number of workload measurement methods in a simulator at NASA Ames for a several commercial airline flight scenarios. In the first study, they found that (pp. 96-99):

1) Subjective measures (NASA TLX and SWAT) demonstrated validity and reliability.
2) Physiological measures were generally disappointing:
    a) Eye movement and eye blink were insensitive to experimental manipulations;
    b) Inter-beat heart rate interval was thought to be reliable and valid, but too prone to other effects; and
    c) Heart rate variability and blood pressure spectral analysis were insensitive to workload manipulations.
3) Primary flight control measures were thought to be good discriminators of workload, but secondary tasks were not good discriminators of workload.

In a second simulator study by the same authors (pp. 159-163), subjective methods again proved valid and reliable. There was some question about whether the simpler, unidimensional Bedford Scale coupled with the Pilot Subjective Evaluation information elicitation technique would not be of greater value for evaluating new aircraft flight deck workload than the more general and data-intensive NASA TLX and SWAT techniques. Again, none of the physiological measures were found to be sensitive to the workload manipulations. In this study, secondary tasks tended to be ignored by the subjects, so only primary flight task (control input) performance was analyzed. The authors felt that the primary-task measure was a useful indicator of workload despite the subjects' tendency to ignore the competing task. In their final report, Corwin, Sandry-Garza et al. (1989) note that workload measurement was still immature, but they recommended the following subjective techniques: Bedford/Modified Cooper-Harper; NASA TLX, SWAT. They also recommended both primary flight control task performance as well as embedded and ecologically appropriate secondary task performance as suitable workload measures for assessing new aircraft. Of the physiological measures studied, only heart rate was suggested as a physiological measure and that was qualified by noting it may be more of a measure of arousal, reflecting stress effects not directly attributable to workload. These results seem typical of evaluations of workload methods. The lack of a clear, positive result for the psychophysiological methods must be disheartening for their advocates.

Most researchers in the field would agree that several psychophysiological measures correlate reasonably well with various aspects of workload and hold promise for objective workload measurement, however, considerable research remains to be done to properly classify and characterize these methods so they can be applied appropriately and assembled into a battery of measurement techniques that could provide a general assessment method. Boucsein and Backs (1999, Table 1.1, p. 9) suggest that a few measures are driven principally by mental strain, although most measures seem sensitive to stress in general rather than just the stress of workload.

Stress related hormones may provide useful, long term measures in either laboratory or field settings, however, they may not be sufficiently sensitive to provide real time interventions (Eggemeier, Wilson et al. 1991). Currently, psychophysiological experts recommend that psychophysiological approaches be applied as a battery of measures to isolate mental workload contributions. The measures selected must be appropriate for the task and the aspect of workload or strain that is of interest; it is not prudent to measure only one or many indiscriminately (Gaillard and Kramer 1999). This is due in part to recognition that the operator's state "... should be regarded as the result of many physiological and psychological processes ..." (Gaillard and Kramer 1999, p. 32). The Boucsein and Backs monograph contains several papers on psychophysiological measurement approaches and the first chapter lists a number of measures with application references and general indicators of the effect (Boucsein and Backs 1999).

Some of the more common psychophysiological methods are noted briefly in the following paragraphs.

3.3.1 Electroencephalography

Measurement of brain activity through electroencephalography (EEG) is used in many fields, not only in workload assessment, and technology is making it practicable for some operational settings. Castor (2003), Boucsein and Backs (1999) note, however, that EEG approaches are prone to artefacts and so have not been used often in field studies. Technical reasons also preclude the use of brain imaging in the field, and probably most human factors laboratories as well. The EEG data are complex waveforms that require sophisticated signal processing equipment. The waveform spectrum is typically divided into a number of frequency bands and workload assessments is made on the power within these bands or on time shifts of event related potentials (ERP).

Freude and Ullsperger (1999) note that movement related readiness potential (BP) and Preparatory Slow Brain Potentials (SP) seem to be complementarily sensitive to attention, demand, and decision making. It is not clear whether the changing amplitudes in these measures can be correlated with workload or performance in a class of tasks to create stable, general predictions of performance changes in practice. Wickens (1992, p. 398) notes that evoked brain potential is better thought of as measure of residual capacity than as a measure of effort, and the reduced P300 amplitude is sensitive to central processing demands, but not response demands, providing an unobtrusive measure of mental workload. Wilson and O'Donnell (1988) note that the P300 amplitude ERP may index the degree of surprise or mismatch of a stimulus with expectation while the P300 latency is more related to the difficulty of a task. Meshkati, Hancock et al. (1995) feel that the Evoked Response Potential family of measures holds the most promise of the physiological measures of workload, despite the technological and interpretation hurdles.

3.3.2 Eye Movement

Measurements of eye activity can be used unobtrusively and much of the technology may already be in place to support these measurements. For example, helmet mounted sighting systems for fighter pilots or proposed helmet mounted display information systems for the infantry may provide a means to obtain these data without further intrusion on the subject. If a stable, helmet mounted display is available for operational reasons, then a number of measures of eye activity can be made unobtrusively, such as: horizontal and vertical eye movement (extent and speed), blink activity (duration, latency and frequency), fixation duration, point of regard and pupil diameter.

Although ocular measures are sensitive to mental demands, they are also sensitive to other factors; in particular, they are sensitive to fatigue. The literature contains contradictory findings that may be due to differences in experimental methods (Sirevaag and Stern 1999). Further, the measurements often require a stable sensor capable of detecting small movements, something difficult to achieve in the field or even in the laboratory at times because of movement of the sensor on the head as the body moves.

Blink measures can be context dependent. Blink rate has been observed to decline with increased workload resulting from processing visual stimuli, however, it has been observed to increase with increased load resulting from memory tasks (Wilson 2004) and the connection between blink rate and workload seems tenuous (Castor 2003). Blink closure duration appears to decrease with increased workload resulting from visual stimuli or gathering data from a wide field of view while blink latency increases with memory and response demands (Castor 2003).

Pupil diameter appears to be sensitive to a number of demands and emotional states, making it less diagnostic, however, the measurements need to be quite precise (on the order of tenths of a millimetre) making application difficult in environments with vibration or requiring considerable eye and head movement. Wilson (2004) notes that pupil diameter generally increases with higher cognitive processing levels and it is sensitive to rapid changes in workload, however, when overload occurs, pupil diameter can become unresponsive to changes or even reverse its response. Nevertheless, research continues (Marshall, Pleydell-Pearce et al. 2002) and a new technique has emerged showing promising results, the Index of Cognitive Activity[11], although no details describing the method or its validation were found.

3.3.3 Heart Rate

Various heart rate measures (such as the rate, its variability, and resulting blood pressure) have been reported to be sensitive to workload. These measures are relatively easy to employ unobtrusively both in the laboratory and in the field. Heart rate measures suffer from interactions with respiration, physical work and emotional strain, and so would likely require unique measures to isolate mental workload contributions. Wilson (2004, p. 4-7) notes that there are numerous coupled control mechanisms and feedback loops in the cardio-vascular system, making definitive interpretation difficult. Meshkati (1988) states that heart rate variability is probably the most used physiological measure in workload measurement and references other literature, noting the varied effectiveness of heart rate variability in workload assessment. Reliable measurement of heart rate and its variability require at least 30 seconds, but not more than 5 minutes for optimal sensitivity with concurrent measurement and correction of respiration effects (Castor 2003).

Mulder, Mulder et al. (1999) note that heart rate measures (particularly heart rate variability in the 0.07 – 0.14 Hz range) are sensitive to task complexity and compensatory effort resulting from stressors (fatigue, noise, etc.), but that cognition and emotion may be too tightly coupled to distinguish effect. Mulder, Mulder et al. (1999, p. 144-145) report that there have been problems with reproducibility of results suggesting more work is required before a comprehensive, formal method can be recommended and work to this end was underway throughout the 1990s. Despite the difficulties associated with the heart rate measures, heart rate variability continues to be studied and commonly used in conjunction with respiratory measures to assess operator state and mental workload.

Blood pressure has been found to correlate with mental demand, however, it does not appear to be very sensitive and it is prone to exercise artefacts. While easily and often measured, blood pressure does not appear to be a principal candidate for workload measurement (Castor 2003).

3.3.4 Respiration

Wilson (2004) notes that respiration is not simply a factor for adjusting heart rate measures; respiration measures offer their own valuable information on operator state. There are several measures that can be recorded, such as the time for inspiration or expiration, the complete cycle time, the volume and flow rate. Several of these measures may be measured or inferred with little intrusion. Respiration rate has been observed to increase while respiration volume decreases as stress and mental workload increase, but it is also highly dependent on physical activity. This suggests that while it provides useful information about operator state, it is not a suitable workload measure on its own (Castor 2003). Nevertheless, because it is a necessary measurement for correcting heart rate measures, it remains a candidate for supplying part of the workload measurement picture.

---

[11] http://www.sci.sdsu.edu/cerf/darpa/darpa.htm

3.3.5 Future Psychophysiological Measure Developments

Formal, coupled models relating various psychophysiological measures and workload need to be developed. Often, physiological phenomena interact with one another such that, although it may be possible to correlate these measures *post hoc*, their independent use as a predictor of workload levels seems quite limited. Nevertheless, this class of methods has potential for unobtrusive, objective measurement of mental workload, particularly for embedded, automated aiding applications, however, few are sufficiently practicable or understood sufficiently for military field operations in their current state of development.

Although bulk and weight have been reduced, the electronic apparatus, sensors and wires associated with psychophysiological methods are seldom acceptable in the workplace, and sometimes not practicable in the laboratory. Modern technology continues to improve devices for a number of different measures and unobtrusive ambulatory recorders with a vast array of sensors are now available that were impracticable a decade ago.

The science of psychophysiological measurement is not static, resulting in better understanding with improved techniques, and advances in technology extending many methods from the laboratory into operational environments more likely. Wilson (2001) reported on several in-flight physiological measures, finding high repeatability among several methods. Some methods did not correlate well with subjective measures; heart rate appeared more sensitive to physical demands of the task rather than mental workload. Others such as electrodermal activity and electroencephalogram measurements showed good correlation with variation in the task cognitive demands while blink rates were found to correlate well with visual demands. Other studies support blink measures, particularly startle eye blink, as being sensitive to workload (Neuman 2002).

Wilson (2004) also reports on several psychophysiological techniques that are being developed for medical applications. Oximetry, Near-Infrared spectroscopy, fMRI (functional magnetic resonance imaging) and stress hormone assessment, while perhaps not practical for human factors or field applications today, may prove useful in the future for workload measurement as technology advances, making these techniques easier to use and less intrusive.

4.0 IMPORTANT REMAINING WORKLOAD MEASUREMENT ISSUES

Regardless of the workload methods selected, formulation of a general theory of workload that can put measurements into context requires many and varied experiments; reports from a single experiment are insufficient (Wierwille 1988, p. 316). It is essential that the results be shown to create a theory that is generalizable. Wierwille notes that individual differences are key features missing from most measurement approaches.

Colle and Reid (1998) note that context has a significant bearing on the measurement of workload, and that this is likely a perceptual rather than judgement issue. This has implications for selecting the range of stimuli, since Colle and Reid indicate that restricted ranges of stimuli can bias the workload rating results. In evaluations with a range of task difficulties at the low end of the difficulty scale, subjects tend to overrate the demands at the high end of this range. Conversely, for a range of tasks at the high end of the demand scale, subjects tend to underrate the demands at the low end of the scale. They conclude that this "... threat to validity ..." necessitates including context as a major consideration in workload experimental design, preferably by presenting a very broad range of stimuli to avoid range bias and by not labelling (anchoring) the measurement scales. Presumably, a similar effect could be accomplished by using doubly-anchored scales in subjective methods where the limiting anchors reflect absolute ratings, but single anchor, relative ratings may also be useful in workload measurement (Vidulich and Tsang 1987).

5.0 RECOMMENDING WORKLOAD MEASURES

When selecting a workload measure, or a battery of measures, the analyst should consider what the objective of the assessment is. If several design options need to be ranked on workload, then perhaps a univariate measure such as an Overall Workload scale is sufficient. If more diagnostic information is required, and this cannot be obtained through interviews, then the NASA TLX measure may be more appropriate. Primary and embedded Secondary task measures relevant to the operational context in which workload measures are desired should also be used. Psychophysiological measures are not recommended for most field analyses at this time; psychophysiological measures require further research to develop formal relationships among the various factors before they will be of use to the general analysis community.

Farmer and Brownson (2003) recommend that a battery of workload measures be selected for simulation-based assessments and provide guidance on such a selection (although the criteria used to select appropriate methods is unclear):

1) Modified Cooper-Harper (MCH);
2) Instantaneous Self Assessment (ISA);
3) Primary and Secondary tasks;
4) Heart Rate;
5) Heart Rate Variability;
6) NASA TLX;
7) Defence Research Agency Workload Scale (DRAWS); and
8) Blink rate.

This does not mean analysts should apply any or all measures from such a list that **might** be useful, but that many should be considered for the insight they can provide. The lack of a formal model relating the various workload measures seriously complicates interpretation. If a shotgun selection of methods is adopted, the analyst might well end up with a bewildering and contradictory set of results. A careful assessment of the task under study and its context is necessary to select an appropriate battery of workload measurement methods. This battery should include at least one objective measure and make use of quantitative subjective assessments (rather than subjective pass/fail ratings).

Wierwille, Rahimi et al. (1985) made the following recommendations for workload measurement techniques based on a series of evaluations.

1) For studies that are predominantly psychomotor (Wierwille and Connor 1983), the recommend the Cooper-Harper scale, the WCI/TE scale[12], and control movements/unit time; two other sensitive, but non-monotonic techniques (time estimation standard deviation and mean pulse rate) could be used to support the other methods, but are not recommended for use alone.

---

[12] See: Donnell, M.L. (1979) The application of decision-analytic techniques to the test and evaluation phase of the acquisition of a major air system: Phase III. Technical Report PR79-6-91. McLean, VA: Decisions and Designs, Inc. – Article not reviewed.

2) In communications studies (Casali and Wierwille 1983), the following methods were found to be sensitive load: Modified Cooper-Harper (MCH), Multi-descriptor Scale[13]; time estimation; pupil diameter; errors of omission; errors of commission; and, communications response time. Of these, the Multidescriptor Scale has been little used and might be advantageously replaced by another, more common subjective scale such as NASA TLX or SWAT.
3) In cognitive (mediation) studies (Wierwille, Rahimi et al. 1985), MCH, WCI/TE, time estimation, fixation fraction (proportion of time on principal display), mean reaction time and mean error rate were judged sufficiently sensitive to be useful, although time estimation was judged to be rather intrusive on the primary task of calculation.
4) Casali and Wierwille (1984) suggested the following measures were sensitive to perceptual loads: Modified Cooper-Harper, Multidescriptor and Workload Compensation Interface/Technical Effectiveness scales; primary tasks (control inputs) and secondary tasks (time estimation variability and rhythmic tapping); respiration rate. They concluded that none of these measurement techniques intruded significantly on the primary task performance, although I would question the user-acceptability of these secondary task measures in many practical applications.

Casper, Shively et al. (1987) created a decision support tool, WC FIELDE (Workload Consultant for Field Evaluations), in the mid 1980s to help researchers select appropriate workload measurement techniques. A web search failed to find much information, although it was referenced on two sites[14] and while it may still be available through HSIAC[15], no additional information on WC FIELDE was found on the NASA web site.

Castor (2003) has presented a method for matching task characteristics to workload measurement methods, but the GARTEUR (Group for Aeronautical Research and Technology in Europe) tool could be elaborated to include more workload methods; similarly, the WC FIELDE tool could be expanded and updated. Lysaght, Hill et al. (1989, pp. 64-65 and Chapter 8) rate a number of techniques according to their sensitivity, cost and diagnosticity, then propose a “matching model” to help researchers select appropriate workload measures. This was to elaborate on the WC FIELDE work, adding an expert system shell and expanding the scope beyond aviation, however, no evidence was found to suggest that this proposal came to fruition in a practical implementation.

The Internet search results (Appendix 1) show a large effort in the psychophysiological arena. Resources such as the Society for Psychophysiological Research are key to providing summary advice as well as supporting evidence for the various methods. This should become a valuable resource for practitioners to keep abreast of developments in psychophysiological methods and how they might be successfully applied to workload measurement in the future.

6.0 CONCLUSION

In the past twenty years, the science of workload measurement has not progressed nearly as far as one might have hoped. Many of the issues and concerns of the early 1980s are with us today. The science is not

---

[13] See: Casali, J.G. (1982) A sensitivity/intrusion comparison of mental workload estimation techniques using a simulated flight task emphasizing perceptual piloting behaviors. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia. – Article not reviewed.

[14] WC FIELDE: http://softwaretechnews.com/stn4-2/stiac.html and http://www.manningaffordability.com/s&tweb/heresource/tool/tool_list.htm

[15] http://www.softwaretechnews.com/stn4-2/hsiac.html

completely static, particularly in the psychophysiological domain, and a critical review of the topic could fill a sizable report. There seems to be more literature appearing frequently and it would be impossible to keep up with all the developments that may be of use unless one specializes in workload measurement.

That mental workload is multidimensional is not seriously challenged today, but whether workload is a scalar or vector has yet to be resolved and may only be relevant to predictive modelling, when the analyst wishes to assess the workload associated with performing two novel tasks together. Attempts to build computational models of human behaviour that are moderated by workload may provide a useful testbed to augment experimental methods attempting to validate proposed measures.

Subjective workload measures that support predictive modelling, such as VACP and DRAWS, usually focus on task demand in multiple channels. When coupled with task duration in simulations, these approaches produce aggregate measures that are sensitive to both task difficulty and time. These results provide diagnostic information of where the high workload is developing in the system and can be used to validate models for other scenarios. Predictive modelling approaches that focus on an overall workload metric such as time pressure, confound task demands with the time available. If task demands and the resulting workload can be characterized by one parameter, then an overall subjective workload measure may be sufficient. In all cases, it is highly desirable to latch predicted objective performance to empirical measurements if suitable performance models are available as a step towards validating the overall simulation.

When measuring workload empirically, the current recommendations are largely the same as twenty years ago: select a variety of workload measurement techniques that seem appropriate to the application and are likely to provide insight; do not select too many redundant measures, as this could produce conflicting results simply by chance. Understanding of the problem under study may require a number of experiments or trials in converging operations to clarify why some the results dissociate among the measures. A number of researchers in the field have created tools to help guide the selection of measurement techniques. An open source, public domain version of these tools, with references to validation data, would be a useful addition to the human factors and research communities.

Based on this review of the literature, psychophysiological measures should not be recommended for applied problems until researchers can develop a formal, unifying theory that explains the interactions of various physiological phenomena and the relationship to workload, despite the recent technological advances made. As a general practice, a global, univariate workload measure is suggested in conjunction with NASA TLX, as well as contextually relevant primary and embedded secondary task measures. SWAT is an alternative to the NASA TLX, although it is more laborious.

7.0 ABBREVIATIONS

BP	Brain readiness Potential	Freude and Ullsperger 1999
DRAWS	Defence Research Agency Workload Scale	Farmer, Belyavin et al. 1995 Farmer, Jordan et al. 1995
DRDC	Defence Research and Development Canada	http://www.drdc-rddc.dnd.ca
EEG	Electroencephalography	
ERP	Event Related Potential or Evoked Response Potential	Wilson and O'Donnell 1988

fMRI	Functional Magnetic Resonance Imaging	
GARTEUR	Group for Aeronautical Research and Technology in Europe	http://www.nlr.nl/public/hosted-sites/garteur/rfc.html
HMD	Helmet/Head Mounted Display	
HRV	Heart Rate Variability	Mulder, Mulder et al. 1999
HSIAC	Human System Information Analysis Center	http://iac.dtic.mil/hsiac/
IP	Information Processing	Hendy and Farrell 1997
IPME	Integrated Performance Modelling Environment	http://maad.com
ISA	Instantaneous Self Assessment	http://www.eurocontrol.int/eec/public/standard_page/1996_note_10.html
MCH	Modified Cooper Harper	
MRQ	Multiple Resource Questionnaire	Boles and Adair 2001
NASA	North American Space Agency	
NASA TLX	NASA Task Load Index	Hart and Staveland 1988
NATO	North Atlantic Treaty Organization	http://www.nato.int/
NATO RTO	NATO Research and Technology Organization	http://www.rta.nato.int/
OW	Overall Workload	
POP	Prediction of Operator Performance	Farmer, Belyavin et al. 1995 Farmer, Jordan et al. 1995
POPIP	Prediction of Operator Performance Information Processing	
SP	Preparatory Slow Brain Potential	Freude and Ullsperger 1999
SWAT	Subjective Workload Assessment Technique	Reid, G. B., C. A. Shingledecker, et al. 1981 Reid, G. B., C. A. Shingledecker, et al. 1981 Reid, G. B., F. T. Eggemeier, et al. 1982
VACP	Visual, Auditory, Cognitive, Psychomotor	Aldrich and McCracken 1984 McCracken and Aldrich 1984 Aldrich, Szabo et al. 1989
WC FIELDE	Workload Consultant for Field Evaluations	Casper, Shively et al. 1987
W/Index	Workload Index	North and Riley 1989

8.0 REFERENCES

Aldrich, T.B. and McCracken, J.H. (1984). A computer analysis to predict crew workload during LHX Scout-Attack Missions Vol. 1. Fort Rucker, Alabama, US Army Research Institute Field Unit, Fort Rucker, Alabama.

Aldrich, T.B., Szabo, S.M., et al. (1989). The development and application of models to predict operator workload during system design. Applications of human performance models to system design. G.R.B. McMillan, D.; Salas, E.; Strub, M.H.; Sutton, R.; van Breda, L. New York, Plenum Press. 2: 65-80.

Annett, J. (2002). “Subjective rating scales: science or art?” Ergonomics 45(14): 966-987.

Biers, D.W. (1995). SWAT: Do we need conjoint measurement. Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting, San Diego, California.

Biers, D.W. and Masline, P.J. (1987). Alternative approaches to analyzing SWAT data. Proceedings of the Human Factors Society 31st Annual Meeting. New York City. The Human Factors Society.

Boff, K.R., Kaufman, L., et al. (1986). Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance., John Wiley and Sons, Inc.

Boles, D.B. and Adair, L.P. (2001). “Validity of the Multiple Resources Questionnaire (MRQ).” submitted to the Human Factors and Ergonomics Society.

Boucsein, W. and Backs, R.W. (1999). Engineering psychophysiology as a discipline: Historical and theoretical aspects. Engineering psychophysiology: issues and applications. R.W. Backs and W. Boucsein. Mahwah, NJ., Lawrence Erlbaum Associates, Inc.: 3-30.

Brookhuis, K.A. and de Waard, D. (2002). “On the assessment of mental workload and other subjective qualifications.” Ergonomics 45(14): 1026-1030.

Byers, C.J., Bittner, A.C.J. et al. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. Proceedings of the Human Factors Society’s 32nd Annual Meeting, Santa Monica, California, USA.

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 emphasizing communications load.” Human Factors 25(6): 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(10): 1033-1050.

Casper, P.A.L., Shively, R.J., et al. (1987). Decision support for workload assessment: Introducing WC FIELDE. Proceedings of the Human Factors Society 31st Annual Meeting, New York City, The Human Factors Society.

Castor, M.C. (2003). GARTEUR Handbook of mental workload measurement, GARTEUR, Group for Aeronautical Research and Technology in Europe, Flight Mechanics Action Group FM AG13: 164.

Colle, H.A. and Reid, G.B. (1997). “A framework for mental workload research and applications using formal measurement theory.” International Journal of Cognitive Ergonomics 1(4): 303-313.

Colle, H.A. and Reid, G.B. (1998). “Context effects in subjective mental workload ratings.” Human Factors 40(4): 591-600.

Colle, H.A. and Reid, G.B. (1999). “Double trade-off curves with different cognitive processing combinations: Testing the cancellation axiom of mental workload measurement theory.” Human Factors 41(1): 35-50.

Corwin, W.H., Sandry-Garza, D.L. et al. (1989). Assessment of crew workload measurement methods, techniques and procedures. Volume II – Guidelines for the use of workload assessment techniques in aircraft certification. Dayton, OH, Wright-Patterson Air Force Base, Wright Research and Development Centre, Cockpit Integration Directorate: 51.

Corwin, W.H., Sandry-Garza, D.L., et al. (1989). Assessment of crew workload measurement methods, techniques and procedures. Volume I – Process, Methods and Results. Dayton, OH, Wright-Patterson Air Force Base, Wright Research and Development Centre, Cockpit Integration Directorate: 237.

Curry, R., Jex, H., et al. (1979). Final report of the control engineering group. Mental Workload: Its theory and measurement. N. Moray. New York, Plenum Press: 235-253.

Damos, D.L. (1991). Dual-task methodology: some common problems. Multiple-task performance. D.L. Damos. London, GB, Taylor & Francis: 101-119.

de Waard, R. (1996). The measurement of driver’s mental workload. Traffic Research Centre (now Centre for Environmental and Traffic Psychology). Heran, NL, University of Groningen: 198.

Derrick, W.L. (1988). “Dimensions of operator workload.” Human Factors 30(1): 95-110.

Eggemeier, F.T., Wilson, G.F., et al. (1991). Workload assessment in multi-task environments. Multiple task performance. D.L. Damos. London, GB, Taylor & Francis, Ltd.: 207-216.

Fahrenberg, J. and Wientjes, C.J.E. (1999). Recording methods in applied environments. Engineering psychophysiology: issues and applications. R.W. Backs and W. Boucsein. Mahwah, NJ., Lawrence Erlbaum Associates, Inc.: 111-135.

Farmer, E. and Brownson, A. (2003). Review of workload measurement, analysis and interpretation methods, European Organisation for the Safety of Air Navigation: 33.

Farmer, E.W., Belyavin, A.J., et al. (1995). Predictive Workload Assessment: Final Report. Farnborough, Hampshire, Defence Research Agency.

Farmer, E.W., Jordan, C.S., et al. (1995). Dimensions of Operator Workload: Final Report. Farnborough, UK, Defence Research Agency: 58.

Fracker, M.L. and Wickens, C.D. (1989). “Resources, confusions, and compatibility in dual-axis tracking: Displays, controls, and dynamics.” Journal of Experimental Psychology: Human Perception and Performance 15(1): 80-96.

Freude, G. and Ullsperger, P. (1999). Slow brain potentials as a measure of effort? Applications in mental workload studies in laboratory settings. Engineering psychophysiology: issues and applications. R.W. Backs and W. Boucsein. Mahwah, NJ., Lawrence Erlbaum Associates, Inc.: 255-267.

Gaillard, A.W.K. (1993). "Comparing the concepts of mental load and stress." Ergonomics 36(9): 991-1005.

Gaillard, A.W.K. and Kramer, A.F. (1999). Theoretical and methodological issues in psychophysiological research. Engineering psychophysiology: issues and applications. R.W. Backs and W. Boucsein. Mahwah, NJ., Lawrence Erlbaum Associates, Inc.: 31-58.

Gopher, D. and Braune, R. (1984). "On the psychophysics of workload: Why bother with subjective measures?" Human Factors 26(5): 519-532.

Gopher, D. and Donchin, E. (1986). Workload – An examination of the concept. Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance. K.R. Boff, L. Kaufman and J.P. Thomas, John Wiley and Sons, Inc: 41-1:41-49.

Hancock, P.A. and Meshkati, N. (1988). Human Mental Workload. Amsterdam, North-Holland (Elsevier Science Publishers B.V.).

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

Hart, S.G. and Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Human Mental Workload. P.A.M. Hancock, N. Amsterdam, North-Holland: 139-183.

Hendy, K.C. and Farrell, P.S.E. (1997). Implementing a model of human information processing in a task network simulation environment. Toronto, Defence and Civil Institute of Environmental Medicine: 110.

Hendy, K.C., Hamilton, K.M., et al. (1993). "Measuring subjective workload: When is one scale better than many." Human Factors 35(4): 579-601.

Hill, S.G., Iavecchia, H.P., et al. (1992). "Comparison of four subjective workload rating scales." Human Factors 34(4): 429-439.

Huey, F.M. and Wickens, C.D. (1993). Workload transition: Implications for individual and team performance. Washington, DC, National Academy Press.

Jex, H.R. (1988). Measuring mental workload: Problems, progress, and promises. Human Mental Workload. P.A. Hancock and N. Meshkati. Amsterdam, NL, Elsevier Science Publishers B.V. (North-Holland): 5-39.

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

Kramer, A.F., Sirevaag, E.J., et al. (1987). "A psychophysiological assessment of operator workload during simulated flight missions." Human Factors 29(2): 145-160.

Luximon, A. and Goonetilleke, R.S. (2001). "Simplified subjective workload assessment technique." Ergonomics 44(3): 229-243.

Lysaght, R.J., Hill, S.G., et al. (1989). Operator workload: comprehensive review and evaluation of operator workload methodologies. Fort Bliss, Texas, U.S. Army Research Institute for the Behavioural and Social Sciences: 262.

Marshall, S.P., Pleydell-Pearce, C.W., et al. (2002). Integrating psychophysiological measures of cognitive workload and eye movements to detect strategy shifts. Proceedings of the 36th Hawaii International Conference on Systems Sciences (HICSS'03), IEEE Computer Society.

McCracken, J.H. and Aldrich, T.B. (1984). Analyses of selected LHX mission functions: Implications for operator workload and system automation goals. Fort Rucker, Alabama, US Army Research Institute Aircrew Performance and Training.

Meshkati, N. (1988). Heart rate variability and mental workload assessment. Human Mental Workload. P.A. Hancock and N. Meshkati. Amsterdam, NL, Elsevier Science Publishers B.V. (North-Holland): 101-115.

Meshkati, N., Hancock, P.A., et al. (1995). Techniques in mental workload assessment. Evaluation of human work. J. Wilson. London, GB, Taylor & Francis, Ltd.

Meshkati, N. and Lowewinthal, A. (1988). An eclectic and critical review of four primary mental workload assessment methods: A guide for developing a comprehensive model. Human Mental Workload. P.A. Hancock and N. Meshkati. Amsterdam, NL, Elsevier Science Publishers B.V. (North-Holland): 251-267.

Mitchell, D.K. (2000). Mental workload and ARL workload modeling tools. Aberdeen Proving Ground, MD, US Army Research Laboratory, Human Research & Engineering Directorate: 35.

Moray, N.E. (1979). Mental workload: Its theory and measurement. New York, Plenum Press.

Mulder, G., Mulder, L.J.M., et al. (1999). A psychophysiological approach to working conditions. Engineering psychophysiology: issues and applications. R.W. Backs and W. Boucsein. Mahwah, NJ., Lawrence Erlbaum Associates, Inc.: 139-159.

Neuman, D.L. (2002). "Effect of varying levels of mental workload on startle eye blink modulation." Ergonomics 45(8): 583-602.

North, R.A. and Riley, V.A. (1989). W/INDEX: A predictive model of operator workload. Applications of human performance models to system design. G.R.B. McMillan, D.; Salas, E.; Strub, M.H.; Sutton, R.; van Breda, L. New York, Plenum Press. 2: 81-90.

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(1): 17-33.

O'Donnell, R.D. and Eggemeier, F.T. (1986). Workload assessment methodology. Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance. K.R. Boff, L. Kaufman and J.P. Thomas, John Wiley and Sons, Inc: 42-1:42-49.

Reid, G.B., Eggemeier, F.T., et al. (1982). An individual differences approach to SWAT scale development. Proceedings of the Human Factors Society – 26th Annual Meeting.

Reid, G.B. and Nygren, T.E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. Human Mental Workload. P.A.M. Hancock, N. Amsterdam, Elsevier Science Publishers B.V. (North-Holland): 185-218.

Reid, G.B., Shingledecker, C.A., et al. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society – 25th Annual Meeting.

Reid, G.B., Shingledecker, C.A., et al. (1981). Development of multidimensional subjective measures of workload. Conference on Cybernetics and Society sponsored by IEEE Systems, Man and Cybernetics Society, Atlanta, Georgia.

Roscoe, A.H., Ellis, G.A., et al. (1978). Assessing pilot workload, NATO AGARD (Advisory Group for Aerospace Research and Development) 7 rue Ancelle, 92200 Neuilly-sur-Seine, France: 84.

