REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
April 2004

3. REPORT TYPE AND DATES COVERED
Final Proceedings (22 Mar 2004 – 25 Mar 2004)

4. TITLE AND SUBTITLE
HPSAA II (Human Performance, Situation Awareness and Automation) Technology Conference
Vol. 2

5. FUNDING NUMBERS
W81XWH-04-C-0080

6. AUTHOR(S)
Dennis A. Vincenzi, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Embry-Riddle Aeronautical University
Daytona Beach, Florida 32114

E-Mail:

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Report published as Volume I and Volume 2

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

None provided

20040909 014

14. SUBJECT TERMS
None provided

15. NUMBER OF PAGES
637

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

BEST AVAILABLE COPY

AD________________

Award Number: W81XWH-04-C-0080

TITLE: HPSAA II (Human Performance, Situation Awareness and Automation) Technology Conference

PRINCIPAL INVESTIGATOR: Dennis A. Vincenzi, Ph.D.

CONTRACTING ORGANIZATION: Embry-Riddle Aeronautical University
Daytona Beach, Florida 32114

REPORT DATE: April 2004

TYPE OF REPORT: Final Proceedings

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

HPSAA II
Volume II

HUMAN PERFORMANCE, SITUATION AWARENESS AND AUTOMATION

Current Research and Trends

Edited by
Dennis A. Vincenzi • Mustapha Mouloua
Peter A. Hancock

Human Performance, Situation Awareness and Automation: Current Research and Trends

Edited by:

Dennis A. Vincenzi, Ph.D.
Department of Human Factors and Systems
Embry-Riddle Aeronautical University
600 S. Clyde Morris Blvd.
Daytona Beach, FL 32114
dennis.vincenzi@erau.edu

Mustapha Mouloua, Ph.D.
Department of Psychology
University of Central Florida
4000 Central Florida Blvd
Orlando, FL 32816
mouloua@pegasus.cc.ucf.edu

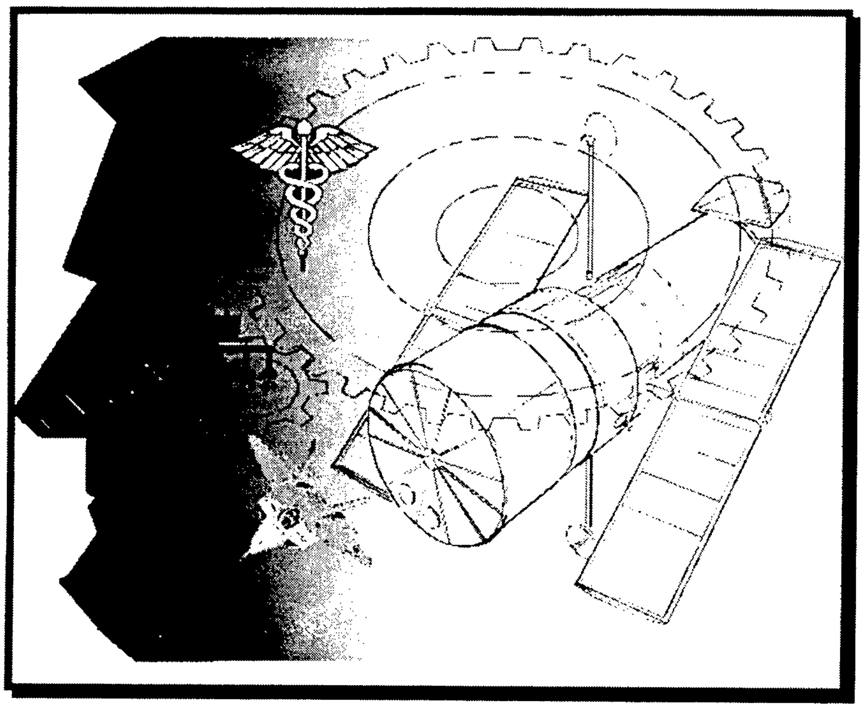
Peter A. Hancock, Ph.D.
Department of Psychology
University of Central Florida
4000 Central Florida Blvd
Orlando, FL 32816
phancock@pegasus.cc.ucf.edu

Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, NJ 07430

ISBN 0-8058-5341-3

HUMAN PERFORMANCE, SITUATION AWARENESS AND AUTOMATION:
CURRENT RESEARCH AND TRENDS

HPSAA II
Volume II



Edited by:

Dennis A. Vincenzi
Mustapha Mouloua
Peter A. Hancock

Proceedings of the Second Human Performance, Situation Awareness and Automation Conference (HPSAA II), held in Daytona Beach, FL, March 22 – 25, 2004.

Conference Chairs:

Dennis A. Vincenzi, Conference Chair, Embry-Riddle Aeronautical University
Mustapha Mouloua, Conference Co-Chair, University of Central Florida

Scientific Program Committee:

Prof. Jean Bresson, *Ecole Nationale De L'Aviation Civile (ENAC)*
John E. Deaton, *Florida Institute of Technology*
Mica R. Endsley, *SA Technologies*
Peter A. Hancock, *University of Central Florida*
David Kaber, *North Carolina State University*
Anthony Majoros, *The Boeing Company*
Prof. Regis Mollard, *Université René Descartes-Paris*
Raja Parasuraman, *The Catholic University of America*
Mark Scerbo, *Old Dominion University*
John Wise, *Honeywell International*

COPYRIGHT

The present volume entitled "*Human Performance, Situation Awareness and Automation: Current Research and Trends*" is intended for limited distribution to the conference attendees. Permission to copy without fee is granted only for non-commercial use. The authors retain all rights to their papers. To copy or republish these articles for any other reason requires specific permission from the authors.

Technical Editing and Graphic Design by:

Dennis A. Vincenzi
Ryan Wasson

Department of Human Factors and Systems
Embry-Riddle Aeronautical University
600 S. Clyde Morris Blvd.
Daytona Beach, FL 32114

TABLE OF CONTENTS

TECHNICAL PAPERS

AUTOMATION AND HUMAN PERFORMANCE

UNMANNED AERIAL VEHICLES AND INTELLIGENT SYSTEMS

MILITARY APPLICATIONS

AUTOMATION AND HUMAN PERFORMANCE

HUMAN REQUIREMENTS IN AUTOMATED WEAPONS SYSTEMS

Jennifer McGovern Narkevicius, Ph.D.
ARINC Engineering Services, LLC

Peggy L. Heffner
Naval Air Systems Command Headquarters

ABSTRACT

Automation is a necessary addition to current and future weapons systems. Although automation is necessary to achieve these goals, its requirements cannot stand alone. Successful acquisition programs require complete requirements definition. Traditionally these requirements are limited to hardware and software elements, failing to account for the human operators and maintainers. Automation technologies have the potential to improve system performance, reduce human error, improve decision making, and highlight situational awareness of both the immediate user and the greater command and control structure. Automation technology will improve decision-making and situational awareness throughout the distributed hierarchy of command and control in a networked battle force.

We will discuss methods for collecting, defining and illustrating human performance requirements, the utility of collecting and integrating human systems requirements into successful systems engineering processes to produce usable and useful automated systems in future weapons systems, recent concept exploration successes, lessons learned and suggestions for future directions.

Keywords: Automation; User Requirements Definition; Distributed Systems; Situational Awareness

INTRODUCTION

Automated systems are a necessary addition, designed more and more frequently into weapons systems. Automation provides the potential for improving human performance by reducing errors and enhancing decision-making and situational awareness. Aviation systems clearly are an important technical area for automated systems [3], but automation is not the sole province of weapons systems. However, these systems are, of necessity, complex and are developed in adherence with the systems engineering principles.