Sirevaag, E.J. and Stern, J.A. (1999). Ocular measures of fatigue and cognitive factors. Engineering psychophysiology: issues and applications. R.W. Backs and W. Boucsein. Mahwah, NJ., Lawrence Erlbaum Associates, Inc.: 269-287.

Vidulich, M.A. (1988). The cognitive psychology of subjective mental workload. Human Mental Workload. P.A. Hancock and N. Meshkati. Amsterdam, NL, Elsevier Science Publishers B.V. (North-Holland): 219-229.

Vidulich, M.A. and Tsang, P.S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proceedings of the Human Factors Society – 31st Annual Meeting. New York City. The Human Factors Society.

Whitaker, L.A., Hohne, J., et al. (1997). Assessing cognitive workload metrics for evaluating telecommunication tasks. Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting.

Wickens, C.D. (1977). Measures of workload, stress and secondary tasks. Mental workload: Its theory and measurement. N. Moray. New York, Plenum Press: 70-99.

Wickens, C.D. (1992). Engineering Psychology and Human Performance. New York, HarperCollins Publishers Inc.

Wickens, C.D. and Hollands, J.G. (1999). Engineering Psychology and Human Performance. Upper Saddle River, New Jersey, Prentice Hall.

Wierwille, W.W. (1988). Important remaining issues in mental workload estimation. Human Mental Workload. P.A. Hancock and N. Meshkati. Amsterdam, NL, Elsevier Science Publishers B.V. (North-Holland): 315-333.

Wierwille, W.W. and Casali, J.G. (1983). "A validated rating scale for global mental workload measurement applications." Proceedings of the Human Factors Society – 27th Annual Meeting: 129-133.

Wierwille, W.W. and Connor, S.A. (1983). "Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator." Human Factors 25(1): 1-16.

Wierwille, W.W., Rahimi, M., et al. (1985). "Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity." Human Factors 27(5): 489-502.

Williges, R.C. and Wierwille, W.W. (1979). "Behavioral measures of aircrew mental workload." Human Factors 21(5): 549-574.

Wilson, G.C., et. al. (2004). Operator functional state assessment. Paris, FR, North Atlantic Treaty Organisation (NATO), Research and Technology Organisation (RTO) BP 25, F-92201, Neuilly-sur-Seine Cedex, France: 220.

Wilson, G.F. (2001). "An analysis of mental workload in pilots during flight using multiple psychophysiological measures." International Journal of Aviation Psychology 12(1): 3-18.

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

Wilson, G.F. and O'Donnell, R.D. (1988). Measurement of operator workload with the neuropsychological workload test battery. Human Mental Workload. P.A. Hancock and N. Meshkati. Amsterdam, NL, Elsevier Science Publishers B.V. (North Holland): 63-115.

Yeh, Y. and Wickens, C.D. (1988). "Dissociation of performance and subjective measures of workload." Human Factors 30(1): 111-120.

Appendix 1 – Internet Workload Measurement Technique Search: Hits by Keyword

An Internet search using the GOOGLE search engine was conducted to gauge the frequency of use of various workload techniques. No attempt was made to eliminate duplicate hits.

		Google Search Results	
	General Search Terms > Specific Search Terms	workload	mental OR cognitive workload
Subjective Ratings			
	Activation Scale	8	8
	Bedford Scale	31	23
	Defence Research Agency Workload Scale (DRAWS, DSTL, QinetiQ)	59	27
	Information Processing/Perceptual Control Theory (IP/PCT)	16	14
	Instantaneous Self Assessment of workload (ISA)	14900	2200
	Malvern Capacity Estimate (MACE)	2600	542
	Modified Cooper-Harper (MCH)	2810	742
	Multiple Resources Questionnaire (MRQ)	84	11
	NASA Task Load Index (NASA TLX)	1600	934
	Observer Rating Form	23	17
	Prediction of Operator Performance (POP, DSTL, QinetiQ)	10	17
	Pro-SWAT	3	3
	Quantitative Workload Inventory (QWI)	60	20
	Rating Scale Mental Effort (RSME)	82	65
	Raw TLX (RTLX)	42	35
	Self report	5230	3750
	Subjective Workload Assessment Technique (SWAT)	3660	758
	VACP	50	41
	W/Index	71	25
Performance Measures			
	Dual task	801	712
	Embedded task	35	27
	Primary Task	4340	1940
	Reaction Time (RT)	32600	8250
	Secondary task	864	684
	Subsidiary task	84	65

		Google Search Results	
	General Search Terms > Specific Search Terms	workload	mental OR cognitive workload
Psychophysical Measures			
	Psychophysiological	2060	1800
	Eye movement measures		
	Blink duration	44	43
	Blink latency	9	7
	Blink rate	211	177
	Endogenous eye blinks (EOG)	606	317
	Eye blink	172	141
	Eye fixations	231	195
	Eye movement	2180	1670
	Glissadic saccades	1	1
	Oculographic activity	2	2
	Pupil diameter	198	172
	Saccade duration	11	11
	Saccadic velocity	15	15
	Cardio-vascular/respiratory measures		
	Blood pressure	43100	15800
	Heart Period (HP)	19	17
	Heart rate	23100	8240
	Heart rate variability	1510	676
	Inter-beat-interval (IBI)	530	156
	Respiration	6460	2770
	Respiratory Sinus Arrhythmia (RSA)	42	38
	Stress-related hormone measures		
	Adrenaline	5500	2060
	Catecholamines	2250	575
	Cortisol	2850	1770
	Epinephrine	2570	730
	Noradrenaline	962	368
	Pprolactin	576	325
	Vanillylmandelic acid	15	5

		Google Search Results	
	General Search Terms > Specific Search Terms	workload	mental OR cognitive workload
Psychophysical Measures			
	Electrical biosignals		
	Autonomic nervous system (ANS)	9390	2720
	Central nervous system (CNS)	9260	4020
	Electrocardiogram (ECG)	12800	4270
	Electrodermal activity (EDA)	3930	498
	Electroencephalogram (EEG)	4320	2900
	Electromyogram EMG	3540	1660
	Event related potentials	409	379
	Evoked cortical brain potential	2	2
	Evoked potential	361	275
	P300 amplitude	59	59
	P300 latency	36	36
	Parasympathetic nervous system	216	98
	Peripheral nervous system (PNS)	868	250
	Skin conduction response (SCR)	4960	698
	Skin resistance level SRL	2320	291
	Skin resistance response SRR	589	95
	Somatic nervous system	30	18
	Speaking fundamental frequency	4	3
	Speaking rate	93	69
	Sympathetic nervous systems (SNS)	1550	596
	Vocal intensity	11	9

Simulator Sickness Research Summary[1]

David M. Johnson
U.S. Army Research Institute for the Behavioral and Social Science
Ft. Rucker, Alabama
USA

Simulator Sickness (SS) is a form of Motion Sickness (MS) that does not require true motion – but does require a wide field of view (FOV) visual display [5, 46, 64]. Like all varieties of MS, an intact vestibular system is necessary to experience SS [12]. It has been called visually induced motion sickness [3, 52, 48] and Cinerama sickness [3, 5, 52]. The term "vection" is used to describe a visually induced sense of self-motion. Vection is "… produced by the nearly uniform motion of a large part of the visual field … When the entire field moves, subjects soon begin to feel that the relative motion is their own" (Young [64], p. 98). Whether found in a flight simulator, Cinerama theatre, IMAX theatre, or virtual reality simulation, vection causes a MS-like discomfort for a substantial minority of participants. This unpleasant experience is now universally referred to as SS. Further, these MS-like symptoms are now referred to as SS whether the simulator is a fixed-base model, and has no true motion, or a motion-base one with a (limited) range of movement. In other words, if the discomfort occurs in a simulator of any kind it will be called SS in the literature.

> Simulator sickness is a term used to describe the diverse signs or symptoms that have been experienced by flight crews during or after a training session in a flight simulator … Motion sickness is a general term for a constellation of symptoms and signs, generally adverse, due to exposure to abrupt, periodic, or unnatural accelerations. Simulator sickness is a special case of motion sickness that may be due to these accelerative forces or may be caused by visual motion cues without actual movement of the subject … (McCauley, [41], p. 1)

A subtle distinction has been made between true MS and SS. MS is caused by motion. SS is caused by an inability to simulate the motion environment accurately enough [23, 33, 48]. If a particular flight profile in an aircraft causes discomfort, this is MS. If the same profile is simulated veridically in a simulator, with the same physical forces present, and discomfort is caused, technically this is still MS. If a particular flight profile in the aircraft does not cause discomfort, but when simulated it does, this is SS. SS is discomfort produced in the simulator that does not occur when the same profile is executed in the physical motion environment. However, this is a logical distinction that apparently has no practical significance. As before, if the discomfort occurs in a simulator it will be called SS in the literature.

1.0 REVIEWS

This problem was duly noted and became the justification for increased research into the magnitude, correlates, causes, and treatment of SS. The results of this work have been reviewed extensively. Crowley and Gower [10] offered an introductory review for the experienced aviator. The excellent books edited by McCauley [41] and AGARD [1] reviewed key areas of this research. Reviews by Kennedy and colleagues described the earlier research with special emphasis on the large Navy database [23, 25, 29, 38]. With the emergence of virtual environment technologies and helmet-mounted displays in the 1990s, the salience of the problem of SS increased again – and this time not just for military training, but for consumer entertainment as

---

[1] This section is an excerpt from a previously published report entitled "Introduction to and Review of Simulator Sickness Research" by David M. Johnson. The report is U.S. Army Research Institute Research Report 1832, April, 2005.

well. Later reviews [5, 12, 26, 31, 48] expanded on the earlier reviews by including these newer technologies, where research was available, and addressing issues related to virtual reality. The detailed review by Wright [63] addressed the problem of SS in the training of Army helicopter pilots.

2.0 SELECTED HISTORY

Signs and symptoms of MS have been produced by visual stimulation alone in persons with an intact vestibular system. "This problem has been known to ophthalmologists and optometrists since the 1840s as the disorder termed asthenopia …" (Ebenholtz, [12], p. 302). Asthenopia remained a little-known optical disorder until 1956 when aviators began operating the first fixed-base (non-motion) helicopter simulator.

2.1 Miller and Goodson [44, 45]

Bell Aircraft Corporation was contracted by the Navy to develop a helicopter simulator for training visual flight skills and hovering. During preliminary demonstrations at Bell, prior to delivery to the Navy, it was found "… that a large number of observers (mostly helicopter pilots) experienced some degree of vertigo during these demonstrations" (Miller and Goodson, [44], p. 7). The observers commented that their discomfort stemmed from the lack of vestibular cues to motion available from the fixed-base device.

Upon installation at the Naval Air Station, Pensacola, two psychologists (Havron and Butler) conducted an initial training evaluation of the device. During this evaluation "… a questionnaire revealed that twenty-eight of thirty-six respondents experienced some degree of sickness" (Miller and Goodson, [44], p. 8). These participants included flight instructors, students, and other personnel experienced both in the simulator and the helicopter. "The more experienced instructors seemed to be the most susceptible to these unpleasant sensations" (Miller and Goodson, [44], p. 8). Sixty percent (60%) of the instructors reported SS symptoms, but only twelve percent (12%) of the students (Miller and Goodson, [45]). This SS usually occurred in the first ten minutes of a training session and frequently lasted for several hours afterward. The incidence and severity of this SS "… became such a serious problem that it was felt that unless it can be remedied in some way the utilization of such simulators as training devices would be limited considerably" (Miller and Goodson, p. 8).

As a part of their evaluation, Miller and Goodson [44] interviewed several of the instructors from the earlier Havron and Butler study. "One of these men had been so badly disoriented in the simulator that he was later forced to stop his car, get out, and walk around in order to regain his bearings enough to continue driving" (Miller and Goodson, p. 9). Miller and Goodson reported positive transfer of training from simulator to aircraft, albeit with a tiny sample size. Later Miller and Goodson conducted an experiment in an attempt to determine the effect of retinal disparity and convergence on SS in this device. They recruited 10 Navy enlisted men as participants. They were unable to find any effect of their independent variables upon SS and concluded that, due to large individual differences in the report of sickness, a "… great many more than ten subjects" (Miller and Goodson, p. 11) were needed to perform behavioral research on this phenomenon. They discussed problems with the device that caused several optical abnormalities. Specifically, Miller and Goodson [45] noted visual distortions and conflicts that could have caused the SS, including: blurring of the image, distorted size perspective, and distorted movement parallax. While Miller and Goodson concluded that the discomfort found could have been caused by some combination of conflicts within the visual modality alone, they also reported that an inter-sensory conflict between vision and proprioception existed. Finally, they listed a number of advantages to using a simulator for aircraft training, including: safety, weather independence, training for special missions, and large economic savings. However, the SS problem "… became so serious that it was one of the chief reasons for discontinuing the use of the simulator" (Miller and Goodson, p. 212).

The events described above represent the first published accounts of SS. Several of the issues identified at the dawn of SS research have remained issues throughout the history of the field. To wit:

1) A substantial percentage of the people who operate the simulator experience SS. This is not a trivial event for simulator-based training – especially for helicopter training.
2) The personnel with more experience in the aircraft appear to have an increased susceptibility to SS.
3) Conflicts both inter-sensory (visual/vestibular) and intra-sensory (visual/visual or vestibular/vestibular) are implicated as the cause of SS.
4) The aftereffects of SS can last for hours.
5) Unless remedied in some way, SS will limit simulator-based training.
6) The Miller-Goodson anecdote. "One of these men had been so badly disoriented in the simulator that he was later forced to stop his car, get out, and walk around in order to regain his bearings enough to continue driving." This anecdote has been repeated frequently throughout the literature as evidence that safety issues are at stake in simulator-based training.
7) Sample size matters. Individual differences in susceptibility to, and reporting of, SS are so large that behavioral research requires large sample sizes.
8) Research shows positive transfer of training from the simulator to the aircraft for many tasks.
9) There are many advantages to simulator-based training besides positive transfer of training, including: safety, independence from (non-flyable) weather, the opportunity to train special missions (mission rehearsal), and large savings in the resources required for flight training.

2.2 McGuinness, Bouwman, and Forbes [42]

The Air Combat Maneuvering Simulator (ACMS) was installed at the Naval Air Station, Virginia Beach, in November 1979; it was commissioned in February 1980; and by March of 1980 reports of SS had found their way to the Naval Training Equipment Center for investigation (McGuinness et al.). The ACMS was a wide FOV, fixed-base, fixed-wing aircraft simulator designed to resemble the cockpits of F-4 and F-14 fighters. Questionnaires were administered to 66 aviators during individual, confidential interviews. The aviators were either pilots or radar intercept officers with flight experience ranging from 250 to 4000 hours. Each had four one-hour training sessions in the ACMS over a period of approximately one week.

Twenty-seven percent (27%) of the participants experienced at least one symptom of SS. The rate for participants with greater than 1500 flight hours experience was 47%, while for those with 1500 or fewer hours it was 18%. The ages of participants were not reported, nor were the incidence rates presented by age. The most common symptom reported was dizziness, followed by vertigo, disorientation, and nausea. There were no reports of flashbacks. Of those who reported symptoms of SS, 61% stated that these symptoms persisted between 15 minutes and 6 hours. Of those who reported symptoms, all symptoms subsided completely after a night's rest. Thirty-three percent (33%) of the aviators reported that the reset function (freezing the visual display and returning to a new set of initial conditions) was the most probable cause of SS onset. There was some evidence of adaptation to the simulator over the course of several sessions. Finally, as a part of their literature review, the authors repeated the Miller-Goodson anecdote.

Several of the findings and explanations reported by McGuinness et al. [42] have been replicated or cited in many other articles since then. For example:

1) The authors explained the SS found in their study with reference to the sensory conflict theory. They argued that there was an inter-sensory conflict between the vection produced by the wide FOV visual display and the lack of any actual motion (vestibular stimulation) in the fixed-base simulator.

2) They explained the differential rate of SS as a function of flight experience, measured by flight hours, in the same fashion. The relative sensory conflict would have been greater for the more experienced aviators because these aviators had a larger neural store of prior flight experience. Therefore, a larger conflict between the current pattern of sensory inputs and the expected pattern would translate into more SS. However, unlike many later researchers, McGuinness et al. did not ignore age entirely. They cited a report by Olive stating that susceptibility to vertigo and disorientation increased with increasing age of Naval aviators. They also stated:

   > Physiological body changes resulting from physical aging may also be a contributing factor to this phenomenon, since those with more flight hours naturally tend to fall into older age groups. (McGuinness et al., [42], p. 25)

3) The SS symptoms reported by the participants, though similar to MS symptoms, were not identical. There were more vision and disorientation symptoms and fewer gastrointestinal symptoms. That is, there was less nausea and no emesis.

4) The symptoms had abated after one night's rest.

5) The freeze/reset function was implicated as causal in producing SS.

6) There was some evidence of adaptation over repeated simulator sessions.

2.3 McCauley [41]

McCauley described several potential operational problems that could result from SS. This discussion (McCauley's four points) was quickly adopted and repeated by later authors.

> 1) Compromised Training. Symptoms experienced in the simulator may compromise training through distraction and decreased motivation. Behaviors learned in the simulator to avoid symptoms (e.g., not looking out the window, reducing head movements, avoiding aggressive maneuvers) may be inappropriate for flight.
>
> 2) Decreased Simulator Use. Because of the unpleasant symptoms and aftereffects, simulator users may be reluctant to return for subsequent training sessions. They also may have reduced confidence in the training they receive from the simulator.
>
> 3) Ground Safety. Aftereffects, such as disequilibrium, could be potentially hazardous for users when exiting the simulator or driving home.
>
> 4) Flight Safety. No direct evidence exists for a relationship between simulator sickness aftereffects and accident probability. However, from the scientific literature on perceptual adaptation, one could predict that adaptation to a simulator's rearranged perceptual dynamics would be counterproductive in flight.
>
> (McCauley, [41], pp. 2-3)

These issues were discussed as potentially significant operational problems. For those who work in the field of simulator-based flight training, it is not a stretch to imagine that SS can affect safety and training. This possibility was noticed immediately (Miller and Goodson, [44, 45]). However, note that McCauley explicitly stated that there was “no direct evidence” suggesting simulators are causally implicated in aircraft accidents. McCauley’s four points appear frequently in published reports of SS.

2.4 Crowley [9]

In August 1984 the AH-1 Cobra Flight Weapons Simulator (FWS) became operational at Hanau U.S. Army Airfield in Germany. Soon thereafter reports of pilots becoming ill were made to Dr. Crowley, a flight surgeon at Hanau. Crowley’s study was performed during the spring of 1985. The FWS was a motion-base simulator, employing a terrain board database, and moderately narrow FOV visual displays (48 degrees horizontal gunner station, 96 degrees horizontal pilot station). Anonymous questionnaires were administered to 115 Army Cobra pilots who were training using the FWS simulator at Hanau. One hundred twelve (112) questionnaires were returned (97%).

Forty percent (40%) of the participants reported at least one symptom of SS. Nausea was the most frequent symptom, followed by sweating, and dizziness. Three pilots (3%) reported vomiting. Pilots who reported SS symptoms had significantly more total flight time than those who did not report symptoms. Pilots with greater than 1,000 hours of Cobra flight time were significantly more likely to report SS than pilots with fewer than 1,000 hours. Experience in the FWS was significantly and negatively correlated with reported SS. That is, more simulator time in the FWS was associated with fewer reports of SS symptoms. Crowley (1987) explained these results in terms of the sensory conflict theory. He quoted the Miller-Goodson anecdote. He also discussed McCauley’s four points and observed that any negative effects of SS upon training remained to be documented.

Because Crowley believed SS to be a potential hazard to aviation safety, a mandatory grounding policy was instituted at Hanau Army Airfield. The most significant portions of the Hanau policy were:

> Aviators flying the AH-1 Flight Weapons Simulator (FWS) are medically restricted from flying duties until the beginning of the next duty day, (normally 0630-0730) … Any aviator forced to stop a simulator period early due to motion sickness is grounded until seen by a flight surgeon and returned to flying duty. (Crowley, [9], p. 357)

3.0 SIGNS AND SYMPTOMS

SS is polysymptomatic [26, 29, 30]. Symptoms include nausea, dizziness, spinning sensations, visual flashbacks, motor dyskinesia, confusion, and drowsiness [41]. Observable signs of SS include pallor, cold sweating, and emesis [41]. The standard measurement instrument for SS, the Simulator Sickness Questionnaire (Kennedy, Lane, et al.), lists 16 symptoms: general discomfort, fatigue, headache, eyestrain, difficulty focusing, increased salivation, sweating, nausea, difficulty concentrating, fullness of head, blurred vision, dizzy (eyes open), dizzy (eyes closed), vertigo, stomach awareness, and burping. Reports of visual flashbacks and visual hallucinations have been documented [41, 63, 64] although they are reported to be exceedingly rare.

The reader will note that the signs and symptoms of SS overlap with those described above for MS. There are several differences, however. The most consistently reported difference is that while major symptoms of MS involve gastrointestinal distress (e.g., burping, stomach awareness, nausea, emesis), for SS there are fewer gastrointestinal symptoms and more visual ones (e.g., eyestrain, difficulty focusing, blurred vision, headache) [23, 26, 30, 31, 38, 61]. Vomiting is a common sign of MS. For example, 75 percent of those suffering from seasickness vomit [26]. By comparison, vomiting is rare in SS – usually occurring in less than one percent (1%) of the cases [26, 30]. Finally, in cases of vection-induced SS, such as a fixed-base flight simulator, closing one’s eyes will end the perceived motion and dramatically reduce the symptoms [30]. Closing one’s eyes, however, will have no such effect on MS, as noted above.

Helicopter simulators have been widely reported to produce more SS than fixed-wing simulators [2, 23, 31, 32, 63, 64]. This is probably because helicopters are usually flown closer to the ground. Discomfort level varies inversely with height above terrain [26, 33, 63]. There is a greater perception of visual flow, caused by greater visual detail, at lower height above terrain.

Several reports of original research include a listing of the most common symptoms found in helicopter simulators. Gower and Fowlkes [14] reported a study of the Cobra AH-1 FWS. This device incorporated a six-degree of freedom (6-DOF) motion base. (These six dimensions of motion are pitch, roll, yaw, vertical [heave], lateral [sway], and longitudinal [surge]). The most commonly reported symptoms from Gower and Fowlkes were eyestrain (37% of the participants) and fatigue (27%).

Gower, Lilienthal, Kennedy, Fowlkes, and Baltzley [17] reported on another simulator of an attack helicopter. This was the Combat Mission Simulator for the Apache AH-64A. The CMS is an interactive, full-mission, 6-DOF simulator. The most commonly reported symptoms were fatigue (43% of participants), sweating (30%), and eyestrain (29%). Braithwaite and Braithwaite [6] reported on a simulator for the British attack helicopter the Lynx. This device included a 6-DOF motion system with a 130 degree (horizontal) by 30 degree (vertical) FOV color projection visual system. The most commonly reported symptoms were disorientation (24% of participants) and difficulty focusing (24%).

Gower and Fowlkes [15] studied the SS potential of a simulator for the UH-60 Blackhawk utility helicopter. This device incorporated a 6-DOF motion base plus forward, left, and right out-the-window views from a collimated visual display. The most common symptoms were fatigue (35% of participants) and eyestrain (34%). Silverman and Slaughter [57] reported on an operational flight trainer for the MH-60G PAVE Hawk helicopter. This was a fixed-base device. It provided a 150 degree (h) by 40 degree (v) out-the-window visual display plus two chin window displays. The most commonly reported symptoms were stomach awareness, dizziness, nausea, fatigue, and sweating in descending order of frequency.

Gower, Fowlkes, and Baltzley [16] reported on the SS symptoms produced by the full-mission simulator model 2B31 for the CH-47 Chinook cargo helicopter. This was a 6-DOF motion device with a 48 degree (h) by 36 degree (v) forward visual display plus a 22 degree (h) by 30 degree (v) chin window display. The most commonly reported symptoms of SS were fatigue (34% of participants), eyestrain (29%), headache (17%), difficulty focusing (13%), sweating (11%), nausea (9%), and stomach awareness (9%).

4.0 MEASUREMENT

Several reviews discussed the difficulties with and tools for measuring SS [7, 20, 26, 33]. Because SS is polysymptomatic one cannot measure just one dependent variable (Kennedy and Fowlkes). Another measurement difficulty is that there are large individual differences in susceptibility to SS. It is common in this research to find that fully 50 percent of simulator operators experience no symptoms at all (Kennedy and Fowlkes [26]). When effects of SS exist, they are often small, weak effects that disappear quickly upon exiting the simulator. Further, because most participants eventually adapt to the motion environment of a particular simulator, researchers cannot reuse the same participants (such as in a within-subjects research design). Thus, researchers are forced to employ between-subjects research designs (Kennedy and Fowlkes [26]). When one combines these factors of large individual differences, weak effects, adaptation, and between-subjects designs it invariably leads to the conclusion that research into SS requires large sample sizes. To get samples of this large size, researchers are forced to survey pilots training in simulators at military training centers (Kennedy and Fowlkes [26]). However, these military centers exist to train pilots efficiently and effectively, not to perform research. This means that the level of experimental control exercised by a researcher is usually low. So research studies investigating SS are either vast surveys of nearly all pilots operating a particular simulator at a particular facility at a particular time, or small-scale experiments with rather more experimental control, but much smaller sample sizes.

There are a number of possible ways to measure SS [7, 20]. One could employ direct observation of participants during a simulator session and note signs such as facial pallor and sweating. This is seldom done for research measurement (cf., Uliano et al., [61]), but often used by instructors at the simulator site to monitor their students. Another option would be self-report measures, such as the Simulator Sickness Questionnaire, that ask the participant to note the type and severity of symptoms currently being experienced. This method is universally performed in some fashion. A third option would be to instrument the participants and measure physiological conditions such as respiration rate and stomach activity. This method has been used upon occasion. Finally, one can employ tests of postural equilibrium to measure simulator-induced disorientation or ataxia. These tests have been widely employed, but with equivocal results.

4.1 Simulator Sickness Questionnaire (SSQ)

The SSQ is currently the gold standard for measuring SS. This instrument was developed and validated by Kennedy, Lane, et al. [30]. The SSQ was developed based upon 1,119 pairs of pre-exposure/post-exposure scores from data that were collected and reported earlier (Baltzley et al., [2]; Kennedy et al., [32]). These data were collected from 10 Navy flight simulators representing both fixed-wing and rotary-wing aircraft. The simulators selected were both 6-DOF motion and fixed-base models, and also represented a variety of visual display technologies. The SSQ was developed and validated with data from pilots who reported to simulator training healthy and fit.

The SSQ is a self-report symptom checklist. It includes 16 symptoms that are associated with SS. Participants indicate the level of severity of the 16 symptoms that they are experiencing currently. For each of the 16 symptoms there are four levels of severity (none, slight, moderate, severe). The SSQ provides a Total Severity score as well as scores for three subscales (Nausea, Oculomotor, and Disorientation). The Total Severity score is a composite created from the three subscales. It is the best single measure because it provides an index of the overall symptoms. The three subscales provide diagnostic information about particular symptom categories. The Nausea subscale is made up of symptoms such as increased salivation, sweating, nausea, stomach awareness, and burping. The Oculomotor subscale includes symptoms such as fatigue, headache, eyestrain, and difficulty focusing. The Disorientation subscale is composed of symptoms such as vertigo, dizzy (eyes open), dizzy (eyes closed), and blurred vision. The three subscales are not orthogonal to one another. There is a general factor common to all of them. Nonetheless, the subscales provide differential information as to symptomatology and are useful for determining the particular pattern of discomfort produced by a given simulator. All scores have as their lowest level a natural zero (no symptoms) and increase with increasing symptoms reported.

An important advantage of the SSQ is that a wide variety of symptoms can be measured quickly and easily with the administration of this one questionnaire. Another important advantage is that it allows quantitative comparisons across simulators, populations, and within the same simulator over time (as a diagnostic to determine if recalibration is needed, for example).

However, Kennedy, Lane, et al., [30] stated restrictions in the use of the SSQ also. First, the SSQ is not to be used with participants who are in other than their usual state of health and fitness. The instrument was developed and validated based on data from healthy, fit pilots. Any scores obtained from participants who arrived for simulator training ill would be uninterpretable. Second, the authors recommended that the SSQ be administered immediately after a simulator session, but not before one. They did not recommend using pre-post difference scores. This is because the high correlation usually found between pre and post can render the difference scores unreliable. Nonetheless, researchers are so comfortable with the SSQ that they sometimes report pre-post difference scores anyway (e.g., Regan and Ramsey, [53]).

4.2 Instrumented Physiological Measures

> Changes in bodily cardiovascular, gastrointestinal, respiratory, biochemical, and temperature regulation functions often arise with simulator sickness. Several physiological measures have been electronically or electro-optically instrumented and transduced directly from subjects in simulator experiments. (Casali and Frank, [7], pp. 9-10).

Heart rate, or pulse rate, has been reported to change from baseline levels as a function of simulator exposure [7]. Unfortunately these reported changes are not sensitive, reliable, or always in the same direction. Respiration rate has proven to be a sensitive index of SS (Casali and Frank). However, the direction of the change is not consistent across individuals. As with MS [52] some individuals increase respiration rate upon simulator exposure, while others decrease rate. Casali and Frank recommend using an absolute difference score. Sweating is a common symptom of SS and this can be measured as an increase in skin conductance or a decrease in skin resistance (Casali and Frank). Facial pallor is also a common symptom of SS. Paleness of the skin can be measured using photo-optical sensors and has been shown to vary as a function of conditions that cause SS (Casali and Frank). Gastric activity can be measured with an electrogastrogram. Gastric activity in the form of tachygastria, a dramatic increase in stomach motility, has been shown to occur along with other symptoms of SS during exposure to vection (Casali and Frank; Hettinger et al., [7]).

4.3 Tests of Postural Equilibrium

Reviews of this methodology can be found in Casali and Frank [7], Kennedy et al., [24], and Kolasinski [33]. Postural equilibrium tests (PETs) exist to provide a behavioral measure of ataxia. Ataxia is a potentially dangerous symptom of SS. It is usually defined generically as:

> An inability to coordinate voluntary muscular movements that is symptomatic of any of several disorders of the nervous system. (Merriam-Webster, [43], p. 137).
>
> Marked incoordination in voluntary muscular movements. (English and English, [13], p. 48).

In the domain of SS research, ataxia is defined as postural instability, postural unsteadiness, or postural disequilibrium (e.g., Kennedy et al., [24]; Kolasinski and Gilson, [35]). It is thought that any disruption of balance and coordination that results from exposure to a simulator may be a safety concern for pilots who need to walk, climb stairs, drive, or fly after a simulator training session. The PETs are used to provide a direct index of postural instability.

> Loss of balance and ataxia are common problems noted by trainees and subjects after exiting a dynamic simulator. The simulator presents an altered sensory environment which usually entails considerable vection, and some adaptation to this environment occurs in the operator’s visual and vestibular sensory systems. Upon return to the “normal” environment, balance and equilibrium may be disrupted until the person progresses through re-adaptation. Such effects may be measured using pre-post simulator postural equilibrium tests. (Casali and Frank, [7], p. 14).

There are several PETs that are described in the literature. They all involve some permutation of the following procedures: standing heel to toe with eyes closed and arms folded across the chest or back; or standing on one leg (preferred leg or non-preferred leg) with eyes closed or open and arms folded across the chest; or walking a straight line (on floor or rail) heel to toe with eyes closed or open and arms folded across the chest. The names and acronyms, where available, for several PETs are listed: Sharpened Romberg (SR), Stand on One Leg Eyes Closed (SOLEC), Stand On Preferred Leg Eyes Closed (SOPLEC, SOPL), Stand On Non-preferred Leg Eyes Closed (SONLEC, SONL), walk toe to heel, Walk On Floor Eyes Closed (WOFEC), Walk On Line Eyes Closed (WOLEC), and Walk On Rail Eyes Open (WOREO).

An example of a method for using PETs in research is described below:

> Standing on Preferred Leg (SOPL): This test of standing steadiness required pilots to first determine which leg they preferred to stand on. Pilots were asked to stand, fold their arms against their chest, close their eyes, lift their non-preferred leg and lay it about two-thirds of the way up the standing leg's calf. They attempted to remain in that position for 30 s. If they moved their pivot foot, moved their raised foot away from their standing leg, grossly lost their erect body position, the trial ended and the time up to that point (in seconds) was recorded as the score for that trial.
>
> Standing on Non-Preferred Leg (SONL): The procedure for this test was identical to that of the SOPL test except that pilots stood on their non-preferred leg. (Kennedy et al., [24], p. 15).

The research literature shows mixed results when using PETs to demonstrate an effect of simulator exposure upon postural stability. Some studies have found no statistically significant effect of simulator exposure upon performance of PETs [15, 16, 18, 36, 61]. Other studies have found a statistically significant effect for some or all PETs used [11, 14, 17, 24, 37, 62].

There are several differences among the reports cited above. Nonetheless possible explanations for these equivocal results present themselves. As mentioned above with regard to SS in general, if an effect is highly subject to individual variability then large sample sizes are required. The mean sample size for the five studies listed above that did not report a significant difference was 61. For the six studies that reported positive results the mean sample size was 120. One cause of variability in performance can be differential rates of learning. Hamilton et al. [18] demonstrated significant learning effects in the performance of four PETs (SR, SOLEC, WOREO, WOLEC). Further, performance on these four PETs continued to improve over the 10 practice sessions they measured. Therefore, when using PETs one must be aware that any improvement in performance occasioned by learning will tend to mask any decrement in performance caused by simulator exposure – if such a decrement exists.

Finally, Kennedy et al. [24] found a statistically significant correlation between the disorientation subscale of the SSQ and performance measures taken from two PETs (SOPL, SONL). The higher the disorientation scores on the SSQ, the poorer the performance on the two PETs. In other words, the subjective self-reports of the pilot participants accurately reflected the behavioral measures taken from them after exiting the simulators. Given the potential measurement problems associated with PETs, the time and effort required in their administration, and the fact that similar results can be acquired more easily and quickly with the SSQ, the use of tests of postural equilibrium should probably be limited to research questions where their specific contribution is necessary.

5.0 INCIDENCE

The incidence of SS varies widely across simulators and conditions. A common method of presenting incidence is to list the percentage of participants who reported at least one symptom. In the review by McCauley [41] incidence was reported to range from 10 to 88 percent. In their review Kennedy and Frank [29] reported that incidence ranged from 27 to 88 percent. In later reviews Kennedy and colleagues [25, 26] reported that the incidence of SS ranged from 12 to 60 percent in Navy flight simulators. Pausch et al., [48] reported in their review that it could range from 0 to 90 percent in flight simulators. Wright [63] limited his review to helicopter flight simulators. He reported that the incidence ranged from a low of 13 percent, when a strict criterion was employed to define SS, to a high of 70 percent, when a lax criterion was used.

It is widely reported that simulators of rotary-wing (RW) aircraft cause participants more SS than simulators of fixed-wing (FW) aircraft. Assuming a constant criterion of at least one reported symptom, there are several studies that report incidence by simulated aircraft type. Kennedy and colleagues [23, 32] collected data from 1,186 simulator exposures. Their sample included data from 10 flight simulators. These simulators represented both FW and RW aircraft, and included both motion-base and fixed-base models. The incidence rates for FW simulators ranged from 10 to 47 percent. The rates for RW simulators ranged from 26 to 69 percent. Baltzley et al., [2] collected data from 742 exposures using a self-report questionnaire. Their sample included data from operators of 11 flight simulators (7 FW, 4 RW). All participants had experience training in flight simulators. The incidence rates reported by pilots training in FW simulators ranged from 6 to 62 percent. The rates reported by pilots training in helicopter simulators ranged from 48 to 57 percent. These results have the advantages of large sample sizes, multiple flight simulators, and a constant method of research and analysis performed by the same investigators.

Magee, Kantor, and Sweeney [40] collected data from a sample of 42 C-130 pilots and flight engineers. The C-130 Hercules is a multi-engine, propeller-driven, FW, cargo aircraft. The C-130 simulator included a 6-DOF motion base and a 120 degree (h) by 40 degree (v) FOV visual display. Participants performed a four-hour simulator session with a short break at the mid-point. Ninety-five percent (95%) of the participants reported at least one symptom of SS upon exiting the simulator.

Crowley reported an incidence rate of 40 percent for the RW Cobra FWS. Braithewaite and Braithewaite [6] reported an incidence rate of 60 percent for 183 Lynx helicopter crewmembers that returned self-report questionnaires. Gower et al., (1987) collected data from 127 participants training in the AH-64 CMS. This simulator represents the AH-64A Apache helicopter. An incidence rate of 44 percent was reported. Gower and Fowlkes [14] collected data from 74 Army aviators training in the Cobra FWS. Thirty-seven percent (37%) of the participants reported at least one symptom of SS. All four of the studies described in this paragraph reported results obtained from participants operating 6-DOF motion-base devices that simulated attack helicopters.

Lerman et al. [37] collected data from 59 armor Soldiers performing tank driver training in a 3-DOF (pitch, roll, yaw) tank simulator. Sixty-eight percent (68%) of this sample reported at least one symptom of SS. Using the SSQ, Lampton et al. [36] measured SS in an M-1 tank driver simulator mounted on a 6-DOF motion platform. They also measured discomfort in the actual M-1 tank. The authors reported significantly greater symptom scores in the simulator than in the tank. Upon interview, thirty-six percent (36%) of their sample reported experiencing discomfort in the simulator. The authors also reviewed the training records of six armor companies that had experienced the device previously. They found that 25 percent of these training records documented SS among the prior trainees. It is plausible that these incidence rates reported by Lampton and colleagues are conservative estimates. Instructors are not likely to mention SS in a written training document unless it is a significant phenomenon.

SS also exists in virtual reality (VR) simulators. For a review of SS from this perspective see Kolasinski [33]. Regan and Ramsey [53] reported a 75 percent incidence rate for subjects in the placebo control group of a VR drug experiment. This level of discomfort was produced by a 20-minute immersion in the VR simulator. Kolasinski and Gilson [35] immersed 40 research participants in a commercially available VR simulator for 20 minutes. Eighty-five percent (85%) of the participants reported at least one symptom of SS. It was because of high sickness rates such as these, produced by relatively short simulator sessions, that the practical future of VR technology became a subject of discussion (e.g., Biocca, [5]; Kolasinski, [34]; Pausch et al., [48]).

It is clear from the literature reviewed above that the incidence of SS varies within a large range. Depending upon the simulator, the conditions of operation, and the criterion definition applied, the rate of SS can vary from low to extremely high.

6.0 RESIDUAL AFTEREFFECTS

The potential for dangerous aftereffects of simulator exposure – including ataxia, loss of balance, flashbacks – has been noted right from the beginning [44, 45]. In fact, the careful reader will meet the Miller-Goodson anecdote frequently in the literature – either quoted directly (e.g., Crowley, [9]; McCauley, [41]; McGuinness et al., [42]; Pausch et al., [48]; Wright, [63]) or, more often, referred to obliquely. McCauley's four points – two of which concern safety – are ubiquitous. Virtually every report refers in some way to these points, usually in the introductory section. So researchers have done their part to alert the community of the potential for dangerous aftereffects of simulator-based flight training.

However, it is only prudent to assure the reader that this potential danger has not manifested itself objectively. Many of the same authors reported that there were no documented cases of flight incidents or automobile accidents linked to prior simulator-based training [9, 29, 41, 63]. The present author has performed a follow-up study on several hundred simulator-trained Apache pilots [21]. Not one aviator has reported an automobile or motorcycle accident within 12 hours of exiting the simulator.

Baltzley et al. [2] reported data from a large study involving 742 simulator exposures across 11 Navy and Army simulators. Overall, 45 percent of the participants reported experiencing symptoms of SS upon exiting the simulator. Of these pilots who reported symptoms, 75 percent said that their symptoms disappeared within 1 hour. Six percent (6%) reported that their symptoms dissipated in 1 to 2 hours, 6 percent in 2 to 4 hours, 5 percent in 4 to 6 hours, and 8 percent reported that their symptoms lasted longer than 6 hours. The most common category of aftereffect was nausea (51%), followed by disorientation (28%), and oculomotor (21%).

Braithwaite and Braithwaite [6] reported that 17 percent of their sample experienced aftereffects. The most frequently stated aftereffects were nausea, which dissipated in 2 hours, and headache, which sometimes lasted as long as 6 hours. Crowley [9] reported that 11 percent of his sample experienced delayed effects of simulator training. The most commonly reported delayed symptom was a perception of illusory movement. Gower et al. [17] reported aftereffects following training in the Apache CMS. Over a series of 10 training sessions, preflight minus postflight performance on 3 PETs decreased until session number 4 and then remained stable for the remainder of the simulator periods. This was interpreted as behavioral evidence of increasing simulator-induced disequilibrium over training trials.

McGuinness et al. [42] reported that 18 members of their sample of 66 aviators (27%) experienced at least one symptom of SS. Of these 18, 11 (61%) stated that their symptoms persisted anywhere from 15 minutes to 6 hours. Silverman and Slaughter [57] reported results from participants operating a wide FOV, fixed-base MH-60G operational flight trainer for the PAVE Hawk helicopter. Data were collected in conjunction with an operational test and evaluation of the simulator. Sortie lengths were at least 3 hours and included a full range of flight tasks. A total of 13 experienced aviators participated and filled-out self-report questionnaires.

Eight (8) of these 13 participants (62%) reported at least one symptom aftereffect. The most commonly reported aftereffects were fatigue, stomach awareness, and vertigo, in that order. Most of these aftereffects came and went within 2 hours of exiting the simulator, although some participants reported symptoms lasting up to "… several hours after the simulator training session" (Silverman and Slaughter, p. 11).

There are some crude conclusions that emerge about the aftereffects of simulator exposure. First, approximately 10 percent of the sample will experience pronounced aftereffects [23, 26]. Second, there is a significant positive correlation between the number and severity of symptoms reported immediately upon leaving the simulator, and the duration and severity of aftereffects [8, 57]. That is, those who experience the most SS during the simulator session usually experience the most aftereffects later. Third, the aftereffects of simulator exposure usually wear off in an hour or two. The persistence of symptoms longer than 6 hours has been documented repeatedly, but fortunately remains statistically infrequent.

It is understood in the training community that a potential exists for residual aftereffects to be a risk to flight safety. For this reason, guidelines recommending a mandatory grounding policy after training in a flight simulator have appeared both in the research literature and the training environment [8, 9, 10, 23, 25, 31, 38, 47]. The minimum recommended period from simulator to aircraft has ranged from 6 to 12 hours and usually includes the admonition to wait until the next day. In cases of severe discomfort, temporary curtailment of other duties for up to 24 hours has been recommended [23]. Currently in the U.S. Army, the policy on how much time must elapse from the end of a simulator training session to flight duty is a matter of unit standard operating procedure and is set by the unit commander [55]. At USAAVNC there is currently no regulation that restricts simulator-to-aircraft time – except for the Longbow Crew Trainer where the required delay is 1 hour [55].

> Allowing a night's sleep before recommencing flying duties should reduce residual risks to negligible proportions. (Chappelow, [8], p. 10).
>
> During initial simulator training sessions or after a long period of not using the simulator, avoid scheduling simulator and aircraft flights on the same day. (NTSC, [47] p. 8).

7.0 ADAPTATION

The concept of adaptation in the literature of SS is identical to that in the literature of MS. Several reviewers have discussed adaptation to a novel simulated motion environment [5, 10, 26, 29, 33, 63]. The theoretical approach used to explain the fact that most participants adapt to the simulator after approximately six sessions (Biocca; Wright) is the sensory conflict theory.

Crowley [9] found that there was a statistically significant inverse relationship between the prior number of hours spent training in the Cobra simulator and the amount of SS reported. The more prior exposure to the simulator, the less SS experienced currently. This was interpreted as evidence of adaptation. Gower and Fowlkes [14] reported the same inverse relationship with a different sample of Cobra pilots and another FWS. Gower et al. [16] reported a significant negative correlation between prior history of hours spent training in the CH-47 flight simulator and SS for a sample of experienced CH-47 pilots. Gower et al. [16] investigated the effects of exposure to the AH-64A CMS on discomfort levels for 127 Apache aviators. Over the course of 10 training sessions, they found that self-reported SS symptoms decreased with increasing sessions in the CMS. They also reported an inverse relationship between the amount of simulator exposure during the prior 3 months and SS. Finally, they noted a significant negative correlation between the amount of recent CMS exposure and disequilibrium as measured by a PET. These results were interpreted as evidence of adaptation to the CMS.

Silverman and Slaughter [57] reported evidence of adaptation to a MH-60G operational flight trainer for the PAVE Hawk helicopter. A sample of 13 experienced pilots executed a full range of flight tasks over several sessions in the simulator. The number of SS symptoms reported on later days were significantly fewer than the number reported on the first day of testing. Uliano et al. [61] required 25 experienced pilots to operate the Vertical Take-off and Landing (VTOL) simulator which represents the SH-60B Seahawk helicopter. Each pilot flew the same flight paths, and performed the same tasks under the same experimental conditions, in counter-balanced order, over 3 days. SS was reported to be significantly worse on day 1 than day 2, and significantly worse on day 2 than day 3. The authors interpreted these results as evidence of adaptation to the simulator.

Besides reviewing the SS literature, Wright [63] reported on his interviews with Army helicopter flight instructors. These instructors trained helicopter pilots daily. Yet, when introduced to a new simulator, they experienced SS symptoms. After a few days the symptoms disappeared or at least subsided to a minor and tolerable level. These instructors also reported that after several months away from the simulator, they had to readapt as if for the first time. Then readapt they did, again, in a few sessions. Wright interpreted these statements as evidence of adaptation to a novel (simulated) motion environment.

All of the studies cited above involved aviators adapting to a helicopter flight simulator of some kind. Lampton et al. [36] reported evidence of adaptation to an M-1 tank driver trainer. They collected data from 115 trainees, all of whom had no prior experience driving a tank. Over the course of several training sessions the amount of SS the trainees experienced decreased. The symptom scores, as measured using the SSQ, were significantly higher after the first training session than after the remainder of the sessions. These results were interpreted as adaptation to the simulator.

Reports and manuals that provide guidelines for the detection and treatment of SS acknowledge adaptation as the best current solution to the problem of simulator-induced discomfort (e.g., Kennedy et al., [25] ; Lilienthal et al., [38]; NTSC, [47]). As with MS, almost all participants eventually adapt to a simulated motion environment. Guidelines often describe procedures to employ during simulator-based flight training to encourage a rapid and reasonably comfortable adaptation period. For example:

> Adaptation of the individual is one of the strongest and most potent fixes for simulator sickness … Do not schedule simulator hops for greater than two hours for any reason. (Kennedy et al., [25], pp. 12, 17).
>
> Persons new to the simulator, and particularly persons with extensive flight time, are at most risk … Decrease the field of view during nauseogenic hops (e.g., initial hops) … Go on instruments. (Lilienthal et al., [38], pp. 277, 279).
>
> Brief simulator flights (short hops with gentle maneuvers) separated by one-day intervals will facilitate adaptation to simulator motion and help prevent sickness, especially during the early stages of simulator training for novices and for experienced pilots with little simulator training … Do not slew while the visual scene is turned on … If all else fails, turn off the motion base or the visual scene and conduct instrument training. (NTSC, [47], pp. 6-7).

8.0 SUSCEPTIBILITY

SS is not only polysymptomatic; it is polygenic [26, 29]. Kennedy and Fowlkes [26] listed 13 factors that are implicated in causing SS. These factors were subdivided into three categories: individual variables, simulator variables, and task variables. In an exhaustive review, Kolasinski [33] described 40 factors that are associated with SS – also categorized as individual, simulator, and task variables. Pausch et al. [48] reviewed several factors that evoke SS, with special emphasis given to simulator design issues.

8.1 Gender

As with MS (e.g., Reason and Brand, [52]), reviews of SS reported that females are more susceptible than males (e.g., Biocca, [5]; Kennedy and Frank, [29]; Kolasinski, [33]; Pausch et al., [48]). The precise reason for this is unknown. Reviewers have cited at least three possible explanations: hormonal differences, FOV differences, and biased self-report data. The hormonal hypothesis is the same as that advanced in the MS literature – females are more susceptible to SS during a portion of their menstrual cycle. This hypothesis is not without its doubters (e.g., Biocca; Pausch et al.). More likely, some think, is the fact that females have a larger effective FOV, and larger FOV is associated with greater SS (e.g., Biocca; Kennedy and Frank; Pausch et al.). Finally, those data upon which gender differences are based are self-reports. Males, it is suggested, may be more likely to under-report the severity of their discomfort (e.g., Biocca; Kolasinski).

However explained, reports of gender differences in SS continue to emerge. Hein [19] reported the results of 22 separate studies, involving 469 participants, over the course of 6 years. All studies took place in a fixed-base, automobile-driving simulator. Hein stated that gender differences in susceptibility to SS were among the most consistent results. “At all ages, female drivers are more susceptible than male drivers” (Hein, p. 610).

8.2 Age

> Walt Disney World’s “Mission: Space” thrill ride left some older riders gulping, “Houston, we have a problem.” In the past eight months, six people over 55 have been taken to the hospital for chest pain and nausea after going on the $100 million ride … It is the most hospital visits for a single ride since Florida’s major theme parks agreed in 2001 to report any serious injuries to the state … Last December, Disney began placing barf bags in the ride … (Schneider, [56], p. B2).

Reviewers have reported that susceptibility to SS varies with age in the same way that MS varies with age (e.g., Biocca, [5]; Kennedy and Frank, [29]; Kolasinski, [33]; Pausch et al., [48]; Young, [64]). That is, below age 2 infants are generally immune. Susceptibility is at its highest level between ages 2 and about 12. There is a pronounced decline between ages 12 and 21. This decline continues, though more slowly, through adulthood until about age 50, after which SS is very rare. These claims are based on the self-report data reviewed by Reason and Brand [52] for MS.

Perhaps the reason reviewers are forced to report conclusions based on decades-old self-report surveys of MS symptoms, is because so little research has been performed examining the effect of age on susceptibility to SS. Very few researchers have attempted to investigate the relationship between age and SS more directly. Braithwaite and Braithwaite [6] administered questionnaires to 230 pilots attending training in a simulator for the Lynx attack helicopter. All were males. Age ranged from 23 to 42 years with a mean age of 32. There was no relationship found between age and reported SS.

Warner et al. [62] assessed SS in two wide-FOV F-16 flight simulators. Twenty-four (24) male pilots participated in total. Sixteen (16) were active-duty military pilots of mean age 28.6 years (the “younger group”). Eight (8) were older active-duty military pilots and former military pilots of mean age 52.1 years (the “older group”). The task was a challenging 50-minute flight through a long, narrow, twisting canyon in each of the two simulators, in counter-balanced order, two weeks apart. One pilot from the younger group (1/16 = 6.25%) terminated a session prematurely due to severe SS. Three pilots from the older group

(3/8 = 37.5%) terminated a session prematurely due to severe SS. The discomfort ratings (early version SSQ) collected from pilots who terminated prematurely were significantly higher than those from pilots who completed the flight. Among those pilots who completed the flight, there was no significant difference in discomfort ratings between the younger and older groups. Among those pilots who completed the flight, there was also no significant difference in postural equilibrium (SOLEC, WOFEC) between the groups.

Hein [19] reported the results of 22 separate studies, involving 469 participants of both genders and a wide range of ages, over the course of 6 years. All studies took place in a fixed-base, automobile-driving simulator. Hein stated that age differences in susceptibility to SS were among the most consistent results. “Younger, male drivers adapt easily. Older drivers and women are severely susceptible to simulator sickness” (Hein, p. 611).

8.3 Age and Experience

Among those (like the present author) who have been involved in the simulator-based training of large numbers of aviators, it is common knowledge that older participants are more susceptible to SS. Further, the small amount of evidence that does exist tends to support these anecdotal observations. Yet researchers investigating SS rarely even aggregate their data by age. Given the importance of age in both behavioral science and medical science research, this appears to be a glaring omission. Then, to confuse matters further, reviewers of the SS literature continue to repeat the conclusions published by Reason and Brand [52] that sickness decreases with age and eventually almost disappears. Why is this so?

This is because researchers are convinced that the demographic variable that influences aviator SS is experience as measured in flight hours, not chronological age. Data are frequently aggregated by the flight hours of the participants. Researchers reviewing the literature discuss the impact of aircraft flight experience on SS. This view is also entirely consistent with the sensory conflict theory, where experience in a particular motion environment is central to the explanation.

However, among aviators age (in years) and experience (in flight hours) are strongly linked. Magee et al. [40] reported a statistically significant correlation between age and flight hours ($r = 0.67$). The present author [21] has also found a significant correlation ($r = 0.75$) between these variables. This is because “As is common in most professions, piloting experience tends to accumulate with age” (Tsang, [60], p. 525). Thus, disentangling age from experience is a knotty problem when examining SS among aviators (see [60]).

It would not, in principle, be such a difficult problem to assess the effect of age upon SS if non-aviators were used as research participants. The present author predicts that among adult non-aviators, SS will increase with age rather than decrease. The chief methodological problems to be solved in order to perform this research would be practical ones. First, gaining access to a sufficiently large sample of non-aviators of a wide range of ages. Second, gaining access to a flight simulator for a period of time sufficient to collect the requisite large amount of data.

8.4 Experience

It is universally understood within this research community that the more experienced aviators are more susceptible to SS than novices. For example, this understanding has been acknowledged in at least 12 reviews covering the period from 1984 to 2003 [4, 10, 25, 26, 29, 33, 38, 41, 46, 48, 63, 64]. In addition, some empirical evidence of this relationship has already been described earlier in the reports by Crowley [9], McGuinness et al. [42], and Miller and Goodson [44, 45].

Braithwaite and Braithwaite [6] found a statistically significant positive correlation between experience as measured in flight hours and SS among pilots training in a Singer-Link simulator for the Lynx attack helicopter. That is, pilots with a greater number of flight hours reported greater SS. Gower and Fowlkes [15] assessed SS (early version SSQ) among 87 Army aviators training in a UH-60 helicopter simulator. They found a significant positive correlation between flight hours and SSQ scores. Gower et al. [16] collected data from 57 aviators with flight experience ranging from 450 to 7,000 flight hours. The pilots were taking currency and refresher training in a 2B31 simulator for the CH-47 cargo helicopter. The authors found no correlation between flight hours and SSQ scores. Gower et al. [16] assessed SS among 127 Apache aviators with flight experience ranging from 150 to 8,400 flight hours. All pilots were training in the AH-64 CMS built by Singer-Link. Again, the authors found no significant correlation between flight hours and reported SS symptoms.

Magee et al. [40] assessed SS among a group of 42 male C-130 pilots and flight engineers operating a CAE C-130 simulator. Twenty-six (26) participants (the “experienced group”) had flight hours ranging from 845 to 10,000 (median 3,166). Sixteen (16) participants (the “novice group”) had flight hours ranging from 50 to 4,340 (median 1,465). There was no significant difference between the two groups in measured SS, either immediately after the simulator session or later. Also, a partial correlation of flight hours against measured SS, with age held constant, showed a small (0.03) and statistically insignificant result.

Silverman and Slaughter [57] collected data from 13 aviators as part of an operational test of a MH-60G PAVE Hawk simulator. The participants’ total flight experience ranged from 350 to 15,327 hours. The authors reported that there was no statistically significant correlation between reported SS and either total flight hours or flight hours for the specific MH-60G helicopter. Uliano et al. [61] assessed SS among 25 male helicopter pilots. Their flight experience ranged from 360 to 2,860 hours (mean 1,071). All participants operated the VTOL simulator, which represented the SH-60B Seahawk helicopter. Aviators with fewer than 900 flight hours experience reported significantly less SS on all measures than those with 900 or more flight hours.

Lerman et al. [37] collected data from 59 male armor Soldiers operating a tank driver trainer. The authors found no significant correlation between amount of prior tank driving experience and SS symptoms.

Sensory conflict theory states that SS is caused when there is a difference between the current pattern of sensory information and what is expected on the basis of past experience. Thus, this theory predicts that the more flight experience an aviator has acquired, the greater will be the disparity between his or her neural store and the pattern presented by the flight simulator – since a simulator cannot perfectly simulate flight – and the more SS will be reported. This is the explanation given when statistically significant differences are found between highly experienced aviators and novices or students.

8.5 Prior History of Motion Sickness

Generally speaking, in the behavioral sciences past behavior is the best predictor of future behavior. It follows that people who have a history of prior episodes of MS or SS will be more likely to experience SS in future simulator-based training. Two reviewers reported that there is empirical evidence in support of this generalization [25, 63]. Kennedy, Fowlkes, et al. [27] discussed some of the methodological issues involved in using the Motion History Questionnaire (MHQ) to predict sickness scores in a simulator.

Braithwaite and Braithwaite [6] reported that among their sample of helicopter pilots training in a Lynx simulator, there was a significant positive correlation between self-reported prior history of motion sickness (MHQ) and SS. That is, those with a history of MS were more likely to experience SS in the helicopter simulator. Gower and Fowlkes [14] reported a significant positive correlation between past history of MS as reported on the MHQ and reported SS while training in the Cobra FWS. Gower and Fowlkes [15] also reported a significant positive correlation between reported history of MS (MHQ) and SS among helicopter pilots training in a UH-60 simulator. Gower et al. [16] found this same statistically significant relationship between MHQ scores and early-version SSQ scores for aviators training in a simulator for the CH-47 cargo helicopter.

Gower et al. [16] collected data from 127 rated aviators training in the AH-64 CMS. They found a significant positive correlation between prior history as reported on the MHQ and SS as reported on a MS questionnaire. Kennedy et al. [23] reported the results of surveying 1186 pilots training in 10 Navy simulators. Five of the simulators were FW and five were RW. They reported a small but statistically significant, positive correlation between MHQ scores and SS symptoms. Warner et al. [62] did not find any significant relationship between MHQ scores and SS symptoms for 24 pilots operating two F-16 simulators. Twenty-four (24) participants, however, is usually too small a sample size for a meaningful study of the correlates of SS.

Lampton et al. [36] reported this same relationship for a sample of 115 male trainees operating an M-1 tank driver simulator. Trainees were asked, “Have you ever experienced motion sickness (such as in a car or bus, on a plane or train, on an amusement park ride, seasickness, etc.)?” Twenty-two percent (22%) responded in the affirmative. Those answering yes were significantly more likely to score higher on the SSQ. Lerman et al. [37] assessed 59 male armor Soldiers during tank driver training in a Link tank simulator. The authors reported a significant positive relationship between prior history as measured by the MSQ and SS as measured by a MS questionnaire.

To summarize, two reviewers as well as eight of nine research studies document that a prior history of MS is positively correlated with SS. Past behavior is the best single predictor of future behavior.

8.6 Miscellaneous: Illness, Drugs, Sleep, Fatigue

There are several health-related conditions that are known to influence susceptibility to SS. As with MS, there is the pathology of an absent or non-functional vestibular system. Persons with this pathology (“labyrinthine defectives”) are incapable of experiencing either MS (e.g., Benson, [3]; Reason and Brand, [52]) or SS (e.g., Kennedy and Frank, [29]; Pausch et al., [48]).

It is widely understood among the research community that individuals should not participate in simulator-based training unless they are in their usual state of health and fitness. Individuals in ill health are more susceptible to SS (e.g., Kennedy et al., [25]; Kennedy and Fowlkes, [26]; Kolasinski, [33]; McCauley, [41]; NTSC, [47]; Pausch et al., [48]; Wright, [63]). Symptoms that make individuals more vulnerable include hangover, flu, respiratory illness, head cold, ear infection, ear blockage, and upset stomach. Individuals exhibiting these symptoms should not participate in simulator-based training or simulator-based research [30]. Similarly, it is widely known that certain medications, drugs, and alcohol can increase an aviator’s susceptibility to SS (e.g., Biocca, [5]; Kennedy et al., [25]; Kennedy and Fowlkes [26]; NTSC [47]; Young, [64]).

Reviewers have stated that fatigue and sleep loss also predispose an individual to SS (e.g., Kennedy et al., [25]; NTSC, [47]; Pausch et al., [48]; Wright, [63]). Gower and colleagues (Gower and Fowlkes, [14]; Gower and Fowlkes, [15]; Gower et al., [16]) have repeatedly reported a significant inverse relationship between the numbers of hours slept the previous night and SS as measured on an early version of the SSQ. That is, the fewer the hours slept, the greater the SSQ score. Gower et al. [16] reported a significant negative biserial correlation between self-reported “enough sleep” (yes or no) and SS. That is, those aviators who reported that they had not had enough sleep last night, scored higher on the SSQ. This relationship between fatigue/sleep and SS is no trivial result. In military aviation training it is common for aviators to be less than fully rested during initial, advanced, or recurrent training.

8.7 Simulator Variables

There are several simulator factors that have been implicated as causal in SS. Arguably the two most thorough reviews of these factors can be found in Kolasinski [33] and Pausch et al. [48]. The review presented below is not an exhaustive listing of known simulator variables.

Wide FOV visual displays have long been associated with increased susceptibility to SS [19, 26, 33, 41, 48]. This is because with a wider FOV there is a greater perception of visual flow or vection. Another visual factor with a long history of association with SS is known as off-axis viewing, design eye point, or viewing region [26, 33, 41]. Every visual flight simulator has a design eye point. This is the location within the cockpit where the visual display can be viewed best and where the pilot should keep his or her head positioned. Moving one's head away from the design eye point, or optimal viewing region – by slouching or leaning forward, for example – will not only guarantee a poorer visual image, but will increase one's likelihood of experiencing discomfort. Perhaps the oldest visual factor known to evoke SS (e.g., Miller and Goodson, [44, 45]) is optical distortion caused by misaligned or poorly calibrated optics [12, 26, 33, 37, 41]. Finally, the general issue of cue asynchrony (visual delay, transport delay, asynchronous visual and motion systems) has been investigated as a source of SS, but with equivocal results [19, 33, 41, 48, 61].

8.8 Task Variables

Not surprisingly, what the participant does while in the simulator, and what is done to him or her, can have a marked impact upon susceptibility to SS. These task factors were particularly well presented in the reviews by Kolasinski [33] and McCauley [41]. The review of task variables presented below is not exhaustive.

First in importance is session duration [14, 16, 26, 33, 41, 63]. The longer the period of time spent operating the simulator, the greater the likelihood of significant discomfort. Another important factor is use, by the instructor, of the freeze/reset command [16, 17, 26, 33, 41, 63]. The more often the instructor freezes the pilot in mid-flight – to prevent a crash or provide instruction, for example – the more likely the pilot will experience SS. Other unusual or unnatural maneuvers, such as moving forward/backward in time or flying backwards, are also associated with increased risk of discomfort (Kolasinski).

Maneuver intensity (aggressive, dynamic, or violent maneuvering) has been implicated in SS, both in flight simulators [41, 63] and automobile simulators [19]. Also, the height above terrain at which pilots fly has been shown to vary inversely with discomfort level [16, 26, 33, 63]. Flying close to the ground (nap of the earth) causes more SS than flying at altitude. This is usually explained in terms of greater perception of visual flow, caused by greater visual detail or density, at lower height above terrain. Degree of control has been associated with increased susceptibility to SS [33, 48, 54]. The pilot in control of the simulator tends to report less discomfort than a passive passenger. Finally, head movements increase susceptibility to SS [26, 33, 41, 54]. This last point has long been a part of simulator-trainee lore. Participants, who find themselves vulnerable to SS, quickly learn to keep their heads stationary.

9.0 SIMULATOR SICKNESS, PERFORMANCE, AND TRAINING

9.1 Performance

Does SS harm the flight performance of experienced aviators while in the simulator? Does exposure to a simulator temporarily harm the cognitive, perceptual, or psychomotor performance of the participants? These are not subjects that have received a large amount of research attention.