Systems engineering follows a fairly rigorous, detailed and documented process. This process ensures that the concerns of all the appropriate and applicable disciplines are considered in the design trade-offs made throughout the development of complex systems. The systems engineering phases, illustrated in Figure 1, provide checks and balances for decision making throughout the design. Exit criteria help decision makers assess programmatic risk (cost, schedule and performance) associated with proceeding in the selected development path. The process also allows opportunities to inject improvements or design changes based on intelligent flexibility in the design trade space.

Successful systems engineering acquisition programs are built on accurate and complete requirements definition. Automated systems require the same careful requirements definition necessary to all complex systems. However, integration of humans into complex automated systems continues to be an issue. The potential benefits of automation are countered by the very real costs of the increased design complexity that is required to accommodate the automated system and the increased potential for human error through operation of an improperly designed automated system. In addition, automated systems are embedded in the increasingly complex structures of distributed decision networks.

Additionally, requirements definition is essential both for successful manpower and personnel acquisition as well as for the necessary and sufficient training required to provide appropriate human performance to mission systems. Traditionally in systems acquisition programs these requirements are limited to the hardware and software elements. These limits fail to account for the requirements the human operators and maintainers bring with them as part of the mission system. However, strategy, tactics, techniques, procedures and accountability all require positive control of the mission system by a human user. It is essential, therefore, that the human user's requirements, and the human maintainer's requirements, as well as the requirements for each hardware and software subsystem, be included in the baseline assumptions for system requirements definition.

Distributed command and control (C^2) systems also require the detailed requirements definition that is warranted by complex systems [7]. There are automated systems embedded in the C^2 systems that take in information from distant automated systems. The integration rules must be specified carefully for each node and each level of the network, balancing the ability to gather data with the ability to use the information. These nested composite systems inside complex systems provide opportunity for requirements to be overlooked or incorrectly captured. This is especially true for the human performance requirements that will be similar in appearance but different in function across systems. To support this requirements definition, there will need to be more research in information processing, social cognition with variable delays, specifically for time critical tasks such as in warfare. These requirements will specify systems and network architectures that appropriately support situational awareness, decision making and reduction of errors throughout the network.

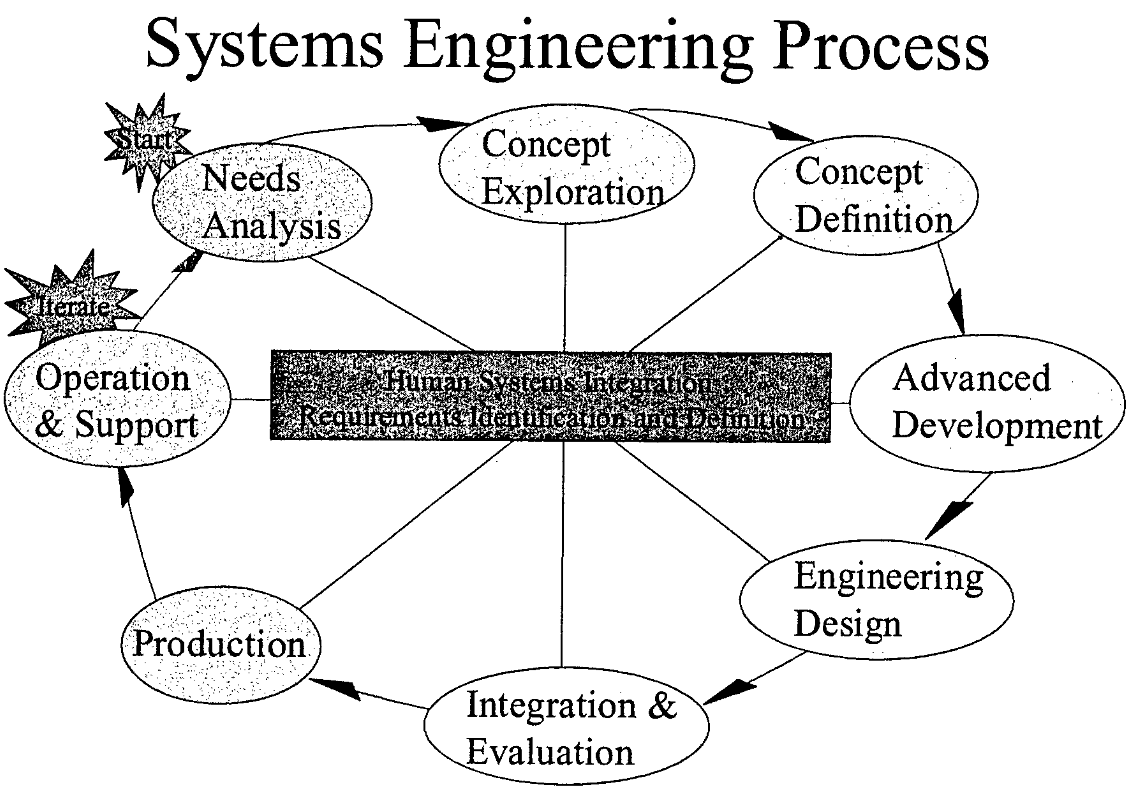


Figure 1. Iterative Phases of the Systems Engineering Process including the HSI elements

"Everything we invent or make is ultimately designed for human use" [2]. To make requirements definition relevant to systems under development for use by human users, the requirements must be documented and utilized.

Transformation and technological developments are allowing many weapons systems to be networked together. These networked systems have the potential to generate new capabilities and new possibilities. The requirements of these networked systems are not the summation of the requirements of the original component systems. Rather, there will be that summation as well as an amalgamation of requirements (and their derivatives) to be defined, designed to, explored through concepts of operations and analyses of alternatives, and met with design decisions.

APPROACH

Successful implementation of appropriate automation in appropriate locations in systems and subsystems will have a positive, force multiplication effect in mission performance. However, successful implementation of this appropriate automation is not simply a software requirement but hinges directly and indirectly on the identification and definition, in operational terminology, of the requirements of the human operators and maintainers. The systems engineering process provides the placeholders for successful human engineering programs and provides a framework for utilization of human systems integration processes and tools.

There are a number of tools available to assist technical professional in developing and implementing the human users' requirements into overall system design. Particularly useful toolsets include modeling and requirements management.

Req

Requirements definition begins with recognizing an operational need [2] or needs. These needs must include those of the users. It is essential to consider not only the immediate needs driving the system under development or modification, but also to consider the application and use of the system with respect to other systems with which it must interact. This is even more important in networked systems that must work together, preferably seamlessly, to achieve a greater capability than the sum of the individual systems' capabilities. As a discipline, Systems Engineering provides a framework within which to approach this requirements definition of the system under development. It also provides the framework within which the more global system can be considered.

Definitions of needs and of requirements are essential in any systems engineering acquisition program [4, 6]. The need illustrates the desired capabilities, accomplishments or achievements. The required performance of the system comes from achieving these desires These requirements must be identified to determine what possible solutions to bring forward in an effort to meet those requirements. Requirements for weapons systems are easily documented for hardware and software but the determination and application of requirements for users is more difficult. Tools, processes and procedures are necessary to apply to users in engineering acquisitions [1].

Because performance of a system depends on the operator as well as the hardware and software [2], it is necessary to translate from the requirements of the overall system to useful, successful human performance in support of that system completing that mission. The primary tools for successful integration of human requirements into systems acquisition and engineering include models, use cases, and requirements management. These tools are necessary to integration human user requirements and their concomitant

Models and modeling

While the requirements detail *what* a system must be able to do to be considered successful, good requirements do not dictate *how* a system must work or operate. It is quite difficult to get from the what of the requirements to the how of design. One useful tool is modeling of potential solutions to the requirements. Modeling can provide a means for asking and answering questions about functional allocation and tasks assignment across the three major elements of the system: hardware, software, and human users. Modeling requires a good understanding of the mission requirements and the means to allocate those requirements within possible solutions. Models must be valid, verifiable, and accurate [3].