Silverman and Slaughter [57] stated that 67 percent of the helicopter pilots in their experiment reported modifying their flight control inputs at some point during the simulator sessions to alleviate the symptoms of SS. Pilots reported that they “slowed control inputs” or “transferred controls” or “closed my eyes during rapid aircraft movements” (p. 16). Uliano et al. [61] had 25 experienced male helicopter pilots perform a series of tasks in the VTOL simulator. All pilots were to perform both an air taxi task and a slalom task according to prescribed standards. Performance in executing these tasks to standards was measured in three spatial dimensions (x, y, z). The authors found that there was a statistically significant negative correlation between the amount of SS reported and performance on the air taxi task. Specifically, the sicker were the aviators, the greater the percentage of time they flew out of tolerance in x, y, or z. The authors did not find a statistically significant relationship for the slalom task. Warner et al. [62] assessed 24 pilots flying two F-16 flight simulators through a challenging 50-minute course. They collected 18 objective measures of piloting performance (e.g., airspeed limits, height above ground level, etc.). These they correlated with SSQ scores. The authors found no consistent relationship between SS scores and piloting performance.

As part of their larger survey of Navy simulators Kennedy et al. [23] performed tests of cognitive, perceptual, and psychomotor capabilities. Three tests (Pattern Comparison, Grammatical Reasoning, Speed of Tapping) were administered both before and immediately after simulator exposure.

> Pre- versus post-performance changes were studied in only six different simulators. In no simulator were group performances poorer after exposure, and indeed, most changes showed learning effects from the first (pre) to the second (post) session. Based on interpolations from other experiments on non-pilot subjects, these changes appear within the range of improvements due to practice which are to be expected over two sessions. (Kennedy et al., [23] 5)

Kennedy, Fowlkes, et al. [28] measured performance on three tasks (Pattern Comparison, Grammatical Reasoning, Finger Tapping) both before and after simulator exposure for 411 aviators engaged in simulator-based training. These data were compared to that from a control group of 16 aviators who were not exposed to a simulator between the first (pre) and second (post) test. Both groups showed improvement (a practice effect) from the pre-test to the post-test for all three tasks. However, the improvement shown by the control group was greater than that shown by the simulator-exposed group. This was a small, but statistically significant, difference. In other words, the simulator exposure attenuated the size of the practice effect for the simulator group. Uliano et al. [61] tested 25 experienced male helicopter pilots on a grammatical reasoning task both before and after a 40-minute simulator flight. They reported that there was no statistically significant effect of the simulator flight on performance of the grammatical reasoning task.

Based on the limited evidence that exists, it appears that simulator exposure has little or no effect on the cognitive, perceptual, or psychomotor abilities of aviators. These results are consistent with a larger set of results from the MS literature.

9.2 Training

With the exception of theme parks, simulators are used for training important and often dangerous skills – such as flying a helicopter or driving a tank. Does SS harm this training? For anyone who has experienced simulator-induced discomfort, it certainly appears reasonable to suggest that SS may interfere with training. But does it? What is the evidence?

The fear that SS would limit the usefulness of simulators for flight training has been in existence since the very beginning [44, 45]. In fact, Miller and Goodson reported that use of the device they evaluated was discontinued. Recall also that two of McCauley's four points concerned this issue [41]. He warned of compromised training and decreased simulator use caused by SS.

When researchers review the literature of SS, the possibility of compromised training and/or decreased simulator use is a common feature. At least 15 times between 1986 and 1997 researchers have mentioned this potential problem of simulator-based training [7, 9, 10, 23, 25, 27, 31, 33, 34, 36, 38, 56, 48, 63, 61] goes farther than other reviewers, however, by describing some of the evidence concerning SS and training.

> Although studies indicate that sickness can occur, little – if any – research has investigated whether such sickness has an impact on training effectiveness. (Kolasinski, [34], p. 151)

Given the primacy of this issue since 1958, it is remarkable how little empirical evidence there is on the subject. Chappelow [8] administered questionnaires to 271 Royal Air Force pilots training in either of two air combat simulators. Respondents who had reported sickness symptoms were asked to assess the effect of the experience on their willingness to use the simulator in the future. A total of 214 pilots answered this question. Four percent (4%) reported that the experience decreased their willingness to use the simulator again. Sixty-eight percent (68%) responded that it had no influence. Twenty-eight percent (28%) stated that the experience increased their willingness to use the simulator again, because they said it provided good training and was fun to operate.

Gower and Fowlkes [14] assessed the effect of SS on training by asking their sample of AH-1 pilots whether simulator-induced discomfort hampers training. They found two related results. First, there was a statistically significant positive correlation between SSQ scores and agreement with the statement that "discomfort hampers training." That is, the aviators who reported the most SS were more likely to agree that discomfort harms training. Second, only 8 percent of their sample agreed, "discomfort hampers training." Four percent (4%) were neutral on the question. Eighty-eight percent (88%) disagreed with the statement. It should be noted that these results were the self-reported opinions of Army aviators. No grades, test results, set-backs, training hours required, or other performance measures were presented to show in an objective fashion that, in fact, those participants experiencing more discomfort learned less than their non-sick counterparts.

Gower and Fowlkes [15] asked the same questions of their sample of UH-60 pilots and found the same pattern of results. First, there was a statistically significant positive correlation between SSQ scores and agreement with the statement that "discomfort hampers training." Second, this was the opinion of a small minority of their sample. Only 1 person (1%) of the 86 who answered this question agreed that discomfort disrupts training. Fifteen percent (15%) were neutral. Eighty-four percent (84%) disagreed with the statement. Again, no data on performance during training were collected that would bear on the issue of SS and amount learned.

Gower et al. [16] found the same pattern of results with their sample of helicopter pilots training in the CH-47 flight simulator. There was a significant positive correlation between SSQ scores and agreement with the statement that "discomfort hampers training." Again, only 1 person (1.5%) agreed with the statement. Two people were neutral (2.9%). Of the total of 68 responses to this question, 65 (95.6%) disagreed with the statement. Finally, as before, no performance data were presented as to SS and amount learned during training.

The results of these four questionnaire studies are clear. The vast majority of the aviators surveyed stated that the discomfort-producing potential of the devices did not detract from the training provided. However, a small minority of aviators – those experiencing the most sickness – held the opposite opinion. Given the centrality of this issue for simulator-based training, more research should be undertaken. Measures of performance in learning the required program of instruction should be correlated with measures of SS such as the SSQ.

In agreement with the quote from Kolasinski [34] above, the present author knows of no published research devoted to this question.

10.0 TREATMENT

As with MS, the surest treatment for SS is simple adaptation. Nearly everyone will adapt to a particular simulator eventually. To aid adaptation to a new simulator, aviators should begin with brief simulator hops, flying gentle maneuvers, with subsequent hops separated by one-day intervals (NTSC, 1988). In this context, "brief" means less than one hour, with breaks as needed. The maximum duration of any simulator session should never exceed two hours. Several other guidelines exist and will be described later in this report.

For those pilots who cannot adapt to a simulator, "… anti-motion sickness medication may be considered for the simulator period" (Crowley,[9], p. 357). Drugs previously used to control the symptoms of MS, such as hyoscine hydrobromide and dimenhydrinate (Dramamine), have also proven effective for relief of SS [3, 52, 53]. In the world of flight training, it is no secret that some aviators with a history of discomfort self-medicate with MS drugs prior to a simulator session. However, no drug can reduce the occurrence of SS for everyone. Further, every drug has side effects. For example, scopolamine administered as a treatment for SS is known to have side effects that could negatively affect learning [9]. An aviator with severe, intractable SS should visit his or her flight surgeon.

11.0 THEORY

SS is a form of MS. The two major theories that exist to explain MS are also used to explain SS. By far the more common is the sensory conflict theory [3, 50, 51, 52]. Virtually all research reports mention the sensory conflict theory by one of its names. Most authors employ it in the explication of their results. Early examples of how this theory has been applied to SS can be found in Kennedy and Frank [29], McCauley [41], and Reason and Brand [52]. The major competitor is the postural instability theory [54, 58, 59]. For a more detailed description of these two theories please see the discussion presented in the Motion Sickness section above.

11.1 Sensory Conflict Theory

The sensory conflict (SC) theory states that sensory inputs from the eyes, semicircular canals, otoliths, proprioceptors, and somatosensors are provided in parallel both to a neural store of past sensory patterns of spatial movement and to a comparator unit. This comparator unit compares the present pattern of motion information with that pattern expected based on prior motion history and stored in the neural store. A mismatch between the current pattern and the stored pattern generates a mismatch signal. This mismatch signal initiates both SS and the process of adaptation.

According to the SC theory, when an aviator is operating a new simulator the pattern of motion information presented by the senses is at variance with past experience in the flight environment. This conflict between the current sensory pattern and that pattern expected based upon past experience causes SS. That is, there is a conflict between the current novel motion environment and past experience. However, with continued sessions operating the device the relative mismatch between current pattern and stored patterns decreases until one has adapted. Flight simulators attempt to simulate flight – that is, to trick the human perceptual system. However, no device can perfectly simulate all the physical forces of flight. It is this inability to simulate flight perfectly that causes SS in experienced aviators.

However, one need not be an aviator to know the discomfort of SS. Anyone with a normal vestibular system is susceptible to SS when operating a novel motion simulator. The key concept is the mismatch between the novel motion environment (the current pattern of sensory stimulation in the simulator) and prior motion history (the patterns of sensory stimulation resident in the neural store). As the reader can see, the SC theory explains SS in exactly the same fashion it explains MS – only the motion environment has changed.

11.2 Postural Instability Theory

The PI theory notes that sickness-producing situations are characterized by their unfamiliarity to the participant. This unfamiliarity sometimes leads to an inability of the participant to maintain postural control. It is this postural instability that causes the discomfort – until the participant adapts. That is, a prolonged exposure to a novel motion environment causes postural instability that precedes and causes the sickness.

PI theory states that there are individual differences in postural stability. Further, an imposed motion presented by a simulator can induce postural instability. The interaction of the body's natural oscillation with the imposed oscillation created by the simulator leads to a form of wave interference effect that causes postural instability. This instability is the proximate cause of SS. Experimental evidence in support of this theory – from participants exposed to simulated motion – has been reported [58, 59]. The PI theory explains SS in exactly the same fashion it explains MS – only the nature of the novel motion environment has changed.

11.3 SS, Age, and Theory

The SC theory and the PI theory make different predictions in some instances. A few examples of these differences are presented earlier in this report in the Motion Sickness section. One issue on which these two competing theories make diametrically opposite predictions concerns the effect of age on susceptibility to SS.

The SC theory states that MS in all its forms must decline with age after about age 12. The reasons for this are that life experiences provide the neural store with a wealth of prior sensorimotor patterns of motion memories and also that receptivity (the strength of the mismatch) declines with age. The SC theory predicts that SS will decline with age. However when research shows that SS increases with age, these results are dismissed as being the product of a confounding with flight experience. Age and flight experience are strongly correlated among pilots. The SC theory predicts that with increasing flight hours the relative mismatch between the sensorimotor pattern of aircraft flight and that of simulator "flight" will be greater and will, therefore, engender more SS. However, this interpretation only exists because the overwhelming majority of simulator research has taken place in the world of aviator training – a world where older aviators are also more experienced aviators. The SC theory would have to predict that a large sample of adult non-aviators of widely different ages would show *decreasing* SS with increasing age.

The PI theory would have to make exactly the opposite prediction. Unlike the SC theory, the PI theory is stated in a way that allows it to be scientifically tested and falsified. According to this theory, SS is caused by postural instability. Postural stability among adults is known to decline with increasing age (e.g., Kane et al., [22]; Lyon, [39]). Therefore, PI theory would predict that a large sample of adult non-aviators of widely different ages would show *increasing* SS with increasing age. Further, within any age cohort this theory predicts that greater instability will be associated with greater SS. So this theory not only makes a general prediction concerning age, but also makes a prediction concerning specific aged adults.

It is not an everyday occurrence in science that two competing theories make precisely opposite predictions. The test suggested above could add to the theoretical understanding of all motion sickness phenomena. Again, the most difficult parts of this experiment would be to gain access both to a large sample of adult non-aviators, as well as to the simulator itself.

12.0 GUIDELINES FOR REDUCING SIMULATOR SICKNESS AND RISKS FROM AFTEREFFECTS

Several authors have taken the time to publish guidelines in an effort to reduce the rate of SS among trainee populations [6, 10, 25, 33, 38, 41, 47, 63]. Arguably the most thorough set of guidelines are those by Kennedy et al. and Wright. These authors not only provide guidelines, but also explain the reasons for the guidelines and the evidence supporting them. If the reader does not have time for a detailed presentation, the best and most entertaining single source is the field manual published by the Naval Training Systems Center (NTSC).

The temptation to include guidelines of one’s own is almost impossible for authors to resist. This is not only because SS is so discomforting to one’s trainees, but also because some policies and procedures are clearly better than others. So in the interests of preventing future discomfort the current author will list some suggestions. This is by no means an exhaustive listing.

12.1 General Rules

- Simulator flights should not be scheduled on the same day as aircraft flights.
- Arrive for simulator training in your usual state of health and fitness.
  - Avoid fatigue or sleep loss, hangover, upset stomach, head colds, ear infections, ear blockages, upper respiratory illness, medications, and alcohol.
  - If you have been sick recently and are not fully recovered, reschedule your simulator training.
- Persons who are new to the simulator, or who have not operated it in months, are at risk.
- Do not schedule simulator sessions for greater than two hours for any reason.
  - Use breaks, time-outs extensively.
  - The more nauseogenic the session, the shorter the session should be.
    - Aggressive, violent maneuvers, near ground level, are more nauseogenic than high, straight-and-level flight.
- Adaptation is one of the most potent fixes for SS.
  - In order to optimize adaptation, there should be a minimum of one day between simulator sessions, and a maximum of seven days.
  - Begin with short sessions, using non-nauseogenic maneuvers.
  - Minimize rapid gain and loss in altitude; minimize abrupt or continued roll; minimize porpoising.
  - Fly the most provocative tasks at the end of the session.
- Minimize head movement, particularly when new or dynamic maneuvers are being trained.
- Tell your instructor if you are experiencing discomfort.

- The instructor should avoid, or at least minimize, use of the freeze command.
  - Have the pilot close his or her eyes before using the freeze command.
  - Have the pilot close his or her eyes before resetting the simulator to another location. Or, turn off visual display before reset.
- The instructor should turn off visual display and turn on cabin lights before asking the pilot to exit the simulator.
- The instructor should decrease the field of view (turn off side displays) during early sessions, nauseogenic maneuvers, or if the pilot shows any symptoms of discomfort.
  - Or, go on instruments at the first sign of discomfort.
- Avoid high-risk activities for at least 12 hours after simulator training.
  - High-risk activities include flying, climbing, driving, riding motorcycles, riding bicycles, or diving.
  - Use handrails to help maintain balance when going up or down stairs.

13.0 SUGGESTIONS FOR FUTURE RESEARCH

This review has uncovered at least two areas where further research into the subject of SS is clearly warranted.

- *The effect of SS on training.* As this review has shown repeatedly, one of the key arguments offered for studying SS is the potential for compromised training. However, there is virtually no evidence to support this argument. There is no evidence showing a statistically significant and substantial difference in the amount learned as a function of reported level of discomfort. Given that most simulator-based research takes place at aviation training sites, this oversight is particularly curious. This research topic is important and should be examined in a quantitative empirical fashion.
- *The effect of chronological age on SS.* Does increasing adult age make one more susceptible to SS or less susceptible? Are older aviators more susceptible to SS because they are older, because they have more flight experience, or some combination of both? Perhaps the best reason to investigate this subject parametrically is because the two leading theories of SS make precisely opposite predictions. The SC theory predicts that SS will decrease with increasing chronological age. The PI theory predicts that SS will increase with increasing chronological age. Thus, performing this research has the added benefit of increasing our theoretical understanding of SS.

14.0 REFERENCES

[1] AGARD. (1988). Motion cues in flight simulation and simulator induced sickness (AGARD Conference Proceedings No. 433). Neuilly-sur-Seine, France: Advisory Group for Aerospace Research and Development.

[2] Baltzley, D.R., Kennedy, R.S., Berbaum, K.S., Lilienthal, M.G. and Gower, D.W. (1989). The time course of postflight simulator sickness symptoms. Aviation, Space, and Environmental Medicine, 60, 1043-1048.

[3] Benson, A.J. (1978). Motion sickness. In: G. Dhenin and J. Ernsting (Eds.), Aviation Medicine (pp. 468-493). London: Tri-Med Books.

[4] Benson, A.J. (1988). Aetiological factors in simulator sickness. In: AGARD, Motion cues in flight simulation and simulator induced sickness (AGARD Conference Proceedings No. 433, pp. 3.1-3.8). Neuilly-sur-Seine, France: Advisory Group for Aerospace Research and Development.

[5] Biocca, F. (1992). Will simulation sickness slow down the diffusion of virtual environment technology? Presence, 1(3), 334-343.

[6] Braithwaite, M.G. and Braithwaite, B.D. (1990). Simulator sickness in an Army simulator. Journal of the Society of Occupational Medicine, 40, 105-110.

[7] Casali, J.G. and Frank, L.H. (1988). Manifestation of visual/vestibular disruption in simulators: Severity and empirical measurement of symptomatology. In: AGARD, Motion cues in flight simulation and simulator induced sickness (AGARD Conference Proceedings No. 433, pp. 11.1-11.18). Neuilly-sur-Seine, France: Advisory Group for Aerospace Research and Development.

[8] Chappelow, J.W. (1988). Simulator sickness in the Royal Air Force: A survey. In: AGARD, Motion cues in flight simulation and simulator induced sickness (AGARD Conference Proceedings No. 433, pp. 6.1-6.11). Neuilly-sur-Seine, France: Advisory Group for Aerospace Research and Development.

[9] Crowley, J.S. (1987). Simulator sickness: A problem for Army aviation. Aviation, Space, and Environmental Medicine, 58, 355-357.

[10] Crowley, J.S. and Gower, D.W. (1988). Simulator sickness. United States Army Aviation Digest, 1-88-11, 9-11.

[11] Duh, H.B., Parker, D.E. and Furness, T.A. (2001). An "independent visual background" reduced balance disturbance evoked by visual scene motion: Implication for alleviating simulator sickness. Paper presented at the SIGCHI '01 conference, March 31 – April 4, 2001, Seattle, WA.

[12] Ebenholtz, S.M. (1992). Motion sickness and oculomotor systems in virtual environments. Presence, 1(3), 302-305.

[13] English, H.B. and English, A.C. (1958). A comprehensive dictionary of psychological and psychoanalytical terms. NY: David McKay.

[14] Gower, D.W. and Fowlkes, J.E. (1989a). Simulator sickness in the AH-1S (Cobra) flight simulator (USAARL Report No. 89-20). Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory.

[15] Gower, D.W. and Fowlkes, J.E. (1989b). Simulator sickness in the UH-60 (Black Hawk) flight simulator (USAARL Report No. 89-25). Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory.

[16] Gower, D.W., Fowlkes, J.E. and Baltzley, D.R. (1989). Simulator sickness in the CH-47 (Chinook) flight simulator (USAARL Report No. 89-28). Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory.

[17] Gower, D.W., Lilienthal, M.G., Kennedy, R.S., Fowlkes, J.E. and Baltzley, D.R. (1987). Simulator sickness in the AH-64 Apache Combat Mission Simulator (USAARL Report No. 88-1). Fort Rucker, AL: USAARL.

[18] Hamilton, K.M., Kantor, L. and Magee, L.E. (1989). Limitations of postural equilibrium tests for examining simulator sickness. Aviation, Space, and Environmental Medicine, 60, 246-251.

[19] Hein, C.M. (1993). Driving simulators: Six years of hands-on experience at Hughes Aircraft Company. Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting 1993, 607-611.

[20] Hettinger, L.J., Nolan, M.D., Kennedy, R.S., Berbaum, K.S., Schnitzius, K.P. and Edinger, K.M. (1987). Visual display factors contributing to simulator sickness. Proceedings of the Human Factors Society 31st Annual Meeting 1987, 497-501.

[21] Johnson, D.M. (in preparation). Helicopter simulator sickness: Incidence, age, flight experience, and training effectiveness (ARI Tech. Rep.). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[22] Kane, R.L., Ouslander, J.G. and Abrass, I.B. (1994). Essentials of clinical geriatrics. NY: McGraw-Hill.

[23] Kennedy, R.S., Berbaum, K.S., Allgood, G.O., Lane, N.E., Lilienthal, M.G. and Baltzley, D.R. (1988). Etiological significance of equipment features and pilot history in simulator sickness. In: AGARD, Motion cues in flight simulation and simulator induced sickness (AGARD Conference Proceedings No. 433, pp. 1.1-1.22). Neuilly-sur-Seine, France: Advisory Group for Aerospace Research and Development.

[24] Kennedy, R.S., Berbaum, K.S. and Lilienthal, M.G. (1997). Disorientation and postural ataxia following flight simulation. Aviation, Space, and Environmental Medicine, 68, 13-17.

[25] Kennedy, R.S., Berbaum, K.S., Lilienthal, M.G., Dunlap, W.P., Mulligan, B.E. and Funaro, J.F. (1987). Guidelines for alleviation of simulator sickness symptomatology (NAVTRASYSCEN TR-87-007). Orlando, FL: Naval Training Systems Center.

[26] Kennedy, R.S. and Fowlkes, J.E. (1992). Simulator sickness is polygenic and polysymptomatic: Implications for research. International Journal of Aviation Psychology, 2(1), 23-38.

[27] Kennedy, R.S., Fowlkes, J.E., Berbaum, K.S. and Lilienthal, M.G. (1992). Use of a motion sickness history questionnaire for prediction of simulator sickness. Aviation, Space, and Environmental Medicine, 63, 588-593.

[28] Kennedy, R.S., Fowlkes, J.E. and Lilienthal, M.G. (1993). Postural and performance changes following exposures to flight simulators. Aviation, Space, and Environmental Medicine, 64, 912-920.

[29] Kennedy, R.S. and Frank, L.H. (1985). A review of motion sickness with special reference to simulator sickness (Tech. Rep. NAVTRAEQUIPCEN 81-C-0105-16). Orlando, FL: Naval Training Equipment Center.

[30] Kennedy, R.S., Lane, N.E., Berbaum, K.S. and Lilienthal, M.G. (1993). Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. International Journal of Aviation Psychology, 3(3), 203-220.

[31] Kennedy, R.S., Lane, N.E., Lilienthal, M.G., Berbaum, K.S. and Hettinger, L.J. (1992). Profile analysis of simulator sickness symptoms: Application to virtual environments systems. Presence, 1(3), 295-301.

[32] Kennedy, R.S., Lilienthal, M.G., Berbaum, K.S., Baltzley, D.R. and McCauley, M.E. (1989). Simulator sickness in U.S. Navy flight simulators. Aviation, Space, and Environmental Medicine, 60, 10-16.

[33] Kolasinski, E.M. (1995). Simulator sickness in virtual environments (ARI Tech. Rep. 1027). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[34] Kolasinski, E.M. (1997). Predicted evolution of virtual reality. In: National Research Council (Ed.), Modeling and simulation: Linking entertainment and defense (pp. 149-153). Washington, D.C.: National Academy Press.

[35] Kolasinski, E.M. and Gilson, R.D. (1999). Ataxia following exposure to a virtual environment. Aviation, Space, and Environmental Medicine, 70, 264-269.

[36] Lampton, D.R., Kraemer, R.E., Kolasinski, E.M. and Knerr, B.W. (1995). An investigation of simulator sickness in a tank driver trainer (ARI Research Rep. 1684). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[37] Lerman, Y., Sadovsky, G., Goldberg, E., Kedem, R., Peritz, E. and Pines, A. (1993). Correlates of military tank simulator sickness. Aviation, Space, and Environmental Medicine, 64, 619-622.

[38] Lilienthal, M.G., Kennedy, R.S., Berbaum, K.S., Dunlap, W.P. and Mulligan, B.E. (1987). Vision/motion-induced sickness in Navy flight simulators: Guidelines for its prevention. Proceedings of the 1987 Image Conference IV, 23-26 June 1987, 275-285.

[39] Lyon, M.J. (2003, October). Aging vestibular system. SUNY Upstate Medical University: Department of Otolaryngology and Communication Sciences. Retrieved from http://www.upstate.edu/ent/faculty.php

[40] Magee, L.E., Kantor, L. and Sweeney, D.M.C. (1988). Simulator induced sickness among Hercules aircrew. In: AGARD, Motion cues in flight simulation and simulator induced sickness (AGARD Conference Proceedings No. 433, pp. 5.1-5.8). Neuilly-sur-Seine, France: Advisory Group for Aerospace Research and Development.

[41] McCauley, M.E. (Ed.). (1984). Research issues in simulator sickness: Proceedings of a workshop. Washington, D.C.: National Academy Press.

[42] McGuinness, J., Bouwman, J.H. and Forbes, J.M. (1981). Simulator sickness occurrences in the 2E6 Air Combat Maneuvering Simulator (ACMS) (Tech. Rep. NAVTRAEQUIPCEN 80-C-0135-4500-1). Orlando, FL: Naval Training Equipment Center.

[43] Merriam-Webster Editorial Staff. (1971). Webster's third new international dictionary. Springfield, MA: Merriam Co.

[44] Miller, J.W. and Goodson, J.E. (1958). A note concerning 'motion sickness' in the 2-FH-2 hover trainer (Project NM 1701-11, Subtask 3, Rep. 1). Pensacola, FL: U.S. Naval School of Aviation Medicine.

[45] Miller, J.W. and Goodson, J.E. (1960). Motion sickness in a helicopter simulator. Aerospace medicine, 31(3), 204-212.

[46] Mooij, H.A. (1988). Technology involved in the simulation of motion cues: The current trend. In: AGARD, Motion cues in flight simulation and simulator induced sickness (AGARD Conference Proceedings No. 433, pp. 2.1-2.15). Neuilly-sur-Seine, France: Advisory Group for Aerospace Research and Development.

[47] NTSC (1988). Simulator sickness field manual mod 3. Orlando, FL: Naval Training Systems Center.

[48] Pausch, R., Crea, T. and Conway, M. (1992). A literature survey for virtual environments.

[49] Military flight simulator visual systems and simulator sickness. Presence, 1(3), 344-363.

[50] Reason, J.T. (1970). Motion sickness: A special case of sensory rearrangement. Advancement of Science, 26, 386-393.

[51] Reason, J.T. (1978). Motion sickness adaptation: A neural mismatch model. Journal of the Royal Society of Medicine, 71, 819-829.

[52] Reason, J.T. and Brand, J.J. (1975). Motion sickness. London: Academic.

[53] Regan, E.C. and Ramsey, A.D. (1996). The efficacy of hyoscine hydrobromide in reducing side-effects induced during immersion in virtual reality. Aviation, Space, and Environmental Medicine, 67, 222-226.

[54] Riccio, G.E. and Stoffregen, T.A. (1991). An ecological theory of motion sickness and postural instability. Ecological Psychology, 3(3), 195-240.

[55] Riley, J. (2004, April 26). Personal communication.

[56] Schneider, M. (2004, May 6). Ride sends six people to hospital in eight months. The Dothan Eagle, p. B2.

[57] Silverman, D.R. and Slaughter, R.A. (1995). An exploration of simulator sickness in the MH-60G operational flight trainer, an advanced wide field-of-view helicopter trainer (Rep. No. AL/HR-TR-1994-0173). Mesa, AZ: Aircrew Training Research Division, Human Resources Directorate.

[58] Stoffregen, T.A., Hettinger, L.J., Haas, M.W., Roe, M.M. and Smart, L.J. (2000). Postural instability and motion sickness in a fixed-base flight simulator. Human Factors, 42(3), 458-469.

[59] Stoffregen, T.A. and Smart, L.J. (1998). Postural instability precedes motion sickness. Brain Research Bulletin, 47, 437-448.

[60] Tsang, P.S. (2003). Assessing cognitive aging in piloting. In: P.S. Tsang and M.A. Vidulich (Eds.), Principles and practice of aviation psychology (pp. 507-546). Mahwah, NJ: Erlbaum.

[61] Uliano, K.C., Lambert, E.Y., Kennedy, R.S. and Sheppard, D.J. (1986). The effects of asynchronous visual delays on simulator flight performance and the development of simulator sickness symptomatology (NAVTRASYSCEN 85-D-0026-1). Orlando, FL: Naval Training Systems Center.

[62] Warner, H.D., Serfoss, G.L., Baruch, T.M. and Hubbard, D.C. (1993). Flight simulator-induced sickness and visual displays evaluation (AL/HR-TR-1993-0056). Williams Air Force Base, AZ: Aircrew Training Research Division.

[63] Wright, R.H. (1995). Helicopter simulator sickness: A state-of-the-art review of its incidence, causes, and treatment (ARI Res. Rep. 1680). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[64] Young, L.R. (2003). Spatial orientation. In: P.S. Tsang and M.A. Vidulich (Eds.), Principles and practice of aviation psychology (pp. 69-113). Mahwah, NJ: Erlbaum.

Dynamic Situations: The Soldier's Situation Awareness

John P. Holmquist
University of Central Florida
Orlando, Florida
USA

Stephen L. Goldberg
U.S. Army Research Institute
Orlando, Florida
USA

The role of the Soldier is not as strictly defined today, as it was in World War II. In place of battlefields, the Soldier is placed in roles from peace keeper to combat Soldier in collapsed first world counties to impoverished third world nations. This fogs the understanding of situations and the role the Soldier is to play in them. Within these changing and dynamic times, a Soldier's situation awareness has become vitally important. Understanding who are combatants, civilians, and allied personnel, as well as, knowing the rules of engagement for the given situation, are all part of Soldiers' situation awareness.

1.0 SITUATION AWARENESS

Researchers over the last two decades have continued to narrow the definition of situation awareness and apply its concepts to different circumstances and personnel. The term situation awareness (SA) has been used with pilots, air traffic controllers, fire fighters, and others who are involved in situations that require quick decisions under stress [11].

A popular definition of situation awareness, offered by Endsley [7], is perception of the elements in the environment within a volume of space and time, the comprehension of their meaning, and the projection of their status in the near future. This was simplified by Howell [14] to read SA involves an operator keeping track of a lot of information from a variety of sources over time and organizing or interpreting this information. Later, Endsley [6] expanded her definition into a model of situation awareness including three levels. The first is a perception of the elements of the current situation. This is an understanding of the physical environment a person is in. The second is a comprehension of the situation. Here the dynamics of the physical elements and people in the situation must be understood, in terms of their movement and purpose. Third, is the projection of future status of the situation. Situation awareness occurs over time; therefore, the effect on current events on the near future is the last level of this definition of situation awareness. SA encompasses not only an awareness of key elements in the situation, it encompasses a gestalt ('big picture') comprehension and integration of that information in light of operational goals, along with the ability to project future states of the system. These higher levels of SA, gestalt understanding of the situation and future prediction, have been found to critical to effective functioning in complex environments, such as those faced by Soldiers [7]. Furthermore, situation awareness, according to the U.S. Army Training and Doctrine Command (TRADOC), is defined as "the ability to have accurate real-time information of friendly, enemy, neutral, and non-combatant locations; a common, relevant picture of the battlefield scaled to specific levels of interest and special needs." This final definition is pertinent to all Soldiers on all battlefields.

2.0 SA AND DECISION MAKING

With an understanding of the Soldiers current situation awareness, application of SA in decision making becomes vital. In order to make good decisions in the combat environment it is necessary to make an accurate assessment of the situation [32].

An area of current research that implements SA in decision making is naturalistic decision making [28, 18]. The Soldier in the field must be prepared to make split-second decisions that could save or lose lives.

One method of planning for split-second decision making is recognition-primed decisions (RPDs). Klein, Calderwood, and Clinton-Cirocco [20] presented the recognition-primed decision model that describes how decision makers can recognize a plausible course of action (COA) as the first one to consider. A commander's knowledge, training, and experience generally help in correctly assessing a situation and developing and mentally war-gaming a plausible COA, rather than taking time to deliberately and methodically contrast it with alternatives using a common set of abstract evaluation dimension. RPDs are hypothesized to work well in naturalistic decision making which encompasses environments with time constraints, changing conditions, and stress, [19]. The findings of RPDs and the ability to make better decisions with RPDs was based, in part, on better situation awareness. Researchers in this area have found that skilled decision-makers usually make a good COA on the first try and that if they change to a secondary choice it is usually not as good as their first choice [22, 16].

In a similar study, Kaempf, Klein, Thordsen, and Wolf [17] investigated how SA influences decision making in a Navy Combat Information Center (CIC) and found that SA is an important factor in decision quality. Furthermore, fluidity of the situations and the incompleteness of available information ensure that the problems attacked by natural decision making are inherently ill-defined [21] which is the exact environment that today's Soldiers find themselves in. RPDs involve an assessment of the situation, recognition of events as typical, and a resultant course of action based on previous experience. A number of features distinguish the RPDs model from classical decision models [20]. These include preplanning of decisions to a given situation. This is a point were rules of engagement need to be clear and defined to allow RPDs to not be hindered by cognitive distance or confusion.

2.1 Uniform Battlefield Decision Making

The US Army has a formal process for planning military operations called the Military Decision-Making Process (MDMP) [4]. This process is long and guided as shown in the seven steps listed in the following table. While MDMP is good for organizational and course of action planning, it does not allow for quick decision making that is needed for Soldiers on the battlefield in combat.

Table 6-1: Steps of MDMP

1. Receipt of mission
2. Mission analysis
3. Course of action
4. Course of action analysis
5. Course of action comparison
6. Course of action approval
7. Orders production

Decision making while units are in combat is characterized by the requirement to make decisions quickly. Often, commanders are bombarded with large amounts of information in various forms and must attempt to form a mental model of the situation to use as a basis for decisions. Even if information is organized and rationally chunked together, the limits of working memory preclude decision-makers from considering all of the information available [33]. This leads back to the benefits of RPDs and the functionality of the first COA of skilled decision-makers as viable courses of action. Past researchers have lead to the same conclusion, that decision making may benefit from following the Recognition Primed Decision-Making model described by Klein [1, 16].

3.0 EXPERIENCE IN SA

More experienced officers demonstrate superior skills in decision making [15, 29, 31]. Klein [19] stresses the importance of situational assessment and the experience of the decision-maker in evaluating the shortcomings of a course of action. In an experiment of identifying locations of units on a battlefield, experienced officers could identify significantly more locations of their own and enemy troops than less experienced officers. Furthermore, experienced officers identify the strongest enemy locations and areas of highest enemy threat, which the less experienced offices could not do [31]. Research suggests that some of the differences between experts and novices in decision making may be due to a difference in the ability to perceive meaningful patterns [30] and to associate certain actions with those patterns [25]. Experts have been shown to use visually-based schema that are specific to their area of expertise [15]. While situation assessment by a skilled worker appears to take place very quickly, the basis for it is built up by continual appraisal [29]. Therefore, the sooner a Soldier can become aware of the forming of patterns in a given situation, the sooner RPDs can be initiated to correctly deal with the situation.

Researchers have indicated that the similarity of trainees' knowledge structure to an expert structure was correlated with skill acquisition and was predictive of skill retention and skill transfer [2]. Training to increase a novice's ability to quickly and accurately assess battlefield situations comes from experience with a variety of situations. Experience alone is not the best teacher, but rather experience with appropriate feedback from an expert coach or mentor. Experience can be gained through training. Effective training can take place in a number of different ways, reading books, participating in field exercises and through use of virtual and constructive training systems [12]. Virtual simulations have been shown as effective means for training decision making and situational assessment [9, 26, 23] and have the advantages of reduced cost, capability to display multiple physical locations, accurate After Action Review capabilities, and less time spent on logistics over training in the field. Virtual simulation provides an opportunity for Soldiers and leaders to go through more scenarios in a given block of time.

4.0 DIGITAL SYSTEMS TO ENHANCE SA

The US Army has a simple definition for situation awareness. SA is seen as the commander's understanding of the battlefield [4]. Frequently the term is used to describe information available on Tactical Operations Center (TOC) displays or SA displays. The purpose of the displays is to provide decision-makers with enough information about what is occurring and likely to occur to make quality decisions. Digitization programs seek to capitalize on networked computer systems to enhance information flow to produce a better common operational picture (COP). Theoretically, this allows decision-makers to maintain a clear, accurate, and shared vision of the battlefield necessary to support both planning and execution [27].

In an Army unit equipped with digital systems, information is typically stored in common databases and can be accessed through a tactical internet, much like the World Wide Web Internet. Much of the information, such as unit positions, can be displayed spatially as graphics, which is much easier to process cognitively allowing for possibly quicker situation awareness and a course to a quicker COA for decision-makers. Through the application of advanced technology on the battlefield, the U.S. Army is well on its way to establishing full situational awareness [3] for the Soldier and of the battlefield. The use of digital automated systems to increase situation awareness is a promising method to allow decision-makers to develop a more accurate mental model of the situation, and consequently increase the quality of decisions [13].

Digital networks allows commanders to maintain an awareness of their subordinate units, known as friendly SA. In mechanized units, for example, each vehicle tracks its geographical position by means of a Global Positioning System (GPS) receiver. Periodically, its position is transmitted back to the unit network where it can be displayed on the commander's computer. This ensures the commander knows the location of all the vehicles in the unit, at all times. A study conducted by McGuinness and Foy [24] found commanders rated this factor to be one of the most helpful for maintaining situation awareness.

As well as friendly SA, commanders also require SA concerning the enemy. Military Intelligence specialists filter advanced imagery and reports to locate enemy units and enter information into the database. Once the location of enemy units on the battlefield can be accurately presented, commanders can recognize patterns of activity and estimate the enemy's intent. With this information, the commander's options become clearer.

This timely sharing of information allows better coordination among units and significantly improves the ability of commanders and leaders to make decisions quickly [3].

5.0 REFERENCES

[1] Adelman, L., Leedom, D.K., Murphy, J. and Killam, B. (1998). Description of Brigade C2 Decision Process. (Army Research Laboratory Technical Report). Aberdeen Proving Grounds, MD: Army Research Institute.

[2] Day, E., Arthur, W. and Gettman, D. (2001). Knowledge structures and the acquisition of a complex skill. Journal of applied Psychology, 85 (6), 1022-1033.

[3] Department of Defense (2000, April). Report on the Plan for Fielding the First Digitized Division and First Digitized Corps: *Presented to the Committee on Army Services, United States Senate, Second Session, 106th Congress Washington: DC. Author.*

[4] Department of the Army (1997). *Field Manual 101-5, Staff Organization and Operations*. Washington, DC: Author.

[5] Endsley, M. (1988). Design and evaluation for situation awareness enhancement. J. Randel et al. 596 *Proceedings of the human factors society, 32nd annual meeting,* pp. 97-101. Santa Monica, CA: the human factors society.

[6] Endsley, M. (1995). Theoretical underpinnings of situation awareness: A critical review. *Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness, Daytona Beach, FL November 1-3 1995*. Daytona Beach: Embry-Riddle Aeronautical University.

[7] Endsley, M. (1999). Situation awareness and human error: designing to support human performance. *Proceedings of the High Systems Surety Conference,* Albuquerque, NM.

[8] Endsley, M., Holder, L., Leibrecht, B., Garland, D., Wampler, R. and Matthews, M. (2000). Modeling and measuring situation awareness in the infantry operational environment. U.S. Army Research Institute: Alexandria, VA, Tech Report 770.

[9] Finkelstein, N. (2000). Charting the retention of tasks learned in synthetic environments. Dissertation Abstracts International: Section B: The Sciences and Engineering. Univ Microfilms International, US, 60 (11-B), 5702.

[10] Garland, D., Phillips, J., Tilden, D. and Wise, J. (1991). Theoretical underpinnings of situation awareness: Towards an objective measure. (Tech. Rep. No. CAAR-15498-91-1). Daytona Beach, FL: Embry-Riddle Aeronautical University, Center for Aviation/Aerospace Research.

[11] Gilson, R., Garland, D. and Koonce, J. (1994). Situational Awareness in Complex Systems. Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

[12] Gorman, P.F. (1991). The Future of Tactical Engagement Simulation. In D. Pace (Ed.), Proceedings of the 1991 Summer Computer Simulation Conference. Baltimore MD: Society for Computer Simulation.

[13] Holmquist, J. and Barnett, J. (2001). Digitally enhanced situation awareness as an aid to military decision making. *Proceedings of the 45th Annual Human Factors and Ergonomics Meeting.*

[14] Howell, W.C. (1993). Engineering psychology in a changing world. Annual Review of Psychology, 44, 231-263.

[15] Hunt, E. (1989). Cognitive science: definition, status, and questions. Annual Review of Psychology 40, 603-629.

[16] Johnson, J. and Raab, M. (2004). "Take the First: Option Generation and Resulting Choices," Organizational Behavior and Human Decision Processes (In press).

[17] Kaempf, G., Klein, G., Thordsen, M. and Wolf, S. (1996). Decision making in complex naval command and control environments. *Human Factors 38*(2) 220-231.

[18] Klein, G. (1989). Strategies of decision making. *Military Review,* May, 1989, 56-64.

[19] Klein, G. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. Klien, J. Rasanu, R. Alderwood and C. Sambok, Eds. *Decision Making in Action: Models and Method,* pp. 306-326. Norwood, NJ: Ablex Publishing Corp.

[20] Klein, G., Calderwood, R. and Clinton-Cirocco, A. (1986). Rapid Decisionmaking on the Fireground, proceedings, *Human Factors and Ergonomics Society 30th Annual Meeting*, Dayton, Ohio, 1, 576-8.

[21] Klein, G., Orasanu, J., Alderwood, R. and Zsambok, C. (1993). Decision Making in Action: Models and Methods. Norwood , NJ : Ablex Publishing Corp.

[22] Klein, G., Wolf, S., Militello, L. and Zsambok, C. (1995). *Characteristics of Skilled Option Generation in Chess*, Organizational Behavior and Human Decision Processes: 62, 63-69.

[23] Lampton, D.R., Riley, J., Kaber, D. and Endsley, M. (2006). Use of immersive virtual environments for measuring and training situational awareness. *Paper presented at the U.S. Army Science Conference*, Orlando, FL.

[24] McGuinness, B. and Foy, L. (2000). Battlespace digitization: SA issues for commanders. *Proceedings of the Human Performance, Situation Awareness, and Automation Conference* (p. 125). Marietta, GA: SA Technologies.

[25] Means, B., Salas, E., Randall, B. and Jacobs, T. (1993). Training decision makers for the real world. In G. Klien, J. Rasanu, R. Alderwood and C. Sambok, Eds. *Decision Making in Action: Models and Method ,* pp. 306-326. Norwood, NJ: Ablex Publishing Corp.

[26] Pleban, R.J., Eakin, D.E., Salter, M.S. and Matthews, M.D. (2001). *Training and assessment of decision-making skills in virtual environments* (Research Report 1767). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA389677).

[27] Program Executive Office Command, Control and Communications Systems (2001) *Our Leadership in Digitization.* [Online] Available: http://peoc3s1.monmouth.army.mil/battlefield_digitization.htm

[28] Randel, J.M., Pugh, H.L. and Reed, S.K. (1996). Differences in expert and novice situation awareness in naturalistic decision making. *International Journal of Human Computer Studies*, 45(5) 579-597.

[29] Sarter, N. and Woods, D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. Special Issue: Situation awareness. *Human Factors 37*(1) 5-19.

[30] Shanteau, J. (1987). Psychological characteristics of expert decision makers. In J. Mumpower, O. Renn, L. Phillips and V. Uppuluri, Eds. *Expert Judgment and Expert Systems,* pp. 289-304. Berlin, Germany: Springer-Verlag.

[31] Strater, L.D., Endsley, M.R., Pleban, R.J. and Matthews, M.D. (2001). Measures of platoon leader situation awareness in virtual decision making exercises. (Research Report 1770). Alexandria, VA: Army Research Institute.

[32] Wellens, A. (1993). Group situation awareness and distributed decision making: from military to civilian applications. In N. Castellan, Jr., Ed. *Individual and Group Decision Making,* pp. 267-291. Hillsdale, NJ: Erlbaum.

[33] Wickens, C.D. (1992). Engineering Psychology and Human Performance (2nd Ed.). New York: Harper Collins.

Team and Collective Performance Measurement

Ebb Smith
DSTL, Policy and Capability Studies
Bedford Technology Park
Thurleigh, Bedfordshire
UNITED KINGDOM

Email: mesmith@dstl.gov.uk

Jonathan Borgvall, Patrick Lif
Swedish Defence Research Agency
Department of Man-System Interaction
SE-581 11 Linkoping
SWEDEN

Email: {jonathan.borgvall, patrik.lif} @foi.se

1.0 INTRODUCTION

The measurement of operator performance has for some time formed the basis of research for those engaged in the field of human system interaction and the use of virtual reality (VR). Performance measurement is particularly relevant when the desire is develop methods and metrics to assess the utility of VR for training purposes and to predict how well that training will then transfer to the real world. Performance measurement becomes even more critical when the VR application is used in a military context, e.g., in preparation for conflict.

This chapter provides descriptions of some of the methods and measures used for measuring task and mission performance in virtual environments. As one of the challenges inherent in assessment of VR is the measurement of team and collective performance, this is the primary focus of the chapter.

2.0 TEAM PERFORMANCE

A team performance measurement system must be able to distinguish between individual and team level performance deficiencies, i.e., both *taskwork* and *teamwork* behaviours [1]. Taskwork behaviours are those performed by individual team members to execute their specific functions, e.g., weapons systems switchology. Teamwork behaviours are those which are related to team member interactions and the co-ordination of team members to achieve a common goal, e.g., communication, compensatory behaviours, information flow and feedback. For example, a team may make an incorrect decision because information was not circulated effectively among the team members (a team level problem). However, the same incorrect decision could be made because an individual made a technical error, which is an individual level problem [2].

A measurement system should assess both team *outcomes* and team *processes.* Outcomes are the end result of team performance (e.g., mission effectiveness – number of targets hit) and processes are the specific behaviours and performance strategies that explain how or why a particular outcome occurs. Sample outcome measures include accuracy of performance, timeliness of action, number of errors; sample process measures include quality of team communications, accuracy of team Situation Awareness, and adequacy of team leadership. Although successful outcomes are the ultimate goal of team training, the measurement of processes is critical for diagnosing performance problems. Feedback to trainees based on outcomes alone may be misleading and detrimental to learning. For example, teams may stumble on the correct decision or course of action despite the use of flawed processes. If feedback is outcome-based, these flawed processes will not be corrected [2].

2.1 Example of Measure of Team Outcomes – UPAS

The US Army Research Institute developed the Unit Performance Assessment System (UPAS) to help eliminate some of the limitations with the feedback capabilities of SIMNET. The UPAS has also been used in a cost-effectiveness evaluation of the Multi-service Distributed Testbed (MTD2) [3].

The system provides students and instructors with timely and useful feedback by performing all statistical analyses in real or near real-time. A UPAS collects and records data packets from SIMNET and translates and organises derived information into a relational database. This information is further manipulated onto map and graphic displays of unit performance that can be used during SIMNET after action reviews. In SIMENT the UPAS collected the following types of Protocol Data Units (PDUs): vehicle appearance, vehicle status, status change, fire, indirect fire and impact (vehicle or ground). The UPAS used five data sources to analyse unit performance in a DIS environment: network data, terrain data, units plans for the operation, radio communication and direct observation of participant behaviour.

2.2 Measures of Team SA and Shared Mental Models

There are two other very important concepts underlying team performance for which measures need to be developed: team Situation Awareness (SA) and shared mental models.

Team SA: SA is important to teams as it allows team members to be attentive to changes in the environment and anticipate the consequences of these variations [4]. A useful definition has been developed for an aviation context: Team SA has been defined as the crew's understanding of flight factors that can have an impact on the mission effectiveness and safety of the crew. Muniz et al. have identified the flight factors and have identified behavioural indicators of low and high team SA [4, 5]. Low team SA includes lack of communication, fixation, deviating from SOPs, violating limitations, using undocumented procedures, etc. Examples of high team SA include confirming information, re-checking of old information, identifying potential problems, noting deviations, having contingency plans, responding quickly to radio messages. Measures are required to evaluate the team SA of participants in a collective training exercise.

Shared Mental Models: For effective team functioning, team members need to be able to predict the needs and information expectations of other team-mates and anticipate actions. This ability is explained by hypothesising that members exercise shared or common knowledge bases, i.e., shared mental models. Shared mental models have been defined as 'Knowledge structures held by members of a team that enable them to form accurate explanations and expectations for the task, and in turn to coordinate their actions and adapt their behaviours to the demands of the task and other team members' [6].

The greater the degree of overlap in team members' models, the greater the likelihood that members will predict, adapt, and co-ordinate with one another successfully. This concept has important implications for scenarios where teams are required to co-ordinate with teams from other services and nations, where the degree of overlap may not be as great as between members from the same units. Measures are required to assess the degree of overlap between participants in a training exercise.

2.3 Example Measure of Team SA – SALIENT

Few methods for measuring team SA exist. This section examines one method known as SALIANT (Situational Awareness Linked Instances Adapted to Novel Tasks) which was developed at NAWC [4, 5].

SALIANT is an event-based approach which evaluates teams based on behaviours associated with team SA. It is similar in approach and format to TARGETs; it provides a behavioural checklist and has been found to have high inter-rater reliability. The SALIANT methodology comprises of 5 phases:

Phase 1: Delineation of behaviours theoretically linked to team SA. 21 generic behaviours have been identified from the literature and these have been clustered into 5 categories:

- Demonstrating awareness of surrounding environment;
- Recognising problems;
- Anticipating a need for action;
- Demonstrating knowledge of tasks; and
- Demonstrating awareness of important information.

Phase 2: Development of scenario events to provide opportunities to demonstrate team SA behaviours. These events were based on SME inputs and a team task analysis.

Phase 3: Identification of specific, observable responses. The behavioural indicators were transformed into observable responses based on 5 flight factors identified as crucial for attaining crew situational awareness, i.e., mission objectives, orientation in space, external support equipment status and personal capabilities.

Phase 4: Development of script. To ensure consistency across teams – when events should be introduced, what information to be provided and how to respond to teams.

Phase 5: Development of structured observation form. The form was developed to rate teams on the number of specific observable behaviours exhibited, i.e., coded whether hit or a miss.

2.4 The Role of Mental Models in Team Effectiveness

Although it has been proposed that shared mental models may hold the key for understanding and explaining team performance, there are few methods for investigating shared mental models. Where reports describe the application of certain techniques, the details of how to administer and analyse are sparse.

Based on a survey of existing research in the area of team behaviour and cognition, UK researchers funded by MoD [6] developed a generic theoretical representation of mental model functionality in command planning teams. This representation provided hypotheses for a pilot trial. They also undertook a comprehensive survey of existing data collection and assessment methods. Some of these methods were modified and new ones were developed to capture mental model data and evaluate hypotheses in a pilot study. One of these looked at the representation of mental models in command teams. The representation developed assumes that mental models can be conceived as a network of interrelated models that pass each other results of their processing. These models can be divided in 3 types:

1) Situation assessment;
2) Taskwork including models of task, equipment, time, team, individual, information; and
3) Teamwork including models of enemy plans, situation development, time, movement, combat, enemy capability, own force capability.

Each model represents a view about some aspect of the team's world. Models contain links in to other models. To complicate things further, there are also ***experiential*** and ***dynamic*** forms of mental models.

Experiential models are built up from past experiences and training; dynamic models are formed from an integration of experiential mental models and information derived from current operating environment.

It is assumed that all models must exist somewhere in team, but not all models need to be held by all members of the team. For an experienced team, the SA and teamwork models are likely to be shared to a significant extent, thus fewer requirements for taskwork models to be shared. Not all team members will have the same constructs nor represent them in same form.

To test these hypotheses, researchers developed a pre-exercise interview aimed to capture experiential teamwork mental models. A cluster analysis was conducted on the lists of rated characteristics to produce quadrant graphs for each team. These show where and how team members think similarly or differently on pertinent issues. The graphs include two variables; the level of consensus for a characteristic, i.e., number of individual who think the characteristic is related to effective teamwork; and the level of criticality for a characteristic, i.e., the degree to which they think the characteristic is critical for effective teamwork. Four quadrants were defined:

- High consensus / high criticality: most people believe characteristics critical for effective teamwork.
- High consensus / low criticality: majority consider relate to, but not critical for, effective teamwork.
- Low consensus / high criticality: one or a minority consider very important for teamwork.
- Low consensus / low criticality: one or a minority considers as not very important.

The graphs provided a profile of thinking within teams and highlight the shared perceptions and potential difference between team members.

Post exercise, teamwork analysis methods were designed to supplement the findings concerning behaviours and dynamic model utility observed and as a mechanism of exposing team's shared perceptions if teamwork. Ratings in importance and extent to which team possessing teamwork characteristics were analysed to assess the levels of disparity between team perceptions and the extent to which opinions were shared.

The results showed that teams mentioned very similar characteristics in particular the ones considered to be core to teamwork, e.g., trust and confidence, situation awareness, individual good taskwork knowledge and skills, providing and receiving performance feedback. The researchers found the method quick and simple to use and provided an effective means for analysing a team's perspective in teamwork.

3.0 COMMUNICATION ANALYSIS

Team members cooperating within or between units and teams need to coordinate their actions. This cooperation is mainly mediated by verbal and written communication, and gestures. In the network centric warfare-oriented defence, the need for communication is apparent, as is the need for communication analysis. Team communication factors have proven to be related to team performance [7]. Some areas of interest are:

- Who is communicating with whom?
- What is communicated?
- What is communicated overtly versus implicitly?
- Are the operators explicitly aware of important situation aspects?
- Which media/channels are used?

- Are there problems/errors in the communication?
- Do the problems have serious consequences on the performance?

Communication analysis can be used in addition to task analysis. It can provide information regarding changes in behavior when modifying/upgrading systems. Many of today's VR systems enable logging of both verbal transactions using push-to-talk-buttons, as well as data transfers on, e.g., enemy positions, which can provide a rich data source for post-event analysis. The communication analysis methods often involve using transcriptions of spoken communication for in-depth examination, including analysis of speech frequencies for different categories of communication, problem occurrences, and communication quality ratings. (See also Sections 4 and 5).

3.1 An Example of Communication Analysis – The IOC Model

UK MoD has funded research to develop metrics to quantify the effectiveness of training applicable to all levels and types of Armed Force training. This work has resulted in a novel approach to representing the performance of teams [8], namely the Integration Organisation and Cohesion (IOC) count analysis model. The overall objective of this research is to develop objective metrics and a methodology that will provide the MoD with a quantitative means of representing collective training in high level Operational Analysis (OA), balance of investment and cost-effectiveness models.

A descriptive model has been developed as a framework on which to base the work. This model proposes that the development of collective performance is based on improvements in integration (I), organisation (O) and cohesion (C) across the relevant set of people, i.e., the IOC model. The model is output-based and aims to assess how well a collective is working together and thus can be used to quantify the extent to which collective training has had an impact.

The IOC model breaks down the team's activities into two types: taskwork and teamwork. It is assumed that successful team outcomes rest on both good taskwork (sub-unit, e.g., formation performance) and good teamwork (processes), and that the primary purpose of team training is teaching good teamwork.

The central idea of the model is that there are three patterns of interaction within the teams:

- Actions based on response to orders;
- Actions based on the need to co-ordinate with other entities; and
- Actions based on loyalty to the team.

These can be defined in terms of three constructs:

- **Integration:** the extent to which realignment of goals arises from interventions by the collective leader. Evidence includes orders/commands coming from the leader of the collective, or information flow between the leader and the team.
- **Organisation:** the extent to which the functions of the entities are distributed and aligned to achieve the common goal. Evidence includes lateral communications used to share situational awareness, or make suggestions to each other.
- **Cohesion:** the extent to which realignment of goals arises from the entities themselves. Evidence includes reinforcing/supporting type communications.

The model hypothesises that the definition of the state of the team in terms of Integration, Organisation and Cohesion would provide an indication of how effectively the collective is likely to perform. It is assumed that appropriate scores for these attributes would lead to patterns of behaviour that support the overall goal of the team. Team training then modifies these behaviours in a manner that enhances the likelihood of achieving the team outcome.

In summary, the IOC Count Analysis technique has demonstrated utility for quantifying the value of team training. However, the technique is probably more applicable to teams within the land and naval domains, where the command structure is more hierarchical, and where communication is central to success.

4.0 DISTRIBUTED VR SYSTEMS – EVENT BASED TRAINING

Some tools have been adapted to measure teamwork in a distributed training environment. These tools were developed in the context of an instructional approach known as Event Based Training (EBT) which links learning objectives, exercise events, performance measures and After Action Review (AAR) or debrief.

Basically the EBT approach involves:

- Specification of Training Objectives (TOs): critical tasks, conditions and standards of performance.
- For each TO, the identification of specific learning objectives: these define the specific focus of exercise (we haven’t talked much about learning objectives in past). Learning objectives represent behaviours which have been deficient in the past, are subject to skill decay, or so difficult to perform need frequent practice.
- Identification of “trigger events” for each learning objective- these create opportunity for participants to demonstrate ability to perform tasks associated with learning objectives. They also provide controlled situations in which evaluators can assess performance.
- Development of performance measures used to assess task performance during each event.
- Examination of measurement data and presentation in manner to support feedback.

Dwyer et al. have been involved in the first systematic application of the EBT approach in a distributed training environment [9]. This was used to develop performance measures, namely the TARGET checklist the TOM instrument. These are outlined below, together with a description of how they were used in two case studies.

4.1 The TARGET Checklist

TARGET stands for Targeted Acceptable Responses to Generated Events or Tasks [10]. The method is event-based and involves the identification of events for a training session which serve as triggers for team members to exhibit examples team behaviours.

In addition, for each of these events, acceptable responses (i.e., the TARGETs) are identified in advance of the exercise. Anticipated behaviours are based on training manuals, SOPS, doctrine and SME inputs. Behaviours are then arranged into a checklist in the approximate order they will occur. As the exercise unfolds, observers score each item as acceptable, unacceptable or unobserved. An implicit assumption in the TARGETs methodology is that behaviours are observable and the instructor can determine them as being present, i.e., a “HIT” or absent, i.e., a “MISS”.

Performance can be assessed in number of ways: the proportion of behaviours correctly performed relative to total set of behaviours can be calculated or behaviours can be grouped into functionally related clusters, which can then be examined to see how well team performed in functional areas.

4.2 The Teamwork Observation Measure

Teamwork Observation Measure (TOM) was derived from performance measurement techniques developed under the US Navy’s tactical decision making under stress [5], and aircrew co-ordination training research. The purpose of TOM is to identify performance strengths and weaknesses and to obtain performance ratings on critical dimensions of teamwork.

TOM includes 4 dimensions of teamwork: communication, team co-ordination, situational awareness and team adaptability. Each dimension is then divided into key factors (see Table 7-1).

Table 7-1: TOM Dimensions and Factors

TOM Dimension	Factors
Communication	Correct format
	Proper terminology
	Clarity
	Acknowledgements
Team Co-ordination	Synchronisation
	Timely passing of information
	Familiarity with other's jobs
Situational Awareness	Maintenance of big picture
	Identify potential problem areas
	Remain aware of resources available
	Provide information in advance
Team Adaptability	Back-up plans
	Smooth transition to back-up plans
	Quick adjustment to situational changes

Assessors are required to provide specific comments based on observations made to be highlighted as critical points during feedback. Assessors also provide ratings of how well participants interacted with each other on each of the four teamwork dimensions.

5.0 COLLECTIVE PERFORMANCE ASSESSMENT

The need to assess and measure performance at a collective[1] level presents researchers with a number of challenges. A collective operates at a higher level than a team and involves different roles co-ordinating their activities, without necessarily being co-located and without necessarily having a single recognised leader or identical goals. Certain skills that are important for teams, e.g., communication, co-ordination and information sharing are also key to collective success. However, in a collective there is less likelihood of shared expectations derived from previous experience and reduced area of overlap in shared mental models compared to an established team [11].

To use an example from the air domain, Collective air mission training may involve many aircraft, fulfilling different roles, some directly involved in a mission and some providing support. For example, a 4-ship in an Air to Ground role needing to co-ordinate with Air-to-Air assets, Suppression of Enemy Air Defence (SEAD) assets and Airborne Warning and Control System (AWACS) aircraft. It is the inter-team

---

[1] 'Collective mission training' is defined as two or more teams training to interoperate in an environment defined by a common set of collective mission training objectives, where each team fulfils a different military role. NATO SAS-013 Study.

rather than intra-team interactions and co-ordination that are important. High level cognitive skills, such as the ability to build and maintain situation awareness or to make tactical decisions in a complex and highly dynamic environment are crucial.

5.1 Implications for Collective Training Assessment Techniques

An understanding of the benefits gained from current collective air training gives an insight into what needs to be captured by training assessment and performance measurement techniques. There is a need for techniques that do not simply capture mission outcomes, but more importantly the underlying cognitive processes and strategies. To truly quantify the training value of collective air training, there is a need to capture some of the less tangible benefits for example positive changes in aircrew's understanding, situational awareness, flexibility and confidence.

Any techniques identified should ideally be of utility in the live environment. For example, whilst observers are able to make valid, albeit subjective judgements of performance and use these to give feedback and guidance to participants, live collective exercises could benefit from a more formal approach. In addition, if techniques could be applied to both live and VR exercises this would enable comparisons of the relative value of both training environments to be made.

5.2 Collective Performance Assessment and Mission Phases

Within the UK, under the sponsorship of the MoD, a programme of applied research has been undertaken to explore the benefits to be gained from using networks of simulators within a VR environment for aircrew collective mission training. Use of networked simulation in this context (in the UK) has become known as UK Mission Training through Distributed Simulation (MTDS) [12]. The approach adopted by UK MTDS researchers advocates a subjective assessment of performance during all phases of the training event [13]. Typically these phases are plan, brief, mission execution and After Action Review (AAR) or debrief. This work has led to the devolvement of tool designed specifically to assess collective performance during all mission phases; the Collective Assessment performance Tool (C-PAT) [15].

C-PAT is being developed by the Defence Science and Technology Laboratory (Dstl), part of the UK MoD. It forms part of an evolving concept of analysis for the UK Mission Training through Distributed Situation (MTDS) initiative and has already demonstrated great utility in providing measures of effectiveness for synthetic collective training. Essentially C-PAT is a 'family' of surveys (listed below in Table 7-2.) designed specifically to facilitate Subject Matter Expert (SME) assessment of collective performance of aircrew throughout all mission phases. Typically these SMEs also undertake the White Force role during both live and virtual collective training events.

Table 7-2: The C-PAT Familiy of Surveys

C-PAT Survey Element	Description
Planning Phase Assessment	WF evaluation of the 'quality' of co-ordination during the planning process on each mission day is an important component of this assessment. This is something that the WF are well used to judging during live collective training exercises. At the end of the planning phase of each mission the WF team were asked to complete a planning assessment questionnaire, giving their expert judgement in areas such as leadership, use of information, time management, thinking about the 'Big Picture', decision making.
Mass Brief Assessment	WF will evaluate the 'quality' of the briefs in terms of clarity, accuracy, big picture information, etc. This survey is still under development.
Assessment Criteria	At the end of each mission the WF are asked to complete an "Assessment Criteria" questionnaire, which asked for assessments on 31 criteria to form a picture of how well a collective exercise is proceeding. Typical criteria are: How effective were the tactics employed during the mission? How appropriate was any review of tactics made as a result of lessons learned? To what extent were the overall objectives of the mission achieved? Were relevant lessons learned and actions thoroughly debriefed?
Mass Debrief Assessment	WF will evaluate the 'quality' of the debrief in terms of clarity, accuracy, and lessons identified. This survey is still under development.
Training Objectives	Participants will be asked to rate to what level the training objectives were supported during the training event. These comprise a number of sub-elements, all of which are given a rating. Scores will then be consolidated to give an overall rating for each of the TOs.
Interoperability	Trust is a vital component of interoperability. One of the benefits of collocation is that it appears to help engender trust in away that may not be possible with distributed players. This survey is still under development.