Modeling tools provide an economical means of exploring solutions in the trade space without negative effects on cost, schedule, or performance. These tools also provide the means to generate a large pool of potential solutions. Then candidate solutions can be further evaluated and final solutions chosen more freely from the available options rather than selecting, in effect, technical "variations on a theme".

It is feasible (and necessary) to model the automated system and to allocate functions to the automation software, the hardware, and to the human user. Modeling also provides a platform to quickly reallocate functions and observe the effect of different allocations on overall system performance. Models can also be developed from networked distributed systems (such as C^2 entities). Again, it is possible to alter the allocation of functions across the distributed network and determine the optimized way to work within the network.

It is equally necessary to model the elements and entities of distributed C2 systems. The interactions of the component systems within the C2 system can be modeled and functions can be allocated to those entities to observe the effects of different allocations on the behavior and success of the network. Distributed systems also require modeling. These models must incorporate the element systems and the distribution or network to fully explore the trade space. But more importantly, modeling distributed systems more fully illustrates unintended consequences (both beneficial and unbeneficial). Modeling may also reveal potential, unanticipated enhancements that are an outgrowth of the distribution of systems and their integration.

Use Cases help support human performance modeling by limiting the possible options to be modeled to an operationally appropriate set. Use cases describe what the system under development must do to achieve the mission from the users' perspective. This focus is at the high level of the system. Use Cases focus on the user as the definition of the scope the project. They can be used to scope the models developed (see above) to ensure that how the user will use the system is included in decision making. Because of the focus on the users' perspective of the functions of the system, the Use Case maintains focus throughout development.

Use Cases are at a low enough level of granularity that they can be used to describe a weapons system and to describe the networked C^2 system in which that weapons system must operate. The use case will facilitate the development of information flow across the C^2 platform and will highlight nodes of information glut that will reduce the performance of the C^2 system and the performance of the weapons system associated with the network.

Use cases should be developed to help select portions of the operational space to be more fully explored in modeling. They provide a consistent set of scenarios to explore throughout development and operation.

Requirements Management

Requirements management tools allow designers and others associated with the development of systems under design to ensure that all identified requirements (hardware, software, and user) are documented and are traceable throughout development. These tools ensure that requirements that are difficult to allocate are not dropped. These tools keep all the requirements on equal footing ensuring that user requirements are not deleted in the face of technical challenges. This is especially essential in automated systems where user requirements make demands that may be difficult to sort out in software architecture development.

DISCUSSION

The US Navy has a renewed interest, driven from the top, in making the sailor the center of the Navy. This will strengthen war-fighting capabilities by including the user of weapons rather than focusing solely on the physics of the weapons themselves. The Human System Integration (HSI) thrust has pushed the user requirements to the forefront. This focus on users of equipment, rather than on the equipment itself, requires a shift in the processes used to acquire warfighting equipment. These changes in focus will include moving to the *integration* of humans as integral parts of the warfighting system rather than the insertion of humans, as has historically been the approach.

This focus on the sailor will require an integration of tools from across disciplines. These disciplines are diverse and include a number of sub-disciplines. Tools come from Manpower, Personnel, Training, Human Factors, Safety, and Health as well as the other elemental disciplines in HSI in addition to tools from more traditional disciplines of hardware engineering, software engineering and systems engineering.

The US Navy continues initiatives to compile and integrate processes, tools and techniques from these various human centered disciplines. These activities work to identify, validate, verify, and integrate the tools and their outputs from different disciplines. This effort will ensure that the information and data applicable to design and exploration of the trade space are useful.

In the E/A-18G electronic attach variant program, the outcome of the HSI approach has directly affected the development of this highly automated system. While in development the E/A program has included a strong reliance on modeling and simulation, use cases, and requirements management. This highly automated, networked system will allow support of distant conflicts with precision, speed, and accuracy (the need for this is highlighted in [5]). Its careful systems engineering approach will allow continued development of systems improvements throughout the lifecycle of the weapons system.

The US Navy continues to explore HSI toolsets and integration of those toolsets. These toolsets will allow successful inclusion of HSI (and its elements' technical requirements considerations) in systems engineering acquisition. This will enhance the use and utility of HSI tools throughout the process.

Highly networked systems will have nested sets of user requirements based on the capabilities of the system. Early use of the tools in the systems engineering acquisition process and follow through with requirements management tools will allow the nested requirements to be incorporated into systems designed to improve situational awareness, decision making, networked work, and reduced error throughout the system.

The human is slow to evolve but the systems around the human can be designed to support decision making, situational awareness, reduced designed induced error, and increased operational effectiveness. The costs associated with these HSI improvements is low, especially if introduced early in the program and carried throughout the acquisition process.

The continued development of highly complex, automated, networked systems will placing increasing demands for modeling, use case, and requirements management in the systems engineering of weapons systems. The E/A-18G program is an excellent example of how this is coming together with the Navy's HSI process. As more complex, networked systems are developed; this approach will become more systematized.

REFERENCES

Booher, H. R. (2003). Handbook of Human Systems Integration. John Wiley & Sons, Inc.

Chapanis, A. (1996). Human Factors in Systems Engineering. New York: John Wiley & Sons, Inc.

Kanki, B. G. (2001). Automation in the Workplace: Lessons Learned in Aviation Operations. Human Factors and Ergonomics Society Computer Science Technical Group Bulletin, *8 (1)*.

Martin, J. N. (1997). Systems Engineering Guidebook. Boca Raton: CRC Press.

Pettre, M. (2003). Close Air Support from Afar. Journal of Electronic Defense, 26 *(6)*, 40 -45.

Rechtin, E. & Maier, M. W. (1997). The Art of Systems Architecting. Boca Raton: CRC Press.

Sheridan, T. B. (2002). Humans and Automation: System Design and Research Issues. John Wiley & Sons, Inc.

THE ADAPTIVE OPERATOR

Hans (J.A.) Veltman, Chris Jansen
TNO Human Factors

ABSTRACT

This paper presents a framework for mental workload that describes the adaptive nature of human beings in interacting with the environment. The framework is a result of many years of mental workload research in different complex task situations. This framework can be used to understand the role of mental workload in complex task situations as well as the dissociation of outcomes of different workload measures that is often observed. These issues are important for effective implementation of adaptive automation. Furthermore, the framework can be of value in discussions about the role of operator state assessment in adaptive automation.

Keywords: mental workload; adaptive automation; operator state

INTRODUCTION

Adaptive automation (AA) is a concept in which dynamic changes occur in the allocation of functions between humans and machines. This allocation can be based on different sources of information (Parasuraman, 2003): (1) critical events, in which certain salient environmental events trigger automation; (2) operator performance; (3) operator state assessment; (4) task and cognitive models and (5) hybrid methods, in which a combination of sources is used. The aim of AA is to improve overall task performance. Several studies indicate that information about the state of the operator is crucial for a functional AA system (e.g. Scerbo, Freeman, & Mikulka, 2000). It is often argued that a system should take over control when the mental workload of an operator becomes unacceptably high. However, this approach faces at least two challenges. First, despite the large amount of publications on measures for operator state assessment, the ultimate measure or set of measures is still not agreed upon. Second, it is not clear how to use the information about operator state effectively. Operators normally adapt to the changing task requirements by regulating their effort expenditure. Many workload measures that are used to detect high workload are often also indicators of a successful adaptation process of the operator. Task reallocation from the operator to the system based on such measures may confuse the operator and will therefore not improve the overall performance. Therefore, we believe that this adaptive behavior of the operator should be taken into account for successful implementation of AA.

We conducted several mental workload experiments in complex task situations such as in cockpits and control rooms of frigates. Different kinds of workload metrics were used in these studies, such as performance, subjective and physiological measures. These measures all provided different information about mental workload. Based on the results, we developed a framework to describe the complex relation between the changing task requirements and the adaptive behavior of the operator. It also provides more insight into the different aspects of workload that are captured by the different workload metrics.