The C-PAT has been developed on the premise that effective collective processes can really only be assessed by an SME with the appropriate level of domain specific knowledge. The thought processes used in making these judgements are often difficult to articulate and considerable effort has been expended in trying to elicit these from tactical/training experts from the Air Warfare Centre (AWC). The tools are continually being refined with inputs from the AWC, and it is hoped that their involvement in the design of C-PAT, will ensure that these are formulated and worded in a manner that will be understood by end users. The ultimate aim is to develop robust metrics that can be utilised to measure the effectiveness of both live and synthetic collective air training exercises, thus enabling the value of UK MTDS training exercises to be quantified.

At the centre of the C-PAC, are collective performance indicators; these have been derived from benefits identified by participants in live collective exercises. Typical collective performance indicators that have been used to assess the utility of a virtual environment to support mission training are presented in Table 7-3.

Table 7-3: List of Typical Collective Performance Indicators

No.	Collective Performance Competency/Indicator
1	Understanding of own team's role and capabilities
2	Understanding of other team's role and capabilities
3	Understanding of where own team fits into the 'bigger picture'
4	Ability to balance risks – exploring the 'what ifs' of the training scenarios
5	Ability to cope with the 'fog of war'
6	Awareness of the tactical situation (multi-level SA)
7	Within role communication and co-ordination skills
8	Between role communication and co-ordination skills
9	Tactical skills
10	Tactics development
11	Utilisation of role specific skills within the collective environment
12	Ability to understand and implement briefed operational procedures
13	Effectiveness in Commander role
14	Decision making
15	Fluidity in a variety of dynamic situations
16	Confidence in own capabilities
17	Confidence in own team's capabilities
18	Confidence in other teams' capabilities

The C-PAT is still evolving. One area requiring further investigation is measurement of aircrew Situation Awareness particularly their awareness of other team member's roles and intentions. Good Situational Awareness is integral to an effective mission execution phase, but it is difficult to quantify. With regard to the surveys themselves, feedback indicates that aircrew may find it difficult to equate their established rating scales with the required percentage responses. The use of anchored rating scales is to be investigated. However, this is not necessarily a simple solution, as ease of use does not necessarily equate to more meaningful data. Recently a mapping exercise was undertaken between assessment criteria and collective training competencies. Understanding these relationships will further help with quantifying training effectiveness.

Data collection and analysis can be time-consuming when carried out manually. One of the future aspirations for the technique is to provide a rapid and reliable measure of effectiveness of UK MTDS training events. To this end, there are plans to administer surveys in an electronic format. This should also permit the automatic data collection of responses in quantifiable terms.

6.0 OBJECTIVE MEASURES

Objective measures are less debatable than subjective measures, but can lack in contextual value. Objective measures can serve as a basis for comparison with subjective measures to reflect whether attitudes reflect what actually happened during the mission. It is important to develop a robust set of objective measures for a more rigorous assessment to performance and to maximise the benefits of the AAR/debrief session. Some form of data logger is thus an important component of overall the VR system.

The logger should log all data that is generated within an exercise or event. For example within a networked VR training event, data is typically output onto the network in the form of DIS Protocol Data Units – (PDUs) that are generated. All PDUs are time stamped with their time of reception at the logger. The logs provided by the logger can then be replayed during debrief at normal speed, slower then normal speed or faster than normal speed to enable the instructor, exercise director or trainee to fast-forward and pause at a critical mission incident and engage the training audience in further discussion and capture lessons learnt.

Information captured on the data logger will also provide valuable insight as to the health of the system and the integrity of the technical and tactical networks. More importantly it will provide measures of individual, team and ultimately collective performance which can be used to aid debriefing and performance assessment on a number of different axes.

In order to be able to make such assessments it may be necessary to have a baseline against which actual performance during the training event could be measured. For example, the air defenders performance in Weapons Engagement Zone management and control and how they ‘pushed the bubble’ could be assessed by comparing it with the baseline parameters; speed, height, sensor information, tactical manoeuvres, etc. Objective assessment is a key to a successful AAR and debrief. Significant progress has been made in this area in recent years and a number of bespoke solutions developed to capture objective data necessary to support a more robust evaluation of performance in a virtual training environment. An example is the work undertaken by the Air Force Research Laboratory in Mesa, US as part of their research into Distributed Mission Training (DMT). AFRL has developed a software tool known as PETS (Performance Evaluation Tracking System). [15]. PETS is capable of capturing the objective data necessary to support a robust and real-time evaluation of performance in a DMT training event. Data is organised at several levels to aid assessment. They include RT graphical displays, performance effectiveness learning curves, and statistical analysis at scenario or shot level.

7.0 INDIVIDUAL PERFORMANCE

Whist the chapter has focused on team and collective performance measures, for completeness some examples of individual performance measures are also included. The measures discussed in this section have been developed by the Swedish Defense Research Agency (FOI) and focus primarily on pilot performance and include both subjective and objective assessment techniques.

7.1 FOI Approach to Performance Measurement

FOI has a long tradition of measuring operative performance. Though varying regarding the specific measures, the general approach has always been the combination of subjective measures (e.g., questionnaires, rating scales), objective measures (e.g., data logging), and psycho-physiological measures (e.g., HRV, EPOG). Since the ambition is to use measures that reflect the dynamics of the situation, attempts to reduce the wide range of variables are necessary. The tradition is to use factor analysis for identification of significant compounded indicators. Linear causal model analyses are then performed by means of structural equation modelling (SEM), for example LISREL [16], to test the validity of different causal flow models possible.

The method of assessing performance that is most commonly used by FOI is a modified version of the Bedford Rating Scale [17]. The pilots answer questions using a 10-point scale. The modified scale can be formulated in either first person or third person. It can also be used pseudo-dynamically, that is, that the scale is being used repeatedly, after important aspects, throughout a mission. The measure has been used in several studies [18, 19].

There are sometimes a difference between pilot ratings and instructor ratings. These differences can be explained by different understanding of what constitutes performance. The ratings has shown correlations with Mental Workload (r = -0.55), Situational Awareness(r = 0.52), and Heart Rate (r = – 0.59) [20].

7.2 The FOI Pilot Performance Scale

The FOI Pilot Performance Scale (FOI PPS) is useful for addressing aspects of difficulty, performance, mental capacity, mental effort, information load, situational awareness, and mental workload. The six dimensions are extracted by means of factor analysis and the number of markers range from 3 to 7. The reliability of the dimensions or indices has been tested by means of Cronbach's alpha and they have been cross-validated. The Swedish questionnaire has not been validated in English. The questions are developed to fit in military fixed wing scenarios and relate to flown missions with specific as well as general questions. Relationships between the indices have been analyzed by means of structural equation modeling [21, 22]. Subjects answer by scoring on a 7-point bipolar scale. This measure has been used in several studies and the reliability ranges from 0.73 to 0.90. Indices change significantly as a function of mission complexity.

The FOI PPS significantly relates to psycho-physiological indices such as heart rate and eye point of gaze changes and it correlates 0.79 with mission/task difficulty level, 0.84 with the NASA-TLX and 0.69 with the Bedford Rating Scale. FOI PPS has mainly been used in training simulators and after missions in real aircraft. FOI PPS is not available in English. Examples of (translated) questions are:

- How complex did you find the mission?
- Did you feel forced to disregard or cancel some of your tasks in order to perform optimally on critical tasks?
- To what extent did you feel disturbed by non-critical information?
- Did you have problems monitoring the information on the Tactical Situation Display (TSD)?

The instrument has 6 dimensions: Operative Performance (r =0.74), Situational Awareness (r =0.80), Pilot Mental Workload (r =0.87), Mental Capacity (r =0.77), Information Handling Tactical Situation Display (r =0.92), and Information Handling Tactical Information Display (TI) (r =0.93).

It takes about 5 minutes to answer the questionnaire. Some subjects find the questionnaire too long and time-consuming. The indices are suitable to use in causal analyses [16].

7.3 Objective Measures of Individual Performance

Task performance measures vary between research groups and research areas. Speed and accuracy are used by most research teams in one way or another. The choice of measure is of course dependent on research area (e.g., visual and audio perception), but also on possibilities in the actual situation. Both controlled laboratory experiments and field studies are of interest, and often complement each other in seeking for new solutions. Below follows some examples of dependent measures of performance commonly used.

Angle Estimation (visual perception): Comparisons between targets with or without monocular depth cues (drop-lines) can be used for evaluating different display settings. Subjects perform angle estimations in a 3D virtual environment where the task is to detect a threat and to estimate the angle of a prioritized target in 3D space [23]. Answer is given by pointing a virtual arrow in the estimated direction, from ownship in direction to target. Both azimuth and elevation are measured and analyzed separately, but they can also be analyzed together. Comparisons between angles are also possible, even though the main interest have been to compare with and without additional monocular depth cues.

- Angle Estimation (audio): Subjects perform angle estimations from an audio signal (speech and noise) with the intent of comparing two 3D-audio systems spatial resolution. One system is an expensive professional hardware solution and the other an inexpensive software application, further developed at FOI.
- Relative Height Estimation: Relative height estimation can be used in a flight situation when evaluating the effect of using monocular depth cues (drop-lines and cone attached to the ground or to a fixed plane). The subject's task is to estimate which target symbol is closest or most distant compared to own ship [24]. One important point using relative estimation is that the measure is non-metric (compared to angle estimation), which can be of importance when using dependent measures in a three-dimensional virtual setting. According to some researchers [25], a 3D virtual environment will create different errors in x, y and z-axis. This kind of problem speaks for the use of non-metric measures in a 3D setting.
- Future Collision Point: In the flight domain future collision points or risk of collisions [26] are of great importance. The subject's task is to select which of a number of the targets has a collision course with the own aircraft. Both speed (Response Time) and accuracy can be measured.
- Deviation from Flight Path: Can be used as a performance measure when flying, e.g., 'tunnel in the sky' or as a secondary dependent measure when performing another task.
- Relative Size Estimation: Can be used when comparing settings were monocular and stereoscopic vision is in focus, including different techniques for stereo presentation and other VE techniques like tactile displays.
- Color Discrimination – Staircase Method to find Just Noticeable Differences (JND): Color perception can be affected during different g-loads during high performance flight. One method is a staircase method with different colors. The baseline for JND at some well known colors is known [27] and can be compared with the JND values acquired in a centrifuge setting.
- Color Identification: Pilots performed identification of well known colors during g-load.
- Symbol Identification: Aircraft vibrates at different amplitudes and frequencies that might cause problems reading text or understanding symbols. To understand the effect of frequencies and amplitudes, experiments were conducted with vibrating symbols at different frequencies [28]. Symbol identification can also be used to evaluate if g-load affects identification during modest or high g-load.
- Balance Measures: Investigation of visual flow effectiveness includes studies of display effects of visual vertical variation on observer balance. Thereby the impact on perceived spatial orientation was studied, with greater postural sway linked to increased proprioceptive and vestibular suppression. Thus, greater postural sway reflects increased display effectiveness [29].
- Reaction Time: Often used as a performance measure in combination with measures of accuracy.
- Alarm Sound Categorization: Subjects performed categorizations of different alarm sounds together with estimated vigilance, duration, and audibility.

8.0 SUMMARY

This chapter has focused primarily on team and collective performance measures. Whilst it has provided some examples of team and collective measures, it should be noted there exists a rich source of information in literature concerned with performance measurement and human factors. Two recent publications which are recommend to readers interested in this topic are; Performance Measurement – Current Perspectives and Future Challenges', edited by Winston Bennett, Charles Lance and David Woehr, published by LEA, 2006 and Human Factors Methods – A Practical Guide for Engineering and Design, by the Human Factors Integration Defence Technology Centre, published by Ashgate, 2005.

9.0 REFERENCES

[1] Naval Air Warfare Center, ‘Performance measurement in teams’, Training Series No. 2.

[2] Smith-Jentsch, K.A., Johnston, J.H. and Payne, S.C. ‘Measuring team-related expertise in complex environments’. In J.A. Cannon-Bowers and E. Salas (Eds.). ‘Making decisions under stress: implications for individual and team training’. 1998. APA Press.

[3] Taylor, H.L., Orlansky, J., Levine, D.B. and Honig, J.G. ‘Evaluation of the performance and cost-effectiveness of the Multi-Service Distributed Training Testbed (MDT2)’. Proceedings of Royal Aeronautical Society Conference, November 96.

[4] Muniz, E., Bowers, C.A., Stout, R.J. and Salas, E. ‘The validation of a team situational measure’. Proceedings of the 3rd Annual Symposium and Exhibition on Situational Awareness in the Tactical Environment, Patuxent River, MD, 62, 1998.

[5] Muniz, E.J., Stout, R.J., Bowers, C.A. and Salas, E. ‘A methodology for measuring team Situational Awareness: Situational Awareness Linked Indicators Adapted to Novel Tasks (SALIANT)’. Proceedings of Symposium on “Collaborative Crew Performance in Complex Operational Systems”, UK, 20-22 April 1998.

[6] Pascual, R.G., Henderson, S.M. and Mills, M.C. ‘Understanding and supporting team cognition-Year 1 report and recommendations’, DERA/CHS/MID/CR980122/1.0, June 1998.

[7] Sexton, J.B. and Helmreich, R.L. ‘Analyzing cockpit communication: The links between language, performance, error and workload’. Paper presented at the Proceedings of the 10th International Symposium on Aviation Psychology, Columbus, OH. 1999.

[8] Belyavin, A. ‘Quantify the effectiveness of collective training: Final Technical Summary Phase I’. QINETIQ/KI/CHS/CR033077. December 2003.

[9] Dwyer, D.J., Randall, L.O., Salas, E. and Fowlkes, J.E. ‘Performance measurement in distributed environments: initial results and implications for training’.

[10] Dwyer, D.J., Oser, R.L., Fowlkes, J.E. and Lane, N.E. ‘Team performance measurement in distributed environments: the TARGETs methodology’ In M.T. Brannick, E. Salas and C. Prince (Eds.). Team performance assessment and measurement: theory, methods and applications. 1997.

[11] McIntyre, H.M. and Smith, E. ‘Collective operational aircrew training research: Progress report FY99 and FY00’. QinetiQ/FST/CSS/TR011124/1.0, November 2001.

[12] Smith, E. ‘Final Report on UK MTDS Study Part 2 – Detailed Full Report’. Dstl/CR07001V1. September 2003.

[13] Smith E. and McIntyre, H. ‘Simulation and collective training – creating a total experience’. The Royal Aeronautical Society, Flight Simulation DERA/AS/FMC/CP010229. June 2000.

[14] Smith, E. ‘Quantitative Techniques to Assess the Value of Training Output’. Dstl/CR17059 v1.0. September 2005.

[15] Gehr, S.E., Schreiber, B. and Bennett, W. ‘Within-Simulator Training Effectiveness Evaluation’. Proceedings of Interservice/Industry Training, Simulation and Education Conference (I/ITSEC) 2004.

[16] Jöreskog, K.G. and Sörbom, D. ‘LISREL8: Structural equation modeling with the SIMPLIS command language’. Hillsdale: Lawrence Erlbaum Associates. 1993.

[17] Berggren, P. (2000). ‘Situational awareness, mental workload, and pilot performance – relationships and conceptual aspects’. Linköping: Human Sciences Division. FOA-R-00-01438-706-SE.

[18] Svensson, E., Angelborg-Thanderz, M., Sjöberg, L. and Olsson, S. (1997). ‘Information complexity-mental workload and performance in combat aircraft’. Ergonomics, 40, No. 3, pp. 362-380.

[19] Borgvall, J., Nählinder, S. and Andersson, J. (2002). WVR-Illustrator Evaluation: ‘Using Pilot Expertise for Future Development’ (Scientific No. FOI-R--0710--SE). Linköping, SE: The Department for Man-System Interaction at the Swedish Defense Research Agency.

[20] Magnusson, S. and Berggren, P. (2001). ‘Measuring pilot mental status’, NAMA conference, Stockholm.

[21] Svensson, E., Angelborg-Thanderz, M. and Wilson, G.F. (1999). ‘Models of pilot performance for systems and mission evaluation – psychological and psycho-physiological aspects’. AFRL-HE-WP-TR-1999-0215.

[22] Svensson, E. and Wilson, G.F. (2002) ‘Psychological and psycho-physiological models of pilot performance for systems development and mission evaluation’. International Journal of Aviation Psychology, Vol. 12 (1).

[23] Andersson, P. and Alm, T. (2003). ‘Perception Aspects on Perspective Aircraft Displays’. Displays, 24, 1-13.

[24] Lif, P. and Alm, T. (2004). ‘Relative Height in 3D Aircraft Displays’. Human Performance, Situation Awareness and Automation Technology Conference.

[25] Smallman, H.S., John, M.S. and Cowen, M.B. (2002). ‘Use and misuse of linear perspective in the perceptual reconstruction of 3-D perspective view displays’. Paper presented at the Human Factors and Ergonomics 46th Annual Meeting, Baltimore, Maryland, US.

[26] Endsley, M.R. (1995). ‘Toward a Theory of Situation Awareness in Dynamic Systems’. Human Factors, 37(1), 32-64.

[27] MacAdam, D.L. (1942). ‘Visual sensitivities to color differences in daylight’. Journal of the Optical Society of America, 32(5), 247-274.

[28] Andersson, P. and Hofsten, C.V. (1999). ‘Readability of vertically vibrating aircraft Displays’. Displays, 20, 23-30.

[29] Eriksson, L. and von Hofsten, C. (2004). ‘Effects of visual flow display of flight maneuvers on perceived spatial orientation’. Human Factors. (Accepted manuscript.)

After Action Review in Simulation-Based Training

Larry L. Meliza, Stephen L. Goldberg and Donald R. Lampton
U.S. Army Research Institute for the Behavioral and Social Sciences
Orlando, Florida
USA

The after action review (AAR) is a mechanism for providing feedback to organizations on their performance of collective tasks. It is an active process that requires unit members to participate in order to benefit. The AAR is a method of providing feedback to units after operational missions or collective training exercises [23]. The AAR is an interactive discussion, guided by a facilitator or trainer known as an AAR leader. During the AAR, unit members discuss what happened, why it happened, and how to improve or sustain performance in similar situations in the future.

1.0 SIMULATION-BASED TRAINING ENVIRONMENTS

Military Training is always a simulation whether it occurs at a sophisticated instrumented range, in a collective training simulator system, or in a command and staff exercise driven by a math model driven war game. Training occurs in live, virtual, constructive, or mixed simulations of battlefield environments. There are always compromises in training with how tasks would be performed in combat. In the live environment, units use operational equipment and actual terrain and perform against an opposition force composed of military personnel (live force-on-force) or targets (live fire), depending upon whether the unit is employing simulated weapons' effects or firing live rounds. In virtual environments, units use simulators to represent equipment and weapons. Weapons effects, terrain and enemy forces are computer generated. In constructive environments, battlefield outcomes (e.g., the unit lost thirty percent of its personnel) are determined by sophisticated math models in order to provide battle effects supporting command and staff training. Mixed environments include elements of two or more of the simulation environments. Training in all of these simulation environments should provide individuals and units with feedback about how their actions contributed to mission success or failure, casualties received, and casualties inflicted on the enemy, the bottom lines of collective performance.

Live simulation training is widely available, in the form of ranges and maneuver areas. The most highly supported form of live simulation training is generally found at ranges that support engagement simulation and vehicle or aviation asset tracking. Examples of these ranges are the US Army's National Training Center or the US Navy's Top Gun program. They differ from local area ranges in that they provide a cadre of observer/controller/trainers, a dedicated opposing force, instrumentation that is capable of collecting position location, firing and status data, and teams of analysts supporting observer/controller/trainers from a data collection and analysis facility.

Virtual Simulation training systems vary widely across services and nations. Virtual training environments can range from driver and gunnery trainers to sophisticated networks of simulators representing aviation assets operating in a coalition format. The introduction of game-based training has expanded the scope of virtual training and made it more widely available because of the games relatively low cost.

Constructive Simulations fall into two categories. Math model based attrition models represent the actions of military units and serve primarily as drivers for training exercises of command groups and staffs.

The constructive simulation provides the inputs to commanders and staff's decision making processes. A second more recent form of constructive simulation is more entity and rule-based. These simulations provide semi-automated forces to be enemy and adjacent friendly forces for virtual simulations. The Close Combat Tactical Trainer (CCTT), a U.S. Army heavy forces virtual trainer, utilizes CCTT SAF to generate enemy forces for the training audience to fight.

2.0 INTRINSIC FEEDBACK

Intrinsic feedback cues and guides unit behavior during task performance [4], whether the task is being performed in an operational or training context.. For example, an infantry unit may call in artillery fire on a target and receive intrinsic feedback when it observes that simulated or actual rounds impact too far from the intended target. Someone from the unit assigned to observe the effects of the artillery fires would then provide the supporting artillery unit with guidance for adjusting their fires. A portion of intrinsic feedback comes from simulations or actual weapons effects (the location of artillery impact) and part comes from unit actions (an observer providing the artillery unit with directions for adjusting fires).

As a rule of thumb, there are more gaps in intrinsic feedback for a unit that is not well trained, because unit members are not providing their portion of the feedback. Continuing with the artillery example, not having an observer in position can result in a gap in terms of a unit's knowledge of the effectiveness of its artillery. From a broader perspective, if there are failures to communicate information up and down the chain-of-command during an exercise , then there will be gaps in feedback needed to cue and guide performance. If unit members are not sure about what aspects of the tactical situations they should be monitoring, then additional gaps in feedback are to be expected.

An important difference between individual and collective performance is that in collective performance much of the information needed to cue and guide performance comes from other people. To be fully trained unit members must learn how to provide this information (i.e., their part of the intrinsic feedback). Improved capability to provide intrinsic feedback at the right time to the right people is evidence that unit performance is improving.

3.0 EXTRINSIC FEEDBACK

For a unit to improve its performance unit members will in most cases need more feedback about what happened during an exercise than that gained by participating in it and observing what happened. Because of the so called "fog of war", when an exercise is over, participants sometimes have a limited perspective regarding what happened, based upon the information available to them and what they saw, heard and smelled. This limited perspective is referred to as perceived truth. Ground truth is the term used to describe the actual events that occurred. Less trained units are expected to demonstrate a greater disparity between perceived and ground truth, simply because much of the intrinsic information that was available was not either perceived or used. Events may be happening quickly and open to differing interpretations. Perceptions and memories of the occurrence, sequence, and timing of events can be greatly distorted leading to generation of causal relationships which are not based on the actual facts [8].

Extrinsic feedback is provided by an outside source, usually observer/controllers or trainers after an exercise ends. It is designed to help participants understand the ground truth situation relative to their perception of perceived truth and to investigate what caused the events to occur as they did. Extrinsic feedback consists of information that the exercise participants don't ordinarily have available to them. It can provide insights into how to improve or sustain performance in the future.

A simulation is effective to the extent that exercise participants can appropriately recognize intrinsic feedback regarding their performance, and extrinsic is provided to clarify misperceptions. Extrinsic feedback, by providing information on exercise outcomes, allows the actions of individuals to be linked to higher level exercise outcomes. Sometimes exercise participants recognize the impacts of their actions via intrinsic feedback, but at other times they are not aware of these impacts until they receive extrinsic feedback.

4.0 EXTRINSIC FEEDBACK METHODS FOR COLLECTIVE TRAINING

Formal post-exercise feedback sessions are one of the types of extrinsic feedback that can be used to improve unit performance [4]. The After Action Review (AAR) is a method of providing extrinsic feedback to units after operational missions or collective training exercises [23]. Simply put, the AAR is the controlled sharing of intrinsic feedback combined with group problem solving. Exercise participants play differing roles and are located at differing points within the battlespace, so each participant receives relatively unique intrinsic feedback. Extrinsic feedback can be used to correct misperceptions and clarify events and effects. The AAR process may provide unit members with a view of collective (team, unit, or organizational) performance that was not apparent to, or viewable by, any one participant during an exercise [14], including trainers who were observing the exercise. The AAR uses a Socratic Method in which a series of leading and open-ended questions are used by an AAR leader to help those in the training audience discover what happened and why.

A debrief or critique conducted by one or more observers of a training exercise is an alternative to the AAR [19, 9] and is a more traditional way of providing feedback by trainers. The person or persons who provide the critique become the source of ground truth as they see it. Their role is to interpret events as they saw them and describe to the training participants what they think happened, why they think it happened, and what they think the unit should do about it. Critiques are an extrinsic source of feedback. A major difference between the AAR and critique is that the critique provides the training participants with conclusions reached by the person giving the critique rather than facilitating the training participants to reach their own conclusions. Critiques can easily be taken as criticism since the opinions expressed are based on perceptions, judgments and possibly misinterpretations of ground truth. Further, the critique is unable to make use of diagnostic information that may be known only to exercise participants.

The AAR leader functions as a discussion facilitator. Training participants are expected to examine their performance through guided self evaluation. They are encouraged to identify their problems and develop approaches to correct them. It is assumed that use of the AAR feedback method results in units claiming ownership of the diagnosis of problems and the corrective actions they identify [19].

Extrinsic feedback regarding unit performance focuses on conceptual knowledge rather than procedural knowledge. Feedback is likely to be more explanatory than directive in nature. The whole process of using interactive discussions to decide what happened, why it happened, and how to improve or sustain performance engenders explanations. Explanatory feedback is superior to directive feedback in terms of conceptual knowledge acquisition [16].

Post-exercise feedback, by definition, is delayed rather than immediate. It could, conceivably be used in conjunction with immediate feedback in the course of an exercise (i.e., through coaching, mentoring, intelligent tutoring, and/or the application of during action review aids) so that a unit can take immediate corrective action and perhaps accelerate the training process [11]. In the case of collective training, corrective actions in mid exercise may help prevent a unit or team from creating a tactical situation that detracts from the intended training objectives of the exercise.

5.0 HOW REALISTIC BATTLEFIELD SIMULATIONS SET THE STAGE FOR AN AAR

The AAR was based upon the "interview after combat" used in World War II by military historian Samuel Lyman Atwood (S.L.A.) Marshall and others to find out what happened during missions [3]. The process was adapted for training events as the capability to provide realistic simulation of weapons effects occurred during the 1970s and 80s [3, 17]. Prior to development of engagement simulation technologies in the 1970s most military collective training exercises casualty exchanges and mission outcomes were based upon the subjective judgments of umpires. Such judgments were insufficient to prepare participants for an "interview after combat," because the participants didn't believe that their status as casualties necessarily resulted from their behavior. The development of tactical engagement simulation technologies provided a means for objective casualty determination [22]. Perhaps the best known example of TES is the use of lasers and laser detectors to simulate the effects of line-of-sight weapons, such as rifles and tank main guns. The later development of virtual simulations such as Simulation Networking (SIMNET) and The Close Combat Tactical Trainer (CCTT) eliminated many of the inaccuracies of live TES casualty assessment and were capable of more fully representing ground truth [7].

6.0 THE ROLES OF AAR AIDS IN SUPPORTING FEEDBACK

To be effective, AAR discussions need to be guided by an AAR leader. The leader needs one or more start points for the discussion and at least a general idea of where the discussion will head. The job of the AAR leader is made easier to the extent that they are already aware of the types of problems the unit has been experiencing. If all an AAR leader knows about a mission is that a unit sustained heavy casualties, the Socratic Method will take a long time to identify the root causes of the problem. If the AAR leader and the unit know that most of the casualties occurred within a few minutes of making contact with the enemy and that few friendly vehicles returned fire upon contact, they are closer to identifying and understanding what happened and why. The AAR does not require an exhaustive review of all aspects of a unit's performance. Instead, trainers are expected to focus on aspects of performance closely linked to key exercise events and outcomes.

At instrumented ranges and in virtual simulations AAR aids prepared from electronic data streams can document or illustrate aspects of performance that are close to root causes of weaknesses and strengths. Developments in battlefield simulations technology have provided trainers with a record of electronic data describing position location, firing events and communications over the course of an exercise. AAR software systems have been developed that allow this data to be converted into a variety of AAR aids describing or illustrating ground truth [14]. For example, a graph showing the number of rounds fired by each vehicle in a platoon over time may make the point that only one of the vehicles in the platoon fire was involved in the early portion of an engagement. To gain this information from the AAR process, a unit would have to slowly reconstruct the sequence of events based on their memories. AAR aids also offer the benefit of providing units with demonstrable ground truth when their recollections are at odds with what actually happened.

To the extent that AAR aids illustrate root causes of exercise events, rather than outcomes, they expedite the AAR process. AAR aid generation capabilities that examine exercise data streams to check specific aspects of performance offer a means of helping AAR leaders and units diagnose strengths and weaknesses.

The most frequently used AAR aid is a sequential replay of exercise events. A replay, however, is not necessarily the most efficient or effective way of illustrating key aspects of performance. Sometimes AAR aids that summarize activity over a period of time can be more effective. A graphic tracing the movement of a unit may be able to quickly illustrate that a unit backtracked, indicating that the route reconnaissance may have been inadequate.

Efforts to develop innovative AAR aids have not always been successful [21]. Efficient aids should provide information that would otherwise have to be gleaned from lengthy reviews of replays. However, poorly conceived aids tend to confuse rather than clarify the situation because they do not present information in a manner that is intuitive or clear to the training participants. The ideal situation would be one in which training participants have learned over time to expect certain AAR aids to be presented after an exercise and they know what information is to be gained from each type of aid. It has been difficult to reach this goal because each training/simulation environment has different capabilities to support AAR aid production. Live simulations are limited by the quality of the data generated by the instrumentation and engagement simulation systems. Constructive simulation often will not represent entity level performance, but rather aggregate performance of units. As mentioned above, virtual simulations represent the best opportunity to provide accurate and appropriate AAR aids because most of the data of interest can be easily collected from a network.

AAR aids provide units with an improved perspective regarding what actually happened during an exercise that more accurately reflects ground truth.. An important goal of the unit is to identify corrective actions it can take that will provide unit members with an improved perspective as it is conducting training and operations in the future (i.e., better intrinsic feedback to cue and guide unit behavior).

7.0 A RECENTLY DEVELOPED AAR TOOL FOR VIRTUAL ENVIRONMENTS

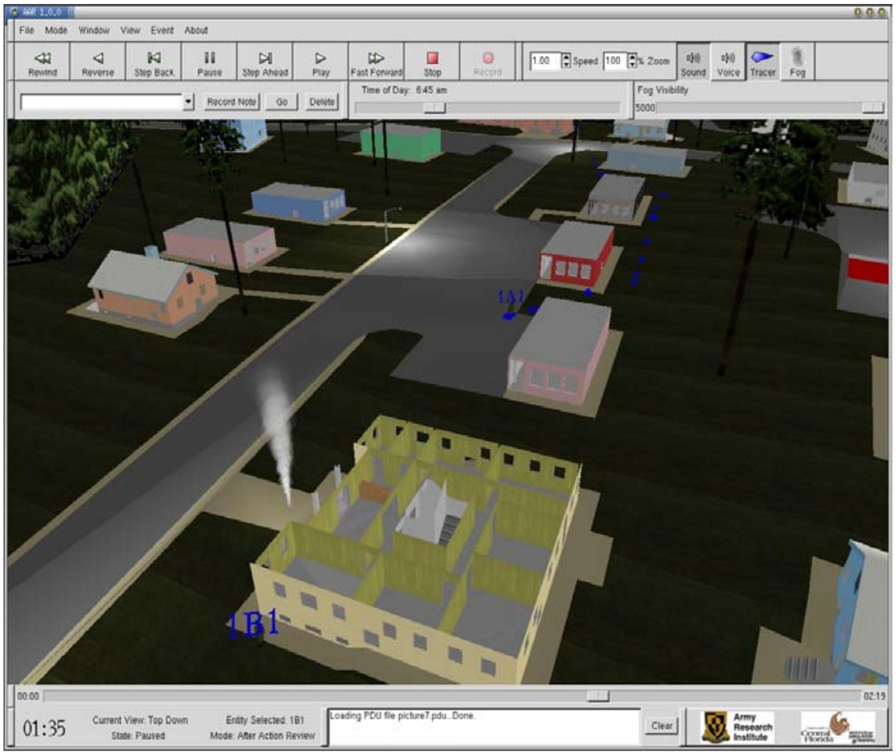
In live and constructive collective training, the AAR is a crucial component of the training process. The same is true of virtual simulation based training. The first collective training virtual simulation, SIMNET, was developed by DARPA, to provide training and better understand the technical requirements of networked simulators. SIMNET was initially developed without an AAR system, a capability to produce AAR aids from recorded movement, communication and firing data. AARs were dependent on the perceptions of the training participants. In the early 1990s the US Army Research Institute with contracting support from Perceptronics and the Institute for Simulation and Training developed the Unit Performance Assessment System [13] to capture SIMNET exercise data and provide AAR aids to support the AAR process. Later the Automated Training Analysis and Feedback System [4] was developed also for the SIMNET system. These AAR systems addressed technical issues of extracting information from the simulation data streams, reducing operator workload, and producing aids and displays that went far beyond a simple replay of the exercise.

In the 1999 – 2002 timeframe, as part of an overall project to develop capabilities for simulation-based training of dismounted combatants, an AAR system was developed called the Dismounted Infantry Virtual AAR System (DIVAARS). The goal was to develop an AAR system that incorporated lessons learned from earlier AAR systems and was tailored to the unique requirements of small unit dismounted.

Infantry training in a virtual environment. An emphasis was placed on being able to meet the special challenges of urban environments on military operations and training. The challenges are primarily visual in that “buildings and other structures break up the visual field and limit the portion of the battlefield that can be observed by any one person.” [12, p. 59]. This required an AAR system that could not just replay an exercise, but could in addition support the AAR goals of presenting exercise events and data in a manner that would facilitate trainee understanding of what happened, why it happened, and how to improve.

For determining "what" happened during a mission, the DIVAARS recreates exactly what happened during the mission. During the replay the unit members can observe the location, posture, and actions of all the other members. And, unlike live field training, DIVAARS can replay mission action exactly as viewed by any of the participants. These features not only support the trainees' explanation of why events happened, but also help the unit members develop shared mental models of individual and unit tasks. Watching the replay may also strengthen group identification and cohesiveness. Finally, several DIVAARS features (such as depicting critical events in slow motion and from multiple perspectives) enhance memory so those lessons learned are more likely to be employed in subsequent training and missions.

Figure 8-1 shows a sample DIVAARS display with many of these features that can be utilized in supporting an AAR.



1.00 Speed
100 % Zoom
Time of Day: 6:45 am
Fog Visibility
5000
00:00
02:19
01:35
Current View: Top Down
State: Paused
Entity Selected: 1B1
Mode: After Action Review
Loading PDU file picture7.pdu...Done.

Figure 8-1: DIVAARS Display.

These features are summarized below.

Playback. A linear beginning-to-end playback is unlikely to the most efficient way to provide the trainees with an understanding of what happened during an exercise. The replay system includes actions such as pause, stop, record, play, step forward, fast-forward, rewind, fast-reverse, and step-reverse. Variable playback speeds are available. Furthermore, the AAR Leader has the capability to mark events during the exercise and jump directly to them during the AAR.

Viewing Modes. Viewing scenario events from different perspectives is essential to understanding what happened. Multiple viewing modes are available to the AAR Leader during both the exercise and the AAR. Many preset views can be selected at any time prior to or during the exercise for immediate use. These can be used for perspectives or positions that the AAR Leader thinks will be useful, such as the view from an enemy position. The variety of viewing modes provides added capabilities during the AAR process.

- Top-Down – A view of the database looking straight down from above. It can be moved left, right, up, down, and zoomed in or out. The AAR Leader can also lock the view onto an entity, in which case it will stay centered directly above that entity as it moves through the database.
- 2D View – This is the traditional plan view display. It is the same as Top-Down except that depth perspective is not shown.
- Entity View – By selecting any entity, (including enemy or civilian), the AAR Leader can see and display exactly what that entity sees. This includes the effects of head turning and posture changes.
- Fly Mode – The AAR Leader can “fly” through the database using the mouse for control.

During the course of a replay the trainees will be able to see the mission from a number of perspectives. The top down, 2D, and fly views, views that are never available to them during the mission exercise, promote seeing the big picture and learning to see the battlefield. The entity view, seeing through the eyes of others, supports a number of training functions. Did the leaders see an action or problem but fail to respond, or were they not looking in the right direction at all? Do squad members maintain 360° security and report promptly? What was the view from likely and actual enemy positions? The DIVAARS replay provides unequivocal answers to those questions.

Movement Tracks. Movement tracks show, in a single view, the path an entity travelled during an exercise. Markers are displayed at fixed time intervals. Every fifth marker is a different shape than the four preceding it. The display of these markers can be turned on and off. The movement tracks provide a clear display of the path and speed of movement of each member of the unit. In addition, they provide indications of the unit formations and of the location and duration of halts in movement. Thus, the AAR Leader may elect to skip or fast-forward through portions of the replay, knowing that the movement traces for those skipped segments will be observable when the replay is resumed.

Entity Identifier. Because friendly force avatars in the DIVAARS and in the virtual simulators are not always easy to distinguish from one another, a unique identifier is shown above the avatar of each unit member. For example, 2SL is the identifier for the squad leader, second squad. The entity identifiers change size to be readable across all levels of zooming.

Digital Recording and Playback of Audio Program. Audio communications within a unit are important scenario events. DIVAARS records and plays back audio content for all scenarios. This system was developed and tested with an ASTi Digital Audio Communications System (DACS: Advanced Simulation Technology, Inc., 2001). The ASTi system converts all voice communications from live participants to digital messages and outputs them on a Distributed Interactive Simulation (DIS) network using Transmitter/Signal/Receiver Protocol Data Units (PDUs). In addition, DIVAARS records environmental audio information (for example gunshots) from the simulator via the DIS Fire and Detonation PDUs. The DIS timestamps are used to play back the audio at the correct moment during the AAR replay.

Viewing Floors Within a Building. The AAR Leader must be able to follow the action in MOUT scenarios even when a unit enters a building. The AAR Leader can select a building and then select a floor of that building to be displayed. Using this feature, the operator can view and display the avatars going through a building without the problem of walls and upper floors blocking the view.

Bullet Lines. This feature helps to determine what objects are being shot at by each entity, and to identify patterns of unit fire. Bullet flight lines are shown for all weapon firings. The line traces a shot's origin and destination. It is the same color as the originating entity. These bullet lines gradually fade away.

Event Data Collection and Display. DIVAARS has the capability to track many events including shots fired, kills by entities, movement, and posture changes. These data can be shown in a tabular format or graphical display. The AAR Leader can use them as needed to make various teaching points. They can also be used to support subsequent data analysis for research and development applications. Custom events defined by the operator are automatically flagged and can be jumped to during playback. Ten different tables and graphs are available:

- Shots fired, by entity and unit;
- Kills, by entity and unit;
- Killer-Victim table that shows who killed whom, with the option to show the angle of the killing shot (front, flank, or back) or the posture of the victim (standing, kneeling, or prone);
- Shots as a function of time, by entity, unit, and weapon;
- Kills as a function of time, by entity, unit, and weapon;
- Kills by distance from killer to victim, by entity, unit, and weapon;
- Rate of movement of each entity, and aggregated at fire team and squad levels;
- Percentage of time friendly units were stationary;
- Percentage of time friendly units were in different postures; and
- Display of user-defined events.

7.1 Evaluation and Utilization

DIVAARS was developed as part of a comprehensive program to develop capabilities for dismounted combatant virtual training. It was evaluated within the context of the exercises conducted as part of the overall research program. In general, DIVAARS has been rated very highly by Soldiers. Table 8-1 contains Soldier ratings of the systems capability to present information. The ratings were collected as DIVAARS was developed and matured. The data represents Soldier's opinions drawn from a number of different projects. After its development trials in 2001 and 2002, DIVAARS has been used as the AAR tool in the Virtual Integrated MOUT Training System testing at Ft. Campbell, KY, in 2004 [10]. It has been used to test wearable computer generated virtual Dismounted Soldier training systems in 2005.

Table 8-1: Ratings of DIVAARS by Soldiers in Dismounted Soldier Simulation Exercises

The AAR system made clear	Ratings	2001	2002	2004	2005
...what happened during a mission	*SA*	*44%*	*82%*	*62%*	*68%*
	A	*56%*	*12%*	*31%*	*32%*
	Total	*100%*	*94%*	*93%*	*100%*
...why things happened the way they did during a mission	*SA*	*44%*	*76%*	*46%*	*62%*
	A	*39%*	*24%*	*35%*	*35%*
	Total	*83%*	*100%*	*81%*	*97%*
...how to do better in accomplishing the mission	*SA*	*28%*	*71%*	*54%*	*69%*
	A	*56%*	*24%*	*38%*	*23%*
	Total	*84%*	*95%*	*92%*	*92%*

***SA = Strongly Agree; A = Agree*

It was recently included in the suite of capabilities making up the US Navy's Virtual Technologies and Environments (VIRTE) program. Within VIRTE it was used to test methods for measuring Situational Awareness in virtual environments. These measures of Situational Awareness may prove to be a highly effective means of tracking the progress of units in providing the intrinsic feedback needed to support unit performance.

8.0 CURRENT AND FUTURE RESEARCH NEEDS

Changes in training environments, operational contexts, and operational systems has driven AAR research resulting in new tools and procedures [17]. Environments, contexts, and systems continue to change, and thus this process of adapting AAR tools and procedures continues. Networked command and control systems and joint, multi-national operations are two of the variables motivating additional AAR research.

Networked command and control systems enable new forms of extrinsic feedback. Software can be used to provide feedback in mid exercise in the form of intelligent tutors [6] or "during action" review aids [1]. These software applications, if used during actual operations, become sources of intrinsic feedback. Unit responses to these forms of feedback become a new source of topics for the AAR, and they also provide their own AAR aids.

The AAR process may need to be tailored to support joint operations, multi-national operations, and distributed training exercises. Joint exercises include participants from a mix of military services and multi-national operations may involve military and civilians representing a mix of nations and/or cultures. For both joint and multi-national AARs, cultural issues may influence the utility of specific design features of the AAR (e.g., is it acceptable for a leader from another service, country or culture to have their mistakes revealed in front of subordinates or outsiders?). In many military training situations where careers and prestige is on the line, it is possible that participants may be more concerned with defending their actions than with learning how to improve their performance in the future. Successful implementation of AARs under these circumstances may require culture changes that allow for open discussion of performance strengths and weaknesses.

9.0 SUMMARY

Realistic battlefield simulations made it possible for the AAR process to replace the critique as the primary method of providing extrinsic feedback after collective training exercises. Realistic simulations provide participants with intrinsic feedback that cues and guides their performance and, to some extent, let them know how well they are performing various tasks. The intrinsic feedback received by individuals depends upon their job, their location in the battlespace, and the quality of the simulation environment. This intrinsic feedback prepares individuals to participate in interactive discussions that can help a unit decide what happened, how it happened, and how to improve performance. A significant part of the extrinsic feedback process is to bring perceived truth regarding exercise events in line with ground truth (e.g., what actually happened), and the sharing of intrinsic feedback enables a view of the situation that is closer to ground truth than is the view of a single individual. In general, less well trained units will have a greater need for extrinsic feedback, because they will have less knowledge about the ground truth situation during exercises to draw upon. In realistic battlefield simulations, a wide variety of participants are able to see, through intrinsic and/or extrinsic feedback, how their actions contributed to the bottom lines of unit performance.

The AAR process makes use of the Socratic method of asking leading and open-ended questions to guide unit discussions. The AAR process can be expedited through the use of aids that use electronic data streams from exercises to document key aspects of performance that are close to the root causes of unit strengths and weaknesses. Designing these aids and implementing their production is a continuing activity.

The AAR process and/or AAR aids have been tailored many times to fit specific instances of the live, virtual, and constructive training environments and to fit changes in unit equipment and missions. This tailoring activity continues to the present.

10.0 REFERENCES

[1] Aiken, D.S., Green, G.E., Arntz, S.J. and Meliza, L.L. (2005). *Real time decision alert, aid and after action review system for combat and training* (ARI Technical Report 1165). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[2] Banks, J.H., Hardy, G.D., Scott, T.D., Kress, G. and Word, L.E. (1977*). REALTRAIN validation for rifle squads: Mission accomplishment* (ARI Research Report 1192). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[3] Bosley, J.J., Onoszko, P.W.J., Knerr, C.S. and Sulzen, R.H. (1979). *Tactical engagement* simulation *training techniques: Two training programs for the conduct of after action review* (ARI Research Product 79-2). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[4] Brown, B., Wilkinson, S., Nordyke, J., Riede, D., Huyssoon, S., Aguilar, D., Wonsewitz, R. and Meliza, L. (1997). *Developing an automated training analysis and feedback system for tank platoons* (ARI Research Report 1708). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[5] Darling, M.J. and Parry, C.S. (2001). After action reviews: Linking reflection and planning in a learning practice. *Reflections,* 3, 2, 64-72.

[6] Domeshek, E.A. (2004). Phase II. Final report on an intelligent tutoring system for teaching battle command reasoning skills (ARI Technical Report 1143). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[7] Goldberg, S.L., Mastaglio, T.W. and Johnson, W.R. (1995). Training in the Close Combat Tactical Trainer. In R.J. Seidel and P.R. Chatelier (Eds.), *Learning without Boundaries: Technology to Support Distance/Distributed Learning* (pp. 119-133), Plenum Press, New York.

[8] Goldberg, S.L. and Meliza, L.L. (1993). Assessing unit performance in distributive interactive simulations: The Unit Performance Assessment System (UPAS). *Proceedings of NATO Defence Research Group Meeting, Panel 8 (Defence Applications of Human and Bio-Medical Sciences). Training Strategies for Networked Simulation and Gaming). Technical Proceedings AC/243 (Panel 8) TN/5*, 173-182.

[9] Hoare, R. (1996). From debrief to After Action Review (AAR*). Modern Simulation & Training*, 6, 13-17.

[10] Knerr, B.W. and Lampton, D.R. (2005). *An assessment of the Virtual–Integrated MOUT Training System (V-IMTS)* (ARI Research Report 1163). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[11] Kulik, J.A. and Kulik, C-L.C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, 58, 79-97.

[12] Lampton, D.R., Clark, B.R. and Knerr, B.W. (2003). Urban Combat: The Ultimate Extreme Environment. *Journal of Performance in Extreme Environments,* 7, 57-62.

[13] Meliza, L.L., Bessemer, D.W., Burnside, B.L. and Shlechter, T.M. (1992). *Platoon-level after action review aids in the SIMNET Unit Performance Assessment System* (ARI Technical Report 956). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[14] Meliza, L.L. (1999). *A guide to standardizing after action review (AAR) aids* (ARI Research Product 99-01). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[15] Meliza, L.L., Scott, T.D. and Epstein, K. (1979). *REALTRAIN validation for rifle squads: Tactical performance* (ARI Research Report 1203). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[16] Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science* 32, 99-103.

[17] Morrison, J.E. and Meliza, L.L. (1999). *Foundations of the after action review process* (ARI Special Report 42). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[18] Rogers, E. (2004). "Pausing for Learning: Adapting the Army After Action Review Process to the NASA Project World" NASA White Paper. Retrieved August 18, 2005, from http://smo.gsfc.nasa.gov/knowman/documents/whitepapers/ Pausing_for_Learning.pdf

[19] Scott, T.D. and Fobes, J.L. (1982). *After action review guidebook I. National Training Center* (ARI Research Product 83-11). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[20] Scott, T.D., Meliza, L.L., Hardy, G.D., Banks, J.H. and Word, L.E. (1979). REALTRAIN validation for armor/anti-armor teams (ARI Research Report 1204). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[21] Shlechter, T.M., Bessemer, D.W., Rowatt, W.C. and Nesselroade, Jr., K.P. (1994). *Evaluating the Unit Performance Assessment System's after action review displays* (ARI Technical Report 997). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[22] Sulzen, R.H. (1986). *Annotated bibliography of tactical engagement simulation 1966-1984* (ARI Technical Report 725). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

[23] U.S. Army Combined Arms Center (1993). *A leader's guide to after-action reviews* (Training Circular 25-20). Fort Leavenworth, KS.

[24] Shriver, E.L., Mathers, B.L., Griffin, G.R., Jones, D.R., Word, L.E., Root, R.T. and Hayes, J.F. (1975). *REALTRAIN: A new method for tactical training of small units* (ARI Technical Report S-4). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Technical Evaluation Report

Peter Werkhoven
TNO Defense, Security & Safety
Delft
THE NETHERLANDS

Peter.werkhoven@tno.nl

1.0 INTRODUCTION

1.1 Background

Recent improvements in computer systems and displays have enabled new simulation technologies such as Augmented, Mixed, and Virtual Environments (AMVE). Increased computer power at low cost, wireless networks, miniaturizations of sensor and computer components, and better visual, auditory and tactile display systems are contributing to the maturation of these technologies. Potential applications in military operations, as well as training and system design are providing requirements that have spurred this technology development.

Today, most of the attention is focused on the development of the technologies themselves. However, to be effective in military operations, the technologies must evolve into systems that provide the information that their human users need to accomplish military objectives. Compared to research on computer architectures, communication protocols, and display devices there has been relatively little research on the perceptual requirements for displays, human-computer-interaction issues, design of effective training approaches, measurement of human performance and cultural and organizational issues. The fundamental knowledge available today already indicates a large potential of AMVE technology for a broad spectrum of military applications.

An important outcome of a previous workshop "What is essential for Virtual Reality systems to meet military human performance goals" of the NATO Research Study Group HFM-021 in the year 2000 was that:

- Baseline applications of VR were solid in the automotive industry and entertainment industry, and military applications were *beginning to emerge* and be evaluated; and
- Successful application of VR depends strongly on:
    - The quality of the interaction methods of humans with VR (train like we fight),
    - On the on the level of *fidelity* of the virtual world, and
    - On the multidisciplinary involvement of scientists, engineers, practitioners and *users*.

This workshop titled "Virtual media for military applications" organized by the NATO Research Study Group HFM-136 focused on the broader set of AMVE technology for military applications. Military users were brought together with academic researchers and industry to discuss if AVME meets operational needs. It provided a unique opportunity to assess advances in AMVE technology and to monitor the progress of the recommended role of human factors research and user involvement in the development of AMVE applications.

1.2 Purpose and Scope of this Workshop

Chairman Thomas Alexander pointed out that the purpose of this workshop was:

- To summarize previous and on-going research (and see if AMVE technology indeed provides the intuitive human-system interaction expected);
- To identify the keys for implementation of ready to use technology; and
- To identify knowledge gaps and thrusts and establish an agenda for future efforts that explore the human dimensions of virtual media for military applications.

The workshop concentrated on the following research areas:

- Training methods,
- Human performance requirements,
- Performance measurement techniques and assessment, and
- Human Factors issues in design and utilization;

with a focus on the following application areas:

- Command, Control, Communication, Computer, Information, Surveillance and Reconnaissance (C4ISR) systems, in particular user interface issues;
- Tele-presence, tele-operation, and tele-manipulation in reconnaissance, surveillance, and target acquisition;
- Military training and simulation;
- Mission preparation and rehearsal;
- Systems acquisition; and
- Mission support (maintenance, decision aiding, logistics, navigation).

1.3 Program Workshop

The workshop took place over three days with the following structure:

Tuesday, June 13, 2006

- Keynote Address by professor Paul Milgram (Toronto University, Canada): How the concept of Mixed Reality Encompasses Augmented Reality and Virtual Environments.
- Session 1: Command and Control
  - Chairs: Thomas Alexander (Research Establishment for Applied Sciences FGAN, Germany) and Patrik Lif (Swedish Defence Research Agency FOI, Sweden).
  - Speakers:
    - Thomas Alexander (FGAN, Germany): Applicability of Virtual Environments as C4ISR.
    - Jared Freeman (Aptima Inc. US): Measuring, Monitoring, and Managing Knowledge in Command and Control Organizations.

    - James Lussier (US Army Research Institute, US): Components of Effective Learning.
    - Neville A. Stanton (Brunel University, UK): Experimental Studies in a Reconfigurable C4 Test-bed for Network Enabled Capability.
    - Andreas Tolk (Virginia M&S Center, US): Challenges and Potential of Service-oriented Architectures for Net-Centric Operations.
    - Hendrik-Jan van Veen (TNO Defense, Security and Safety, The Netherlands): SIMNEC, research platform for studying human functioning in NCW.

- Session 2: Tele-Operations
  - Chairs: Ebb Smith (Defence Science and Technology Laboratory DSTL, UK) and Professor Robert Stone (University of Birmingham and Director of the Human Factors Integration Defence Technology Centre (HFI DTC)).
  - Speakers:
    - Robert J. Stone (University of Birmingham, UK): Serious Gaming Technologies Support Human Factors Investigations of Advanced Interfaces For Semi-autonomous Vehicles.
    - Jan B.F. van Erp (TNO Defense, Security and Safety, The Netherlands): Tele-presence: Bringing the Operator Back in the Loop.
    - Michael Barnes (US Army Research Institute, US): Understanding Soldier Robot Teams in Virtual Environments.
    - Boris Trouvain (FGAN-FKIE, Germany): Tele-operation of Unmanned Vehicles, The Human Factor.
    - Robert Taylor (DSTL, UK): Human Automation Integration for Supervisory Control of UAVs.

Wednesday June 14, 2006

- Keynote Address by COL James Shufelt (US Army, TRADOC program Integration Office): The Future Role of Virtual Simulators/Simulations in U.S. Army Training.
- Session 3: Vehicle Simulation
  - Chairs: Lochlan Magee (DR&D, Canada) and Sylvain Hourier (IMASSA, France).
  - Speakers:
    - LTC Wil Riggins (US Army, Program Executive Office, Simulation, Training and Instrumentation): Requirements for Virtual Vehicle Simulation.
    - Leonhart Vogelmeier (EADS, Military Aircrafts): Interaction Method for Virtual Reality Applications.
    - Brian Schreiber (US Air Force Research Lab and Lumir Research Institute, Arizona): Evaluating Mission Training Fidelity requirements, Examining key Issues in Deployability and Trainability.
    - Bernd de Graaf (TNO Defense, Security and Safety, The Netherlands): Military Simulation Centre: Vehicle for Validation of Military Flight Simulation.

    - Mark Espanant (CAE, Canada): The Application of Simulation to Study Human Performance Impacts of Evolutionary and Revolutionary Changes to Armored Vehicle Design.

- Session 4: Dismounted Simulation
  - Chairs: Nico Delleman (TNO Defense, Security and Safety, The Netherlands) and LtCmdr Joseph Cohn (Naval Research Laboratory, US).
  - Speakers:
    - Phillip Jones (MYMIC Low Liability Company, US): Emerging Requirements for Dismounted Virtual Simulators.
    - Amela Sadagic (MOVES Institute, US Naval Postgraduate School): Combined Arms Training, Methods and Measures for a Changing World.
    - Bruce Knerr (US Army Research Institute, US): Current Issues in the Use of Virtual Simulations fro Dismounted Soldier Training.
    - Frederick P. Brooks, Jr. (University of North Carolina, US): Challenges of Virtual Environment Training of Dismounted Teams.
    - James Templeman (Naval Research Lab, US): Immersive Simulation to Train Urban Infantry Combat.
    - John Frim (Defence R&D Canada): Use of the Dismounted Soldier Simulator to Corroborate NVG Studies in a Field Setting.
- Keynote Address by Bowen Loftin (Texas A&M University, US): The Future of Simulation.

Thursday June 15, 2006

- Session 5: Mixed and Augmented Reality
  - Chairs: Stephen Goldberg (US Army Research Institute, US) and Lisbeth Rasmussen (Danish Defense Research Establishment, Denmark).
  - Speakers:
    - Mark A. Livingston (Naval research Lab, US): Battlefield Augmented Reality.
    - Jarrell Pair (University of Southern California, Institute for Simulation and Training, US): Performance Assessment in Immersive Mixed and Virtual Environment Systems.
    - Ed Bachelder (Systems technology Inc., US): Helicopter Air Crew Training using Fused Reality.
    - Brian F. Goldiez (University of central Florida, Institute for Simulation and Training, US): Human Performance Assessment when using Augmented Reality for Navigation.
    - Sheila Jaszlics (Pathfinder Systems Inc., US): The DARTS Augmented Reality System.
    - Matthew Franklin (QinetiQ, UK): The Lessons Learned in the application of Augmented Reality.
    - Stephen Ellis (NASA-ARC, US): The User-interface to virtual or augmented environment displays for the Air Traffic Control.

- TER and Keynote Speaker Comments
  - Prof Peter Werkhoven (TNO Defense, Security and Safety, The Netherlands).
  - Prof Paul Milgram (Toronto University, Canada).
  - COL James Shufelt (US Army, TRADOC Program Integration Office).
  - Prof Bowen Loftin (Texas A&M University, US).

1.4 Attendees

Distribution of attendees across nationality and affiliation:

Country	**Total**	**Defense Research Institutes**	**Military**	**Industry**	**Academia/ Civil Res. Inst.**
Canada	3		1	1	1
Denmark	1	1			
France	1	1			
Germany	3			1	2
The Netherlands	6			1	5
Sweden	2	2			
United Kingdom	5	3			2
USA	20	9	3	3	5
Total	**41**	**16**	**4**	**6**	**15**

2.0 TECHNICAL/SCIENTIFIC SITUATION OF MILITARY VR APPLICATIONS

2.1 Introduction

2.1.1 Technological Developments

- We see a rise of mixed reality (Milgram) technology. With current technology real and virtual worlds can be mixed. The amount of virtual varies from seeing a virtual bird in your real environment to seeing your real hand in a virtual environment. In the first example only the bird has to be modeled. In the second the complete environment. More complex examples include virtual in real in virtual. Milgram presented a theoretical framework to categorize various forms of mixed reality. From a technical point of view three dimensions can be distinguished: virtual versus real, ego-centricity versus exo-centricity and display congruency. It would be interesting to include social dimensions and other modalities than visual.
- Industry claims that robust Augmented Reality (AR) – representing virtual entities in real environments – is technically feasible (Jaszlics). Furthermore, the AR community is supposed to be "big enough" to increase Technology Readiness level (TRL) and that robust AR is technically feasible (Jaszlics). Wireless stereoscopic networked AR demonstration systems are available (e.g., DARTS system). However, some

technological challenges with respect to tracking and occlusion are still remaining. Accurate tracking of the body parts of a soldier in the open field is still years away (Ellis) – but we can start looking at applications that don't need this accuracy or applications in instrumented environments. The ability to occlude parts of the real world that are covered by virtual elements generally still relies on immature technology such as real time 3D reconstruction of the real word and filtering the real world (video merging or occlusion displays). Furthermore the resolution of (see-through) head mounted display is generally not high enough, for example for AR training of Forward Air Controllers (Franklin). Research into the visual requirements for augmented reality displays for the airport tower has shown that virtual information that occludes the real world may leave out currently used dynamic visual cues that provide lead information used for spacing and sequencing during air traffic management tasks (Ellis). This warning probably has a more general scope.

- In controlled conditions (e.g., cockpits and helicopter training) the concept of fused reality may be an alternative for real time 3D reconstruction of the real world. Fused reality makes use of live video capture and VR merged by chroma-key (magenta) and luma-key technology allowing real objects to move in virtual space. For example in helicopter training it eliminates the need for expensive computer models (Bachelder).
- The value of 3D sound in virtual training generally seems to be underestimated. For example, in the case of forward air control 3D sound can compensate for low resolution displays: we hear the plane before we see it (Franklin).
- AR applications are starting to look at embedded training of team operations. The Battlefield Augmented Reality System (BARS) support dismounted soldiers in room clearing operations. The information provided through head mounted displays changes on criticality and need. Voice and gesture commands can be used to interact with the system. The research focus is now on multiple users and team performance. The use of head-mounted displays is still cumbersome and limits human communication.
- Training in mixed reality may have specific effects on realism, presence, affective appraisal, etc. This poses high demands on cognitive performance tests. Test batteries are needed to assess cognitive performance in mixed reality. A test battery to assess attention, spatial ability, memory, reasoning abilities will be completed in 2007 (pair).
- It has become clear from many discussions on application development that application developers move away from high-end high-fidelity simulators towards low-end simulation. On one hand head mounted displays are still cumbersome and high-end simulators are still costly. On the other hand many training tasks turn out not to rely on realism and can be done using powerful commercial off–the-shelf-game engines (Stone). It most be noted however, that high-end simulators (CAVE, motion platforms) are still required for high-fidelity skill training including eye-hand coordination tasks (e.g., surgery, flight simulation) and multi-user spatial awareness tasks (e.g., room clearing with teams). We always need to go back to "what are the training tasks?". High fidelity isn't always the answer. We need to design to optimize training for specific tasks (Brooks).
- Interface technology has not made much progress. The workshop has not revealed any innovations in this field. This is remarkable because a strong recommendation of a NATO workshop in 2000 discussed the need to improve the quality of interfaces with VR. Tracking is still a problem, head-mounted displays are still cumbersome, 3D audio is not used much, force feedback is still in its infancy. Although tactile displays are available for the torso we haven't seen applications that make use of it. Requirements generators and hardware developers still express the need for intuitive interfaces allowing "train as you fight" simulations. In particular there is a need for improved ergonomic design of mobile AR applications for hands free multimodal interfaces (Bachelder).

2.1.2 Scientific Disciplines Involved

- Since the recognition of the importance of intuitive man-machine interaction and intelligent system adaptation to human conditions, simulator engineers have involved psychologists and biologists to improve human performance capabilities within these systems.
- Today we see the need for more disciplines to be included in development of effective AMVE technologies. The research field of simulation deals with modeling (virtual and constructive) worlds, objects and behavior, scenario generation, training methods, multi-sensory interfaces and didactic concepts, but also with drama, style and emotions (the gaming element). These components have only had limited application to AMVE applications – but this is why, for example, screenwriters now collaborate with the U.S. intelligence community on "war gaming" and why Paramount Digital Entertainment collaborates with the US Department of Defense on "crisis-management simulation". While it has been claimed that these components have substantial impact on user experience, learning and training transfer, there is only limited research to backup the claims. So, the creative disciplines and liberal arts are beginning to be involved in serious simulation.
- With the introduction of new doctrine and networked (joint and combined) command and control capabilities (NEC) new requirements for the use of AMVE technologies are being developed.
  - We need to cope with a deluge of data and information overload at all levels (tactical, operational and strategic) (Alexander). How will information be "funneled" to the appropriate user? How will information be filtered according to classification and sensitivity to ensure it gets to the right people at the right time and in the right format? Obviously other dimensions than technology become important such as culture, organization, leadership, etc.
  - There is a need for a theoretical framework and predictive models on the complex social and organizational dimensions as a direct consequence of the potential of dynamic task allocation, common operational pictures, commanders talking directly to soldiers and aircrew, etc. There are no sufficient theoretical frameworks to model such complex socio-technical systems. First socio-technical systems that should help leaders to measure monitor and manage are being developed (Freeman), but not yet evaluated.
  - There is a gap in knowledge on how quickly we can learn and adapt to the use of new technologies. How well do people learn in highly dynamic structures? Decisions may get easier, but actions may get harder.
  - Scale and complicity drive us towards concept development and experimentation environments with an important role of VR for creating large scale distributed multi-user virtual environments including real and virtual humans, sensors and systems and complex decision structures and dependencies.
- Conclusion: New disciplines must urgently get involved: creative disciplines (scenario and gaming elements), social disciplines (dimensions of socio-technical systems), organizational disciplines (mechanisms and mathematics of self organizing networks).

2.1.3 Evolved Application Fields

Command and Control:

  - Current virtual command and control training systems generally just provide monitoring functionality and lack sufficient feedback and instructional mechanisms (Lussier). It is crucial to include feedback mechanisms and to be able to measure thinking skills and judgments of tactical situations.