Workload framework

The framework (see Fig. 1) is based on perceptual control theory (PCT; Powers, 1973) that is also used in models of Hockey (2003) and a model of Hendy, East and Farrel (2001). The model of Hockey uses PCT to describe state regulation, whereas Hendy et al. use the PCT to describe information processing. The present framework is a combination of these models. The PCT assumes that the difference between a required situation (goal) and actual situation (sensor information) is crucial for the adaptive behavior of biological systems. Adaptive changes will occur when such differences (error signals) exist. Goals can be defined at several levels and an error signal is often a new goal for a lower order system.

The framework in Fig. 1 includes two levels: *task goals* at the highest levels and *required state* at a lower level. More levels can be included; for example the difference between the required and the actual state can be described as the required blood pressure (a goal for the cardiovascular control system).

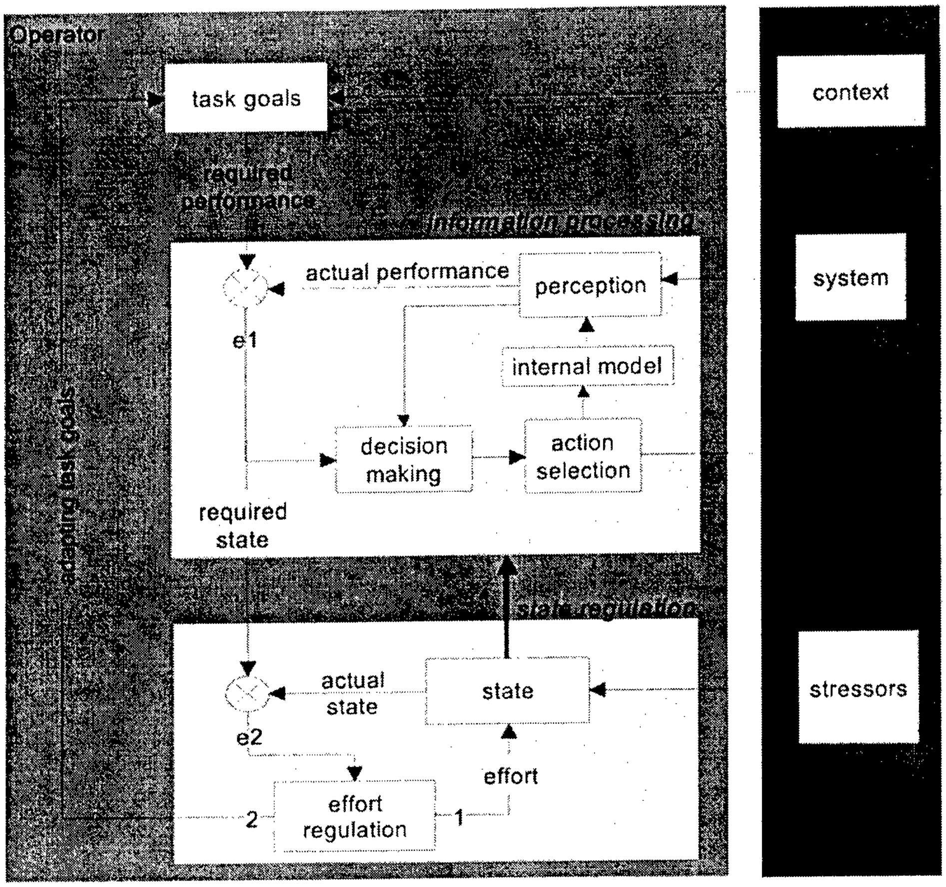


Fig. 1 framework for operator state assessment (see text for explanation)

The framework includes an *information processing* loop and a *state regulation* loop. The state is crucial for the information processing. This is often neglected in information processing models. It is well known that it is difficult to perform a cognitive demanding task when we are in a sub-optimal state, for example due to sleep loss or fatigue. The information-processing loop includes the stages of information processing of an operator dealing with a system (perception, decision making and action selection). Information to be processed can come from the environment (system) or from an internal model of the system that is built up by the operator. The perceived information, and in particular, the perceived actual performance is compared with the required performance (task goals). The intensity of the information-processing loop is adjusted depending on the difference between the required and perceived actual performance. For example, if the perceived actual performance is poor, but the operator does not have the intention to perform well, there will be no error signal (e1) and as a consequence the intensity of the information processing will not change. On the other hand, if the performance is good, but the operator has the intention to perform perfect, there will be an error signal.

If the error signal (e1) persists, the required state needs to be adjusted. If this does not match with the actual state then another error signal (e2) will increase. There are two main processes available to reduce e2. The most direct one is investing more mental effort to adjust the actual state to the required state. This process can be observed by physiological changes such as an increase in blood pressure and heart rate and a decrease in heart rate variability (Veltman & Gaillard, 1998). However, there are costs involved in effort investment. Operators will become fatigued and as a consequence they will feel resistance for further effort investment. An indirect way to reduce e2 is to change the task goals. For example, operators will slow down the task execution, will skip less relevant tasks or accept good instead of perfect performance. In this way, they reduce the intensity of the information processing and hence, the required state.

The framework assumes that there is no direct relation between information load and physiological measures that are used as 'state' estimators. Making a task more difficult will not automatically result in changes in physiological reactions. This is because an increase in information load may also result in setting lower task goals instead of putting in more effort. For example, the operator can take more time to perform the task, skip some tasks, or will be satisfied with more errors.

Effects of context

The likelihood of adapting the task goals is affected by the context. For example, in a flight simulator, reducing the task goals often does not have serious consequences. In a real aircraft this can have serious consequences and therefore, the effort investment is often much higher in a real aircraft (e.g. Wilson et al., 1987). However, when the context of the flight simulator is a selection to become a pilot, then the mental effort, measured with physiological measures, is the same as in a real aircraft (Veltman, 2002).

Another example of the effect of context on task goals is the existence of other goals. In many situations, the task goals are just one set of goals among many other goals such as keeping rest, going to a toilet, have a conversation, going away for a cigarette etc. The context is important for keeping the task goals the primary one. During vigilance for example, performance will often deteriorate after some time because it is difficult to keep the task goal the primary goal among other competing goals as getting rest or countering boredom.

Effects of stressors

External stressors such as G-load, noise, vibration and extreme temperatures are assumed to affect the state of the operator. External stressors disrupt state regulation, making the operator less able to adapt to changing task demands. The same mechanisms as describes above, can compensate for a reduced state. The operator can invest additional effort, or he can change the task goals.

Because stressors do have an effect on the state of the operator, they are important for the interpretation of physiological workload measures. Physiological workload measures that seem to work well in laboratory situations are often difficult to use in applied situations because of the many stressors that operators have to deal with.

Applying the framework: some examples

Level of information processing: novice versus expert operators

Information can be processed at different levels. Rasmussen (1986) described three levels: skill-based, rule based and knowledge based. When the operator is well trained, he can process most information at the skill-based level, which does not require much attention and effort. An increase in information will hardly affect the intensity of the information processing and no change in operator state is required. However, the same information can result in knowledge-based processing for a novice operator. Increasing the amount of information will then result in a more intensive information processing and an increase in mental effort, as is reflected in the physiological state of the operator.

Effects of an incorrect mental model

When there is a discrepancy between the information from the system and the mental model, the perceived performance is strongly affected. The increased error signal (e1) results in a considerable increase in the intensity of the information processing (and the 'required state'). In a study on mental workload during helicopter missions, Veltman and Gaillard (1999) found that this factor was more important for the effort investment than the total amount of information presented to the crew.