- Network enabled command and control deals with dynamic roles and functions, less well defined context for decision making, real time information sharing and a global scale of operations. Central in network enabled command and control is to ability to share relevant information across networks at the right time and in the right form. Although technology currently enables common operational pictures (shared information), it is still a challenge to create common mental models (shared situation awareness). Tolk pointed out that progress in this field relies more on cultural factors (e.g., the willingness to share information, trusting others who are at a distance and allowing for misinterpretations due to cultural differences) than on technology. Further, progress relies on dealing with organizational issues rather than technological ones.
- Challenges for the use of network enabled capabilities (NEC) are the dynamic reallocation of functions during NEC-operations, and tools for supporting multiple command levels (van Veen). When experimenting with Networked Enabled command and control (C2) concepts in simulated worlds it must be possible to vary these aspects and to monitor behavior and performance. The human factors community currently lacks a theoretical framework and predictive models for designing NEC-concepts in laboratories or test beds. . Simulated worlds provide experimentation environments for exploring concepts and development processes. Advanced simulated test beds facilitate the translation of operational needs into new concepts and produce new research questions. Test beds also provide a means to try out new technological innovations to evolve further new concepts and capabilities.
- Public sources such as “Google Earth” have become useful for creating simulated training environments for urban operations (Stanton).

Tele-Operations:

- High fidelity virtual training environments for tele-operations making use of head-mounted displays have not always brought the success expected. Stone opined that early VR technology ‘failed to deliver’. Low-fidelity gaming environments may provide a new promise. Stone used various commercial game engines and experiments favoring a “quick and dirty” approach to human factors research. This approach relies on high frequency short cyclic concept development in which the quality of man-machine interface is implicit and develops in an evolutionary manner. This seems to be the way the commercial gaming world works. However, it should be noted that the game industry has been successful in adopting many of the research results that have come from thorough academic human factors studies. It is a misunderstanding that evolutionary processes are sufficient by themselves and that funding of n academic human factors studies is a poor investment. For example, within some communities high-fidelity systems are still thought to be essential for effective training transfer of perceptual-motor skills (flying, surgery, etc.) when some recent research has indicated that this in not necessarily the case.
- Given our current technology, fully autonomous robotic systems do not seem to be the right approach.. Experiments (van Erp) have shown that human intelligence is still needed and that the operator must be kept in the loop facilitating some sort of tele-presence. The use of HMDs has not been shown to be successful for tele-presence. Attention has shifted towards BOOM-displays. It should be noted that many experimental results cannot be generalized, but only hold for the specific design tested.
- Having a human in the loop does not mean that human can necessarily control everything simultaneously. Successful tele-operations must allow for variable and task dependent autonomy (Stanton).
- Alternatives to tele-presence environments for remote control are declarative user interfaces (Trouvain). Specialists operate robots, but commanders must specify navigational and manipulation

tasks at a higher and more overall task objective oriented level. This may be an advantage for operating swarms of robots in which tele-presence at multiple locations becomes impossible.

Vehicle Simulation:

- An interesting application of VR in the field of vehicle simulation is assembly and maintenance training. These applications require intuitive manipulation techniques for object handling, such as virtual hand control. The state of the art of existing interface technologies are still far from meeting functional requirements (Vogelmeier). We still have to work with data gloves, head-mounted displays inaccurate tracking devices and rudimentary force feedback systems. No interaction method can generally be applied. Given the limitations of current technology we have to fall back to task and subtask specific interface designs. For many purposes, real mockups are still the best way to go. There is an urgent need to explore new ways of interfacing with virtual worlds, for example brain machine interaction.
- An important aspect of vehicle simulation is motion cuing, the simulation of forces that a vehicle exerts on the user. In particular, for virtual training motion, cuing may be of crucial importance for training transfer. Evidence for this, however, is sparse. A very advanced experimental motion platform (Desdemona) with all degrees of freedom and the option of sustained 3 G (max) is currently being built at TNO in the Netherlands (de Graaf). It will be used for research regarding the characteristics of human motion senses and for studying the added value of motion cuing for various driving and flight simulations. For the study of joint mission training requirements it has been linked to real F16 cockpits, C2 facilities and tele-presence control systems.
- Fully virtual training environments seem to no longer meet the functional requirements for vehicle simulation. Furthermore, head mounted displays have been shown to destroy communication between crew members. Consequently a mix of live, virtual and constructive environments (LVC) are being explored for the evaluation of future armored vehicles (Espenant). LVC enables rapid prototyping of vehicles and can be used to assess future technologies that are currently planned. Unfortunately, objective performance measures to determine the effectiveness of these new simulation concepts are still lacking.

Dismounted Soldier Simulation:

- The dynamics of real missions has not yet been sufficiently captured in dismounted virtual simulators. Although technology is not rigorous enough for effective simulations of many dismounted tasks, the main causes of limited support for the development and adoption of training simulators for dismounted soldiers are cultural (Jones). Leaders want to train in real environments, soldiers want to “go out and get cold”. It is simply not the same in a simulator. Further research should reveal to what extent exciting scenarios, new forms of role playing and the “x-factors” from the gaming industry can attract game savvy soldiers into virtual simulators Experiments made clear that game playing experience has implications for the motivation of trainees and the effectiveness of training (Knerr). The next generation trainees are gamers who don’t like games that have been designed for training (not enough fun). Currently a big gap exists between gaming and training applications. If tomorrow’s military are experienced gamers, will training solutions that use game technology be sufficiently engaging for training purposes?
- Formerly, human factors experts thought they knew better what was good for the user than the user did. Nowadays users are heavily involved in all phases of the evolutionary development process of successful training simulators. It is important to let training systems adapt organically and to evaluate

on the fly (Sadagic). The combination of many users and short design cycles provides the feedback necessary to design and develop effective training simulators.

- We form cohesive teams of soldiers by letting them share experiences and stories. VR is a potentially powerful training tool to support these processes (Brooks). An important question, however, is how to built *cost-effective* training systems? An approach proposed by Brooks is to first built effective systems and than cost reduce, not the other way around. The challenges for the coming years are to bring team members in one space. How to provide personal displays without head-Mounted Displays? How to track each of the members (diagnostics) and how to build in "pinball scores" to make the training attractive are questions that need to be answered. A success factor is to involve users from the start to ensure successful early adoption of the training system. Technological challenges are rapid and cheap scenario generation, model acquisition by laser and video and scenario capture by computer vision.
- Dismounted soldier simulators are in urgent need of intuitive interfaces for navigating the virtual world. A good example of human factors engineering has led to an innovative Sony game console based navigation interface (Templeman) that matches the natural motion of soldiers in the field. High-fidelity navigation, however, remains an issue and requires the combination of real movements in virtual environments which needs further research.
- Mission rehearsal is a subset of training and may require less fidelity than training. Simulation has also been set up for pilot selection and is considered a part of tests for promotion. In general the emphasis on dismounted soldier training is on cognitive skills. For this purpose, low-end simulations and game use are on the rise.

Emerging Applications Fields:

- VR for selection of personnel (suggestion Milgram);
- VR for medical assessment, cognitive tests, and treatment (Pair); and
- Gaming environments for recruitment (US Army game).

2.1.4 Requirements Perspective

- The US Army TRADOC Program Integration Office, Virtual (TPIO,Virtual) (Shufelt) has developed a clear vision and strategy on future combat training systems. Key for future virtual training systems will be that they are: combined arms, full spectrum, embedded and all terrain. In 2025 approximately 10% of the total force should be able to train according to these principles. We will see fading boundaries between training and operations. Embedded training devices will also serve to realize operational networked enabled capabilities for net centric warfare. The most important technology trends identified are the rise of joint experimentation networks, the use of (commercial) game technology, the development of virtual humans and augmented reality displays. Requirements for these developments are interoperability, "plug and train" functionality, the availability of terrain databases within 48 hours, multilevel security and cognitive models for virtual instructors.
- The US Army's PEO,STRI (Riggins) wants to strengthen expeditionary capabilities by developing trainings systems with a highly modular composition. The strengths of current training systems are mobility and tactical and spatial fidelity. Areas for improvement are virtual instructor functionality, train as you fight functionality (combinations of live and constructive) and mission rehearsal functionality. Progress is not solely dependent on technology development. It has been recognized that all lines of

development should progress in balance: doctrine, materiel, leadership, facilities, operations and personnel. The focus will be on collective (not just individual) training functionality.

2.1.5 Application Development Approach

- Much of the conclusions of human factors research on the effectiveness of VR and mixed reality are still in terms of “don’t know” and “depends on”. This illustrates the complexity of the research questions. Human factors knowledge traditionally focused more on individual tasks in controlled conditions. There are insufficient theoretical frameworks and predictive models for more complex socio-technical systems and a system of systems approach. For training situations that deal with collective performance most human factors questions are still unanswered because cognitive models of group behavior are lacking (Sadagic and Darken).
- In order to progress on research questions more explorative concept development and experimentation methods are being developed. It is a challenge to find new methods for structuring concept development and experimentation such that existing human factors knowledge is brought optimally into the design and such that we do not arrive at suboptimal solutions.

2.2 Bottlenecks and Opportunities

Based on the observations listed in Section 2.1 we summarize the following opportunities:

- Agendas and strategies of defence organizations world wide reveal a high priority for the realization of embedded virtual training capabilities to strengthen the ability to train while deployed to locations around the world. The following key technology trends have been identified: joint experimentation networks, the use of (commercial) game technology, the development of virtual humans and augmented reality displays. Embedded training devices will also serve to realize operational networked enabled capabilities for net centric warfare. The focus will be on collective (not just individual) training functionality.
- Emerging applications fields are VR for selection of personnel, VR for medical assessment, cognitive tests, and treatment and gaming environments for recruitment.

Section 2.1 also showed the following bottlenecks and challenges:

- Mixed reality (merging of real and virtual worlds) is promising, but the technology for tracking body positions in the open field and real time real world modeling are still immature.
- Validated test batteries are needed to assess cognitive performance in mixed reality (attention, spatial ability, memory, reasoning abilities).
- Interface technology has not made much progress in the last six years (head-mounted displays are still cumbersome and destroy communication, force feedback is still in its infant years). There is a need for improved ergonomic design of mobile AR applications and hands free, intuitive, multi-modal interfaces allowing “train as you fight” simulations.
- For training situations that deal with large scale collective performance in networked operations most human factors questions are still unanswered because predictive cognitive models of group behavior are lacking. Progress urgently requires the involvement of social disciplines (dimensions of socio-technical systems, cultural factors), organizational disciplines (mechanisms and mathematics of self organizing networks) and creative disciplines (scenario and gaming elements).

- The main causes of limited use of the potential of training simulators for dismounted soldiers are cultural: leaders want to train in real environments and soldiers want to "go out and get cold". The case for training in virtual worlds to prepare for training in live environments has yet to be effectively made.
- Current training systems usually meet the immediate training needs, but are not sufficiently designed to allow for flexibility and evolution.
- The development of many virtual training applications has not yet balanced all lines of development: doctrine, materiel, leadership, facilities, operations and personnel.
- More interchange of ideas is needed between research community, system developers, requirement people and military users, especially in C2/NEC. Often there is good evidence and data to support a given decision, but requirement people supporting another decision may not be aware of it.
- Services need to communicate better resulting in effective transfer of knowledge, lessons learned, etc.
- There are no funds for longitudinal studies. Longitudinal studies are essential for measuring long term adaptation of users to new technology.
- Human factors conclusions are often design specific. There is a strong need for more generic guide lines to guide developers and optimize system design.

2.3 Recommended Actions

We have come to the following recommendations:

- Create fully integrated joint Concept Development and Experimentation facilities for the development and training of Networked Enabled Capabilities (dynamic reallocation of functions, switching micro/macro command levels, trust in command at distance, etc.).
- Focus research on:
  - Theoretical framework on socio-technical systems and predictive cognitive models for the effective structuring and evaluating of Concept Development and Experimentation (CDE) processes with a focus on collective performance;
  - Virtual characters and virtual instructors;
  - Non-obtrusive intuitive multimodal (brain machine) interfaces for hands free navigation and manipulation, allowing (non-verbal) team interactions and "train like you fight";
  - Rapid scenario generation and scenario capture;
  - Non-obtrusive mixed (augmented) reality displays;
  - Tracking technology for the open field;
  - Validated test batteries to assess cognitive performance in mixed reality (attention, spatial ability, memory, reasoning abilities); and
  - Generic human factors guidelines versus design specific results.
- Create budgets for longitudinal studies on adaptation of users to new technology.
- Involve organizational, social and cultural disciplines.
- Stimulate dual use of commercial gaming technology and military simulation technology.

- Design training systems to allow for flexibility and evolution.
- Balanced all lines of system development: doctrine, materiel, leadership, facilities, operations and personnel.
- Better organize the dialog between the research communities, system developers, requirement people and military users, especially in the field of C2 and NEC.
- Better organize the dialog between services resulting in an effective transfer of knowledge and lessons learned.

The enthusiasm of the workshop attendees and the evident willingness to share ideas and to discuss their findings provides a promising base for working on these recommendations.

Please note that recommendations below generated by NATO study group HFM-021 in 2000 are still unfulfilled needs today:

- The military should develop a vision on the use of VR technology and more clearly specify their needs.
- Industry should work on standardization and should substantially bring human factors into their design and development processes.
- Academia and research institutes should coordinate and accelerate their long-term research efforts to focus on natural interfaces (innovative metaphors) and on how to model human and object behavior.
- In the short term academia should focus on human factors metrics and metrics for team and collective performance (cognition, communication), and a standard evaluation methodology.
- In general, better coordination between military organizations, industry and academia is necessary in order to identify gaps in current knowledge and coordinate research.

3.0 CONCLUSIONS

The workshop “Virtual media for Military Applications” uniquely has brought together prominent people from academia, industry and the military. Workshop attendees enthusiastically and knowledgeably:

- Exchanged operational requirements and on-going research in the field of augmented, mixed and virtual reality technology (AMVE) for military applications;
- Identified success factors and bottlenecks for implementation; and
- Have made recommendations for future research agenda’s, methodological approaches and organizational issues.

The main conclusions are:

- For training situations that deal with collective performance most human factors questions are still unanswered because predictive cognitive models of group behavior are lacking. Serious research budgets for longitudinal studies, fully integrated concept development and experimentation facilities and the involvement of social, organizational and creative disciplines are necessary to find the answers.
- Developers of virtual media applications make significant use of public data bases and commercial off the shelf gaming technology. Gaming technology evolves quickly with extremely short cycles and

massive numbers of users. As such, it generates implicit human factors knowledge together with technological advances which seems to put military training system developers and the traditional human factors community at the "tail" of the gaming community. Dual use of gaming and military simulation technology should be further stimulated. Military researchers should not try to keep pace with gaming technology, but rather work to adapt technology advances to unique military applications.

- Exciting pushes are seen into "new" territories such as serious gaming, unmanned vehicles and robotics (tele-presence), dismounted training capabilities and emerging applications in the field of personnel selection and recruitment and medical treatment.
- The dialog between the research communities, system developers, requirement people and military users and between services themselves has to be improved for a more effective transfer of knowledge and lessons learned.

3.1 Future Meetings

In 2000, HFM-121 held a workshop to look at the applications of virtual technologies to military use. In many respects HFM-136 was a re-look at an expanded realm in which AMVE technologies are being used by the military. Six years have gone by, but as noted above, much work still needs to be done to provide NATO militaries with AMVE systems that are usable by soldiers, sailors and airman and effective for the purpose they were designed.

The direction of change has taken some unanticipated turns since 2000. Full scale VR does not seem as likely a solution as once thought. Game technology, augmented reality, and embedded capability seem to be the direction requirements and technology is going. The need to ensure that human considerations are taken into account is a continuing requirement. There will need to be a follow on to this workshop that explores in more detail the current trends as noted in this TER. In particular the embedding of training in weapon systems is a topic that will include considerations of live, virtual and constructive simulation and will be worthy of further study.

Conclusions

Thomas Alexander
FGAN-Research Institute for Communication
Information Processing and Ergonomics (FKIE)
Wachtberg
GERMANY

Stephen L. Goldberg
U.S. Army Research Institute for the Behavioral and Social Sciences
Orlando, Florida
USA

This report summarizes the results and conclusions achieved during the three year period of operation of HFM-121 / Research Task Group 042 and an associated, application-oriented workshop (HFM RWS-136) on Virtual Media for Military Applications. HFM-121's goal was to identify and explore how human factors considerations figure in the design, use, and performance evaluation of Augmented, Mixed, and Virtual Environments (AMVE). The task group was also to present an overview of how these considerations are impacting present and future military applications. The group was to characterize the scientific, technological, and application development of AMVE.

The overall conclusion reached by the task group was that AMVE technologies have become much more useful for a variety of military application areas than they were when a NATO task group last examined the area in 2000. In comparison to the situation in 2000, AMVE technologies have become more capable, available, and affordable.

Display capabilities and rendering performance have increased dramatically. According to Moore's law, computing power doubles every 18 months. Whilst only a rough estimate, it characterizes the development in computer graphics as well. For example, previously, Virtual Environment (VE) systems required special and expensive graphic workstations. They were replaced by personal computer clusters, and more recently by widely available, relatively inexpensive personal computers with very capable graphics cards. This has made VE a more affordable technology causing a broad push forward for all kinds of applications. There are many success stories of VE applications, primary in military education and training that show that VE can be useful, and sometimes necessary to achieve the goals of the application. However, AMVE cannot be considered an intuitive technology, which can be applied instantly. Instead, a detailed analysis has to be performed in order to identify the applicability of AMVE-technology to represent complex processes or tasks. This analysis includes the consideration of human factors on different levels in order to make optimal use of the capabilities of the new technology.

AMVE systems represent complex human computer interactions. The presentation of information through the senses requires a design that must consider fundamental perceptual human factors. Perceptual requirements drive the technology and display components used. Perception, e.g., visual acuity, color and luminance distinctions, have to be taken into account to decide which display to use for a specific system design. The same is true for acoustic and haptic displays. Other sensory modalities, e.g., olfactory displays, are rarely used today. User characteristics work together with the technical design and application specific capabilities to define the system. If there is a mismatch between these three factors, the performance of the human-computer-system is degraded and negative side-effects like ocular strain, headaches, or simulator sickness could occur.

On a higher conceptual level, the combination of user involvement, system characteristics, and task demands can create a feeling of immersion or of being physically present in the computer-generated scenario. This complex phenomenon called "presence" has been intensively researched, but it still not well understood. There are disagreements over how one defines and measures presence. Moreover, there are disagreements over its benefits in the context of achieving the AMVE application's goals. For this reason it should be investigated.

Another complex factor that can affect the use of AMVE technology is workload. There are many different methods for measuring human operator workload. The methods vary from subjective rating scales to psychophysical measures, e.g., heart rate. Which method to choose strongly depends on the application and tasks that must be performed. Based on the review included in this report psychophysiological measures are not recommended for applied research because of the lack of consistent findings regarding these measures and performance of cognitive tasks. As a general practice, a global, univariate workload measure such as the NASA TLX, is suggested as well as contextually relevant primary and embedded secondary task measures.

Like presence, simulator sickness is the outcome of a complex interaction among user, system, and task characteristics. Simulator sickness is a form of motion sickness that does not require true motion, but often involves a wide field-of-view display. There are two major theories to explain simulator sickness: they are the sensory conflict theory and the postural instability theory. The sensory conflict theory states that there is conflicting information from sensory inputs from vision, semicircular canals, otoliths, proprioceptors, and somatosensors which causes simulator sickness. The postural instability theory notes that sickness-producing situations are characterized by their unfamiliarity to the participant. Regardless of which theory is supported simulator sickness can reduce performance and usability of AMVE technologies. The section on simulator sickness includes recommendations for reducing its effects.

There need to be valid measures to estimate the effect of the technology on performance. Relevant measures identified are discussed in the sections on situational awareness, collective performance measures, and after action review methods.

Attaining situational awareness has become vitally important for mission success. Understanding who are combatants, civilians, and allied personnel, as well as, knowing the rules of engagement for the given situation, are all part of soldiers' situation awareness. AMVE technologies can represent the complex nature of the battlefield and can be used to provide the training needed to acquire situational awareness. There have to be appropriate measures to estimate and quantify the effect. The group has summarized the main measures and approaches in this field.

For sometime, the measurement of operator performance has been an important aspect of research for those investigating human system interaction and the use of VE. Developing methods and metrics to measure performance have contributed to assessing the utility of AMVE technologies and to predict how well training in simulation will transfer to the real world performance. A variety of subjective and objective measures of individual and team performance measures were identified and summarized in this report.

An important method for providing performance feedback is the after action review (AAR). Simulation-based training is particularly suited to AAR because ground truth can be easily displayed. AAR is an important part of the collective training process. It is an active process that requires unit members to participate in order to benefit. After-action review is a method of providing feedback to units after operational missions or collective training exercises. Realistic simulations provide participants with intrinsic feedback that cues and guides their performance and, to some extent, let them know how well they are performing various tasks. The intrinsic feedback received by individuals depends upon their job, their location in the battlespace, and the quality of the simulation environment. AAR systems have been developed to collect network data and provide units with information on what happened and allows them to discuss and discover why it happened and what they should do about it the next time the situation arises.

The main benefits of AMVE technologies were found to be their relatively low cost and the large number of possible applications they could serve. Compared to performing in the real-world environments, VE can be cost effective. Computer-generated scenarios can be cheaper to generate than real-world ones, and commercial off-the-shelf technologies can now provide effective interfaces to synthetic environments, sometimes providing a richer set of sensory cues than prior or conventional simulator technologies. Often, the same AMVE hardware can be used for several purposes since it is less dependent on physical representations of the human interface than conventional approaches for simulation. Given the flexibility of the AMVE hardware and software, the same system can be used to model and train a variety of tasks if they have common requirements for the human interface.

The effectiveness of AMVE-technologies depends on the task to which they are applied and its human factors. Although VE is ready for some applications or part task training, research is still needed in other areas. Training of tasks that rely on manual dexterity is not possible due to missing haptic feedback. There are still no realistic and practical means of providing locomotion in a VE. Representation of comprehensive environmental stimuli does not include “mud and dirt.” AMVE-technologies can contribute to effective training strategies, but they will not totally replace other education and training methods. This is particularly true for live training. Even though technology is capable of simulating many aspects of the real world, there are limitations. Soldiers will always need to experience the physical demands of the real world. Simulation allows soldiers to get the most out of real world training since it has the potential to allow them to experience many of the situations they will encounter in theatre.

By taking into consideration the human factors issues and variables described in this report, the resulting simulations and training applications would be more usable and effective. Effectiveness would be determined by testing the system with a man in the loop and use of a variety of measures. In addition to simple temporal (time to complete mission) and error messages (mission goals met) more sophisticated measures of situational awareness, team and collective performance measures should be applied.

1.0 FUTURE CONSIDERATION

Military education and training have been identified as a main application area for AMVE-technologies. A number of factors are influencing training policies, procedures and technologies. An important factor is the need for units to deploy with little or no notice. When deployed they do not have the facilities and infrastructure needed to acquire needed skills, to plan and rehearse complex missions. Recent advances in computer and display technologies strongly miniaturize them while increasing their performance and functionality, make embedding training and mission rehearsal capabilities in highly mobile military hardware both practical and effective.

Embedded training is a concept which integrates training functionality into operational equipment. It allows military personnel to train and rehearse while deployed to an operational area. Embedding training allows skills to be developed and maintained close to the battlefield. To date, embedded training has been successfully applied by NATO armed forces primarily to large computer controlled systems such as air defence systems, and naval vessels. The recent increases in power and capability and decrease in size of virtual environment (VE) and augmented reality (AR) technologies allows virtual simulation to be embedded in smaller ground and air systems. There are also potential applications for training dismounted soldiers. By integrating network-enabled capabilities, collective as well as individual training (for war fighting, peace keeping and maintenance skills) may also be possible.

Human-centred design and integration of embedded virtual simulation technologies requires a thorough review and analysis on conceptual and technical levels. It covers a broad spectrum of Human-System Integration, which spans from novel pedagogical concepts to innovative techniques of human-computer interaction. There is a growing need for research that explores use of virtual simulation in embedded training systems. As indicated by the results of the present RTG, AMVE technology has the potential to also enhance the effectiveness of embedded training and rehearsal systems if both simulation and human performance are addressed.

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	**2. Originator's References**	**3. Further Reference**	**4. Security Classification of Document**
	RTO-TR-HFM-121-Part-II AC/323(HFM-121)TP/62	ISBN 978-92-873-0077-7	UNCLASSIFIED/ UNLIMITED

5. Originator

Research and Technology Organisation
North Atlantic Treaty Organisation
BP 25, F-92201 Neuilly-sur-Seine Cedex, France

6. Title

Virtual Environments for Intuitive Human-System Interaction – Human Factors Considerations in the Design, Use, and Evaluation of AMVE-Technology

7. Presented at/Sponsored by

Final Report of Task Group TR-HFM-121.

8. Author(s)/Editor(s)	**9. Date**
Multiple	July 2007
10. Author's/Editor's Address	**11. Pages**
Multiple	178

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

AMVE (Augmented, Mixed and Virtual Environments)
Augmented reality
Computer graphics
Computerized simulation
Display devices
Human computer interface
Human factors engineering
Psychometrics
Simulation
Simulator sickness
Virtual environments
Virtual reality
Visual perception
Workload

14. Abstract

Changing missions and development of new virtual and augmented reality technologies require the alliance to investigate and integrate new concepts for preparing their militaries and for evaluating human-system integration (HSI) for the design of new equipment. New paradigms and media of Augmented, Mixed, and Virtual Environments (AMVE) provide a realistic training environment and natural HSI using complex realistic or abstract synthetic environments. There are still many open research issues which limit practical application of these technologies. This report addresses relevant human factors issues that must be considered in the design, operation and performance assessment of AMVE-systems. The report discusses how AMVE technologies have been applied to solve real world military problems. While AMVE technologies have proved useful they still need further development before they can become a truly intuitive technology.

NORTH ATLANTIC TREATY ORGANISATION

RESEARCH AND TECHNOLOGY ORGANISATION

BP 25
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

Les publications de l'AGARD et de la RTO peuvent parfois être obtenues 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 soit à titre personnel, soit au nom de votre organisation, sur la liste d'envoi.

Les publications de la RTO et de l'AGARD sont également en vente auprès des agences de vente indiquées ci-dessous.

Les demandes de documents RTO ou AGARD doivent comporter la dénomination « RTO » ou « AGARD » selon le cas, suivi du numéro de série. Des informations analogues, telles que le titre est la date de publication sont souhaitables.

Si vous souhaitez recevoir une notification électronique de la disponibilité des rapports de la RTO au fur et à mesure de leur publication, vous pouvez consulter notre site Web (www.rto.nato.int) et vous abonner à ce service.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE
Streitkräfteamt / Abteilung III
Fachinformationszentrum der Bundeswehr (FIZBw)
Gorch-Fock-Straße 7, D-53229 Bonn

BELGIQUE
Royal High Institute for Defence – KHID/IRSD/RHID
Management of Scientific & Technological Research for Defence, National RTO Coordinator
Royal Military Academy – Campus Renaissance
Renaissancelaan 30
1000 Bruxelles

CANADA
DSIGRD2 – Bibliothécaire des ressources du savoir
R et D pour la défense Canada
Ministère de la Défense nationale
305, rue Rideau, 9e étage
Ottawa, Ontario K1A 0K2

DANEMARK
Danish Acquisition and Logistics Organization (DALO)
Lautrupbjerg 1-5
2750 Ballerup

ESPAGNE
SDG TECEN / DGAM
C/ Arturo Soria 289
Madrid 28033

ETATS-UNIS
NASA Center for AeroSpace Information (CASI)
7115 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)
Defence Industry & Research General Directorate, Research Directorate
Fakinos Base Camp, S.T.G. 1020
Holargos, Athens

HONGRIE
Department for Scientific Analysis
Institute of Military Technology
Ministry of Defence
P O Box 26
H-1525 Budapest

ISLANDE
Director of Aviation
c/o Flugrad
Reykjavik

ITALIE
General Secretariat of Defence and National Armaments Directorate
5th Department – Technological Research
Via XX Settembre 123
00187 Roma

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25
NO-2007 Kjeller

PAYS-BAS
Royal Netherlands Military Academy Library
P.O. Box 90.002
4800 PA Breda

POLOGNE
Centralny Ośrodek Naukowej Informacji Wojskowej
Al. Jerozolimskie 97
00-909 Warszawa

PORTUGAL
Estado Maior da Força Aérea
SDFA – Centro de Documentação
Alfragide
P-2720 Amadora

REPUBLIQUE TCHEQUE
LOM PRAHA s. p.
o. z. VTÚLaPVO
Mladoboleslavská 944
PO Box 18
197 21 Praha 9

ROUMANIE
Romanian National Distribution Centre
Armaments Department
9-11, Drumul Taberei Street
Sector 6
061353, Bucharest

ROYAUME-UNI
Dstl Knowledge Services
Information Centre
Building 247
Dstl Porton Down
Salisbury
Wiltshire SP4 0JQ

TURQUIE
Milli Savunma Bakanlığı (MSB)
ARGE ve Teknoloji Dairesi Başkanlığı
06650 Bakanliklar
Ankara

AGENCES DE VENTE

NASA Center for AeroSpace Information (CASI)
7115 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 Acquisitions
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)

NORTH ATLANTIC TREATY ORGANISATION

RESEARCH AND TECHNOLOGY ORGANISATION

BP 25
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

DISTRIBUTION OF UNCLASSIFIED RTO PUBLICATIONS

AGARD & RTO publications are sometimes available from the National Distribution Centres listed below. If you wish to receive all RTO reports, 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 reports may also be purchased from the Sales Agencies listed below.

Requests for RTO or AGARD documents should include the word 'RTO' or 'AGARD', as appropriate, followed by the serial number. Collateral information such as title and publication date is desirable.

If you wish to receive electronic notification of RTO reports as they are published, please visit our website (www.rto.nato.int) from where you can register for this service.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Royal High Institute for Defence – KHID/IRSD/RHID
Management of Scientific & Technological Research for Defence, National RTO Coordinator
Royal Military Academy – Campus Renaissance
Renaissancelaan 30
1000 Brussels

CANADA
DRDKIM2
Knowledge Resources Librarian
Defence R&D Canada
Department of National Defence
305 Rideau Street, 9th Floor
Ottawa, Ontario K1A 0K2

CZECH REPUBLIC
LOM PRAHA s. p.
o. z. VTÚLaPVO
Mladoboleslavská 944
PO Box 18
197 21 Praha 9

DENMARK
Danish Acquisition and Logistics Organization (DALO)
Lautrupbjerg 1-5
2750 Ballerup

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)
Gorch-Fock-Straße 7
D-53229 Bonn

GREECE (Point of Contact)
Defence Industry & Research General Directorate
Research Directorate
Fakinos Base Camp
S.T.G. 1020
Holargos, Athens

HUNGARY
Department for Scientific Analysis
Institute of Military Technology
Ministry of Defence
P O Box 26
H-1525 Budapest

ICELAND
Director of Aviation
c/o Flugrad, Reykjavik

ITALY
General Secretariat of Defence and National Armaments Directorate
5th Department – Technological Research
Via XX Settembre 123
00187 Roma

LUXEMBOURG
See Belgium

NETHERLANDS
Royal Netherlands Military Academy Library
P.O. Box 90.002
4800 PA Breda

NORWAY
Norwegian Defence Research Establishment, Attn: Biblioteket
P.O. Box 25
NO-2007 Kjeller

POLAND
Centralny Ośrodek Naukowej Informacji Wojskowej
Al. Jerozolimskie 97
00-909 Warszawa

PORTUGAL
Estado Maior da Força Aérea
SDFA – Centro de Documentação
Alfragide
P-2720 Amadora

ROMANIA
Romanian National Distribution Centre
Armaments Department
9-11, Drumul Taberei Street
Sector 6, 061353, Bucharest

SPAIN
SDG TECEN / DGAM
C/ Arturo Soria 289
Madrid 28033

TURKEY
Milli Savunma Bakanlığı (MSB)
ARGE ve Teknoloji Dairesi Başkanlığı
06650 Bakanliklar – Ankara

UNITED KINGDOM
Dstl Knowledge Services
Information Centre
Building 247, Dstl Porton Down
Salisbury, Wiltshire SP4 0JQ

UNITED STATES
NASA Center for AeroSpace Information (CASI)
7115 Standard Drive
Hanover, MD 21076-1320

SALES AGENCIES

NASA Center for AeroSpace Information (CASI)
7115 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 Acquisitions
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 2216
UNITED STATES
(also available online in the NTIS Bibliographic Database or on CD-ROM)

ISBN 978-92-873-0077-7