Differences between physiological and subjective effort measures

The framework provides insight into differences between subjective and physiological workload measures. It often happens that subjective effort measures such as the Rating Scale Mental Effort (RSME; Zijlstra, 1993) or the effort sub-scale of the TLX (Hart & Staveland, 1988) show differences between conditions, whereas physiological effort measures such as heart rate and heart rate variability show no effects or effects in the opposite direction.

Experiments showed that subjective workload measures are very sensitive to increases in the error signal (e1 and e2), whereas physiological measures are more sensitive the state changes (Veltman & Jansen, 2003).

The role of State assessment in Adaptive automation

Physiological measures can be of great value in adaptive automation as is shown in several experiments (e.g. Scerbo et al., 2000; Parasuraman, 2003). However, information about the state of the operator can only be used successfully when it is combined with other information such as the difficulty of a task, the output of the operator, context and stressors. This conclusion has been drawn by others as well. However, based on the presented framework, we would like to emphasise the importance of the 'adaptability' of the operator. State changes are often a result of a successful adaptation of the operator to changing task demands. When operator tasks are reallocated to the system when the operator is doing a great job, the overall performance will not improve. Having an adaptive system working together with an adaptive operator will likely be unsuccessful. An adaptive system is more likely to work successfully when it starts reallocating tasks as soon as the operator is no longer able to adapt properly to changing task demands. In other words, only reallocate tasks in an adaptive automation setting when there are signs that the operator is unable to adequately adapt to changing task demands.

REFERENCES

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.), Human Mental Workload, (pp. 139-184). Amsterdam: Elsevier.

Hendy, K. C., East, K. P., & Farrell, P. S. E. (2001). An information processing model of operator stress and performance. In P.A. Hancock & P. A. Desmond (Eds.), Stress, workload and fatigue (pp. 34-82). Mahwah (NJ), London: Lawrence Erlbaum Associates, Publishers.

Hockey, G. R. J. (2003). Operator Functional State as a Framework for the Assessment of Performance Degradation. In G.R.J. Hockey, A. W. K. Gaillard, & A. Burov (Eds.). NATO science series, Amsterdam: IOS Press.

Norman, D. A. & Bobrov, D. G. (1975). On data-limited and resource-limited processes. Cognitive psychology, 7, 44-64.

Parasuraman, R. (2003). Adaptive Automation Matched to Human Mental Workload. In G.R.J. Hockey, A. W. K. Gaillard, & A. Burov (Eds.), Operator Functional State Assessment: The assessment and Prediction of Human Performance Degradation in Complex Tasks. NATO science series, Amsterdam: IOS Press.

Powers, W. T. (1973). Behavior: The Control of Perception. Chicago: Aldine.

Rasmussen, J. (1986). Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering. Amsterdam: Elsevier.

Scerbo, M. W., Freeman, F. G., & Mikulka, P. J. (2000). A biocybernetic system for adaptive automation. In R.W. Backs & W. Boucsein (Eds.), Engineering Psychophysiology: Issues and applications (pp. 241-254). Mahwah (N.J.), London, Lawrence Erlbaum Associates.

Veltman, J. A. (2002). A comparative study of psychophysiological reactions during simulator and real flight. International journal of aviation psychology, 12, 33-48.

Veltman, J. A. & Gaillard, A. W. (1998). Physiological workload reactions to increasing levels of task difficulty. Ergonomics, 41, 656-669.

Veltman, J. A. & Gaillard, A. W. K. (1999). Mental workload of the tactical co-ordinator of the Lynx helicopter. (Rep. No. TM-A-036). Soesterberg, The Netherlands: TNO Human Factors.

Veltman, J. A. & Jansen, C. (2003). Differentiation of mental effort measures: consequences for adaptive automation. In G.R.J. Hockey, A. W. K. Gaillard, & O. Burov (Eds.), Operator Functional State: The Assessment and Prediction of Human Performance Degradation in Complex Tasks. (pp. 249-259). NATO science series, Amsterdam: IOS Press.

Wilson, G. F., Purvis, B., Skelly, J., Fullenkamp, P., & Davis, I. (1987). Physiological data used to measure pilot workload in actual flight and simulator conditions. Proceedings of the Human Factors Society, 31st Annual Meeting; 779-783.

Zijlstra, F. R. H. (1993). Efficiency in Work Behaviour: A design approach for modern tools. Technical University

OVERTRUST DUE TO UNINTENDED USE OF AUTOMATION

Makoto Itoh
University of Tsukuba

Hiromasa Inahashi and Kenji Tanaka
University of Electro-Communications

ABSTRACT

In this study, we investigate how operator's overtrust in automation can be reduced. We have developed a model of trust in automation in order to discuss how trust becomes overtrust. Based on this model, we have conducted an experiment to examine how operators come to rely on automation too much. Previous analyses showed that it is necessary to give operators information on limit of capability of an automated system and its reason. However, giving such information was not sufficient to prevent overtrust completely. In this paper, we analyze how operators who know the reason of the limit of automation changes their understanding of the automation limit, and show that unintended use of automation causes those changes.

Key words: Trust; Overtrust, Mental Model, Automation

INTRODUCTION

Reducing overtrust in automation is becoming one of important issues in human-machine systems. Many automated systems are becoming intelligent and powerful; still, their capability is limited. It is necessary to understand how operators become reliant on automation too much in order to clarify how to reduce overtrust in automation.

Previous studies related to overtrust have focused on 'complacency' (e.g., see, Moray, 2003; Parasuraman, et al., 1993). However, several aviation accidents suggest that human operators rely on an automated system inappropriately when they misunderstand the limit of the capability of the automation. Such kind of over-reliance may occur even when an operator is highly motivated.

In this study, we investigate how a human operator comes to expect that an automated system can perform a task successfully even beyond the limit of automation. We have developed a model of trust in automation by which we are able to discuss how operator's trust in automation becomes overtrust (Itoh, Tanaka, 2000). On the basis of the model of trust, we conducted a cognitive experiment using a microworld of an automated mixed juice processing system to examine whether the range of user's expectation exceeds the limit of the capability of automation. The results showed that operators tended to rely on too much, when the operators were not informed the reason for the limit of the capability of automation (Itoh, et al., 2003). However, it was not sufficient for preventing overtrust to inform the automation limit and its reason. There were a few operators who became completely reliant on automation even though they knew the reason for the automation limit.

The structure of this paper is as follows. We give a brief description of our model of trust. The method of our experiment and the summary of previous analyses are shown. We also analyze how operators changed their understandings on the automation limit even though they were informed the automation limit and its reason.

OVERTRUST

Structure of Trust

Itoh and Tanaka (2000) proposed a model of trust in automation as shown in Figure 1. The horizontal axis in Figure 1 represents the level of difficulty for an automated system (LDA) to perform a task. It is assumed that there exists a functional limit within which the automation may work successfully (actual automation range: aAR). However, it is often restricted that operation should be done within easier situation than within the functional limit. Thus, it is assumed that the second limit (designed limit) is set to guarantee the automation to work correctly. In this paper, the area within the designed limit is called designed automation range (dAR).

Muir (1994) proposed that the notion of trust in automation has three dimensions, such as predictability, dependability, and faith. In this paper, faith (F) is regarded as situations in which a human operator expects that the automation should work.

As shown in Figure 1, F can be divided into D (dependability), UD (undependability), and UP (unpredictability). A human operator feels that the automation is reliable and dependable in D on the basis of his or her past experiences. On the other hand, the operator feels the automation to be untrustworthy in UD based on his or her experiences. Behavior of the automated system in both D and UD are predictable for a human operator. There exist some unpredictable conditions (UP) in which a human operator is not sure whether the automation is dependable or not.

The vertical axis in Figure 1 represents the level of willingness of a human operator to rely on the automation (LWRA). LWRA is assumed to range from 0 (complete distrust) to 1 (complete trust).

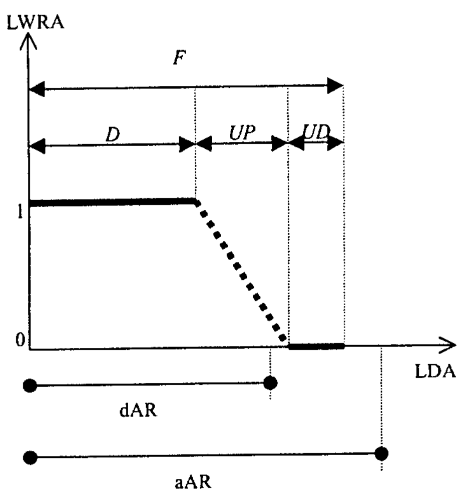


Figure 1. Structure of trust

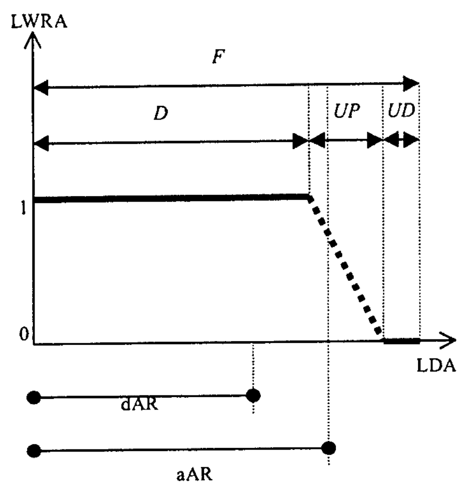


Figure 2. Example of overtrust

Overtrust

If aAR is the subset of D, we can say that the trust is one of overtrust. Moreover, it can be also regarded as overtrust when the upper bound of Faith is greater than the functional limit of the automation (Figure 2). If the operator's trust in automation is as shown in Figure 2, he or she may rely on the automation beyond aAR.

Causes of Overtrust

In many cases of accidents, an automated system was used even though the situation was not suitable for the automation. In other words, the situation was beyond the functional limit of automation. In order to improve systems safety, it is necessary to clarify why some human operators rely on automation beyond its capability.

It is assumed that human operators receive training in use of automated systems and that the operators understand the designed automation range (dAR). On the other hand, understanding of the actual automation range (aAR) is not necessarily adequate. There are two types of failure of understanding of aAR.

(1) The functional limit of automation is not explicitly informed to human operators.
(2) The functional limit of automation is given to a human operator. The reason for the functional limit, however, is not given to the operators. An operator may regard that 'true' limit is greater than the given functional limit.
(3) Both the functional limit of an automation and its reason are given to the operators. However, their understanding of aAR changes on the basis of their experiences of using the automation.

METHOD

Mixed Juice Processing Plant

The experiment in the present study is applied to computer-controlled simulation of a mixed juice pasteurizing plant as shown in Figure 3 (Itoh, et al., 1999).

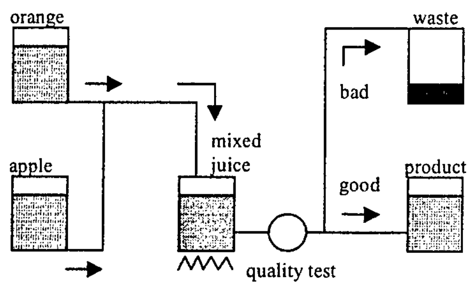


Figure 3. Mixed juice processing plant

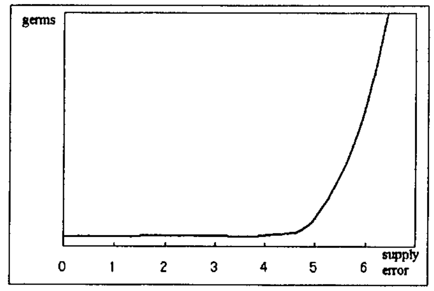


Figure 4. Supply error and residual germs

The production process of the mixed juice is automated. This automated process, however, is not always successful. The quantity of raw juice that flows into the mixture vat does not always equal exactly that specified in an order sheet. In the present paper, *supply error* (E) is referred to as the difference between the desired mass and the actual mass in the mixture vat. The automatic pasteurization is assumed to be successful in most cases if E is within five 5% of the desired mass. However, if $E > 5\%$, the pasteurization time should be manually recalculated according to the actual mass, otherwise the automatic pasteurization fails due to residual germs in most cases (Figure 4). If $E < 3\%$, the automation is guaranteed to pasteurize the juice successfully.

The task imposed on an operator is the supervision of the automation. Operators are encouraged to rely on the automatic pasteurization system as much as possible, because orders to produce mixed juice must be filled as fast as possible and automatic pasteurization is faster than manual pasteurization. Only if an operator believes that the automation has not set the pasteurization time properly, the operator should intervene and set an appropriate pasteurization time.

Participants

Thirty-three undergraduate and graduate university students volunteered to participate. Volunteers were paid for their participation.

Design and Procedure

Three types of information on limit of automation capability are compared. Participants were randomly assigned to one of the following groups.

Group 1 (G1): Operators are informed that the automation will succeed in pasteurizing the juice when the supply error is less than 3%.

Group 2 (G2): In addition to information given to G1, operators are informed that the automation may succeed in pasteurizing the juice when the error is less than 5%.

Group 3 (G3): In addition to information given to G2, operators are informed that automation will fail to pasteurize the juice when the supply error is greater than 5% because the germs are not eliminated from the juice as shown in Figure 4. Operators are also shown this figure.

The experiment lasted three days, in which it took about an hour each day. Participants were requested to perform 100 trials each day. On the first day, a participant was notified the purpose and the procedure of the experiment. Each participant received some training trials to understand when and how he or she should intervene into control.

Measure

In each trial, an operator has to decide whether he or she uses the automation for the pasteurization. Each decision on use of automation was recorded.

RESULTS AND DISCUSSIONS

For each subject, we made a plot to visualize the degree of reliance on the automation as shown in Figure 5. The horizontal axis and the vertical axis represent the trial number and the supply error at each trial, respectively. Open circles mean that the operator used the automation at the trial. Filled squares, on the other hand, are trials at which the operator intervened into control manually. Figure 5 is an example of those plots for participant 3b, who relied on the automation when the supply error was less than about 3.7% for three days.

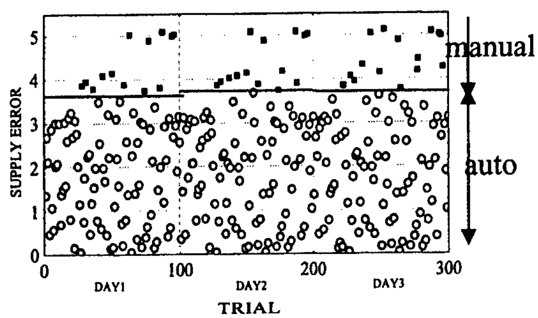


Figure 5. Degree of reliance on automation

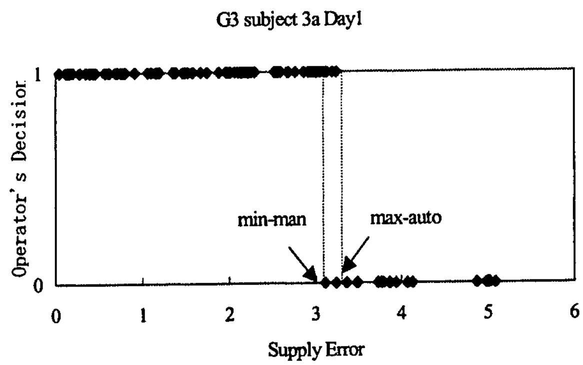


Figure 6. Mode threshold

Based on those plots, participants can be distinguished into four types (Table 1).

Type A: Operators used the automation when the supply error was less than 5%.
Type B: Operators completely relied on the automation and used in all 300 trial.
Type C: Operators used the automation only when the supply error was less than 3%.
Type D: Operators became completely reliant on the automation on the second or the third day based on their experience.

Table 1. Number of participants for each type of reliance

Group	Type			
	A	B	C	D
G1	7	1	1	2
G2	7	1	1	2
G3	7	0	2	2

On Type A, we obtained the following two values on each day for each subject (Figure 6).

(1) The maximum value of the supply errors when he or she used the automation (max-auto)
(2) The minimum value of the supply errors when he or she intervened into control (min-man)

We define *mode threshold* as the mean value of the above two. Figure 7 depicts trend of the mode thresholds. A two-way ANOVA on the mode threshold was conducted. The design was a 3 x 3 factorial, mapping onto Group and Day. Group was a between-operator factor, and Day was a within-operators factor. The ANOVA showed that a main effect of Day, $F(2,36)=10.28$, $p=0.0003$, and a main effect of Group, $F(2,18)=10.79$, $p=0.0008$).

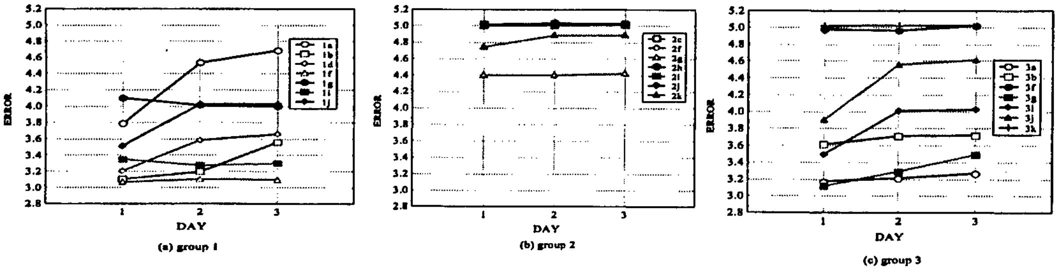


Figure 7. Trend of mode threshold

The main effect of Day can be interpreted that the mode threshold is increasing. Group 1 is a typical example. The main effect of Group suggests that the mode threshold is higher in G2 than in G1 and G3. There was not significant difference between G1 and G3 by the Tukey's HSD test; nevertheless, we can say that trend of mode thresholds in G1 is different from that in G3. According to the interview after completion of all trials, two subjects (3i, 3j) in G3 thought that the automation could be used when the supply error was less than 4.0, 4.5, respectively. Thus, we can claim that operators may rely on the automation too much when they are not informed the functional limit of an automation and/or the reason for the limit.

However, Table 1 also suggest that informing both limit of automation and its reason is not always perfect to prevent overtrust in automation. Even in G3, in which operators received the information on limit of automation and its reason, there were two persons in Type D, who became completely reliant on the automation on the second or the third day based on their experience.

According to interviews after completion of all the 300 trials, they had experiences in using the automation even though they did not intend to do. Because their mode thresholds were relatively high, they hit the button to use the automation in most trials. Thus, they mistakenly hit the button to use the automation even when the supply error was greater than 5%. The automatic heating was successful at that trial because the supply error was just slightly greater than the functional limit. This experience resulted in change of understanding on the functional limit of the automation. Typical example of this change of the mode threshold is shown in Figure 8.

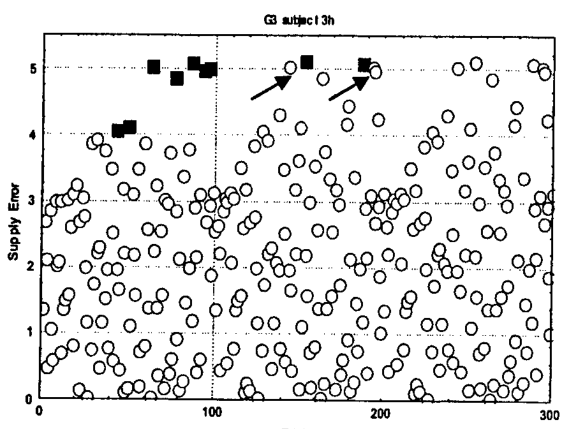


Figure 8. Example of change of understanding of automation limit due to unintended use

CONCLUSION

Overtrust is not necessarily due to overtrust-prone or complacent characteristics of people. Our results suggest that people may rely on automation too much if information on the functional limit of capability of automation and its the reason is not appropriately given.

However, it is not always sufficient to inform operators the functional limit of automation and its reason. Even though operators had understood the limit of automation correctly, some operators changed their understanding of the automation limit based on their experiences of using automation. This phenomenon can occur in the real world. It may be difficult for operators to distinguish whether current operating condition is within the functional limit or not. If an operator uses the automation mistakenly when the current operating condition seems to be beyond the functional limit, the operator may change their understanding on the functional limit which result in overtrust.

In order to reduce overtrust due to unintended use of automation, it is necessary to support situation awareness on the relationship between current operating condition and limit of capability of automation.

ACKNOWLEDGMENTS

This work was supported by Grant-in-Aid for Scientific Research 11780299 of the Japanese Ministry of Education, Science, Sports and Culture, Nissan Science Foundation, and Secom Science and Technology Foundation.

REFERENCES

Itoh, M., Abe, G., and Tanaka, K. (1999). Trust in and Use of Automation: Their Dependence on Occurrence Patterns of Malfunctions, Proc. IEEE-SMC Conference, III, 715-720.

Itoh, M., and Tanaka, K. (2000). Mathematical Modeling of Trust in Automation: Trust, Distrust, and Mistrust, Proc. IEA2000/HFES2000, 1, 9 -12.

Itoh, M., Inahashi, H., and Tanaka, K. (2003) Informing Limit of Ability and its Reasons for Reducing Overtrust in Automation, Proc. XVth Triennial Congress of the Inernational Ergonomics Association, 1, 141-144.

Moray, N. (2003). Monitoring, Complacency, Scepticism and Eutactic Behavior, Int. J. Industrial Ergonomics, 31, 175-178.

Muir, B. (1994). Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems, Ergonomics, 37(11), 1905-1922.

Parasuraman, R., Molloy, R., Singh, I. L. (1993). Performance Consequences of Automation-Induced "Complacency," Int. J. Aviation Psychology, 3, 1-23.

EVENT RATE CHANGES AND MONITORING PERFORMANCE USING A BIOCYBERNETIC ADAPTIVE SYSTEM

Mark W. Scerbo, Frederick G. Freeman, Peter J. Mikulka
Old Dominion University

ABSTRACT

The present study was designed to examine whether an adaptive, biocybernetic system could generate a pattern of event rate changes in a vigilance task that would enhance performance. In session 1, participants performed a 40-min vigil while an index of task engagement was derived from their EEG activity. This index was used to change the presentation rate of events among three values: 6, 20, and 60 events/min. Event rates were changed according to a negative or positive feedback contingency. The schedule of changes among event rates was recorded and in session 2, half of the participants were yoked to their own prerecorded schedule and half were yoked to the prerecorded pattern generated by someone in the opposite contingency. In session 1, there was a trend toward better performance under negative feedback. In session 2, the performance of participants operating under the schedule of event rate changes that they generated under negative feedback was significantly better than that of those operating under the schedule of event rate changes that they generated under positive feedback. These findings demonstrate that the schedule of event rate changes established with a brain-based, adaptive automation system can produce performance benefits that transcend the initial period of interaction with the system.

Keywords: adaptive automation, vigilance, psychophysiology

INTRODUCTION

Adaptive automation refers to systems where decisions regarding initiation, cessation, and mode of operation are shared between the human operator and the system in real time (Parasuraman et al., 1992; Scerbo, 1996). The object of adaptive systems is to adjust situational demands, restructure the environment, and maintain more stable levels of workload thereby enhancing operator performance. Interest in adaptive automation is fueled by concerns over the difficulties operators have when working with complex systems that have multiple modes of automation (Woods, 1996). Byrne and Parasuraman (1996) suggested the use of physiological measures in the design and regulation of adaptive systems because such measures are relatively unobtrusive as compared to subjective or secondary task measures and can allow a real time assessment of workload and effort. Several studies have now shown that a brain-based, adaptive system that uses the operator's own EEG can moderate workload and improve performance on a compensatory tracking task (Freeman, Mikulka, Prinzel, & Scerbo, 1999; Freeman, Mikulka, Scerbo, Prinzel, & Clouatre, 2000; Prinzel, Freeman, Scerbo, Mikulka, & Pope, 2000).

Recently, Mikulka, Scerbo, and Freeman (2002) investigated whether the same brain-based, adaptive automation system shown to improve tracking performance might also improve vigilance performance. In their study, participants were asked to monitor the repetitive presentation of a pair of white lines on a computer screen for occasional increases in length. Each participant's EEG was recorded and used to compute an engagement index in which the relative power in the beta bandwidth (13-30 Hz) was divided by the relative power in the alpha (8-12 Hz) and theta (4-7 Hz) bandwidths (Pope, Bogart, & Bartolome, 1995). This index was used to control the presentation rate of stimulus events. Three different rates were used: 6, 20, and 60 events per minute. In addition, two feedback contingencies were studied. Under negative feedback, if the participant's engagement index increased the rate of presentation was decreased and if the index decreased, the rate of presentation increased. The opposite was true for positive feedback. Each experimental participant was paired with a yoked control participant who received the same pattern of changes in event rate, but whose EEG had no effect on the pattern of changes in event rates. Mikulka, et al. found that *both* the experimental and yoked participants performed significantly better under negative as compared to positive feedback, but the interaction between type of feedback and time was limited to the first and fourth periods.

Another way to examine the effects of positive and negative feedback contingencies on vigilance would be to record an individual's pattern of event rate changes and have the individual perform a second vigil using the schedule of changes from his/her previous session. If the benefits of a negative feedback contingency are tied to real-time adaptive conditions, then one would expect performance to be optimal when the schedule of event rate changes is coupled to the individual's engagement index. Likewise, performance should be particularly poor under positive feedback in the real-time, adaptive condition as compared to the uncoupled condition. On the other hand, it is also possible the schedule of event rate changes derived from one's own EEG would have beneficial effects under negative feedback (and detrimental effects under positive feedback) that transcend the session in which they were recorded as was observed by Mikulka et al. (2002). The goal of the present study was to examine these two possibilities.

METHOD

Participants

Twenty undergraduate students served as participants in this study. Their ages ranged from 18 to 35 years ($M = 23$). Seventy percent of the participants were female, but comparable numbers of males and females were assigned to each condition. All participants had normal or corrected-to-normal vision.

EEG Recording and Engagement Index

EEG was recorded using a montage of four sites: F3, F4, 01, and 02. The left mastoid was used as the reference site. Each amplified EEG channel was digitized at a rate of 200 samples per second in a circular buffer array. These samples were taken from the buffer in four vectors, one per input channel (site), with each vector containing 512 data points resulting in 2.56 seconds of data per channel. Each vector was smoothed using a Hanning windowing procedure. The power spectrum was computed using a Fast Fourier transformation. Bin powers were combined to calculate total power in three bandwidths (theta: 4-7 Hz, alpha: 8-12 Hz, and beta: 13-30 Hz). Bin powers are the estimates of the power spectrum within bins between discrete Fourier frequencies of 0-256 Hz. Bandwidth powers were divided by total power to produce percent power. The array of percent power for the four sites by the three bandwidths was used to compute the engagement index, 20 beta/(alpha + theta). The index was first computed over a 20-second period and then updated every two seconds using a sliding 20-second window. The engagement index, 20 beta/(alpha + theta), has been shown to vary between 2 and 20 (higher values reflect higher levels of engagement) and is the most effective of several indices employed by Freeman, et al. (1999) and Pope, et al. (1995).

Apparatus

EEG was recorded using an Electro-cap International lycra sensor cap. The cap consists of 22 recessed tin electrodes arranged according to the international 10-20 system. EEG was recorded using a BIOPAC EEG100A differential amplifier module consisting of four, high gain, differential input, bio-potential amplifiers. The low and high pass filters were set at 100 and 1 Hz, respectively.

The amplifier was connected to a Macintosh Quadra. A LabVIEW Virtual Instrument (VI) calculated total EEG power in the three bandwidths: alpha, beta and theta. The VI also calculated the engagement index and commanded the task mode changes through serial port connections to the task computer.

An artifact rejection subroutine examined the amplitudes of each epoch from the four digitized channels of EEG and compared them with pretrial tests in which the participant's eyes were open and closed. A power spectral distribution was then derived and if the voltage in any channel exceeded the threshold by more than 25%, the epoch was excluded when computing the index in subsequent analyses. Less than 1% of any participants' data file was rejected.

Task

The task consisted of a 40-min vigil analyzed in four consecutive 10-min periods. Participants were asked to monitor the repetitive presentation of a pair of 3mm (W) X 38mm (H) white lines separated by 25mm. The lines were presented against a blue background and appeared in the center of the computer screen. Critical signals were pairs of lines that were 2mm taller and occurred once a minute at random intervals. All stimuli were presented for 300 ms. Participants were required to respond to the presence of critical signals by pressing the space bar on the keyboard. Responses made to critical signals within 1000ms of stimulus onset were considered correct detections. All other responses were logged as false alarms for the signal detection analyses (see below).

Three different event rates that could be considered slow, moderate, and fast (6, 20, and 60 events per minute) according to Davies and Parasuraman's (1982) original taxonomy were used. The occurrence of critical signals was tied to a predetermined schedule of seconds for each minute of the vigil. Thus, when the event rate was 60 and a critical signal was scheduled to appear at the 41st second within the minute, the 41st event would be presented as a critical signal. However, under slower event rates (6 and 20) if no stimulus event was presented when a critical signal was scheduled to occur (e.g., the 41st event under an event rate of 6), a critical signal would be substituted for the next stimulus event (i.e., the event presented at 42 seconds into the minute would be a critical signal).

The mean and standard deviation of the engagement index were derived from a 5-min baseline practice period with an event rate of 20. This value of the index was then used to determine event rate changes. If the value of the index moved 0.2 *sd* or more above or below the baseline level, the event rate was shifted. Pilot testing showed that a *sd* of 0.2 was sufficiently sensitive to switch among task modes. For the participants in the negative feedback condition, the event rate increased to 60 when the index dropped 0.2 *sd* below the baseline value and decreased to 6 when the engagement index rose 0.2 *sd* above baseline. Conversely, for participants in the positive feedback condition the event rate increased to 60 when the engagement index rose 0.2 *sd* above the baseline value and decreased to 6 when the engagement index fell 0.2 *sd* below the baseline value. The schedule of event rate changes was recorded for all participants.

Procedure

The experiment took place in an electronically shielded room in a secluded and quiet experimental suite. The room was illuminated by two 75 watt bulbs contained in ceiling fixtures. All participants were run individually. They were fitted with the electrode cap and had their scalps prepared to reduce the impedance levels for the four recording sites and the reference site below 5 kOhms. The participants were seated about 0.5 meters in front of a desk containing the computer with a display placed at eye level.

The participants were given instructions on the vigilance task and then began a 5-min practice session to become familiar with the task and to establish a baseline value for the engagement index. They were asked to press a response button every time they detected a critical signal. The signal detection score, A′, (see below) was calculated and if their practice score fell below 0.7, they were required to complete another 5-min practice session. All participants met this criterion. After the practice session the participants were given a brief 1-min rest and then completed the first experimental session. Half of the participants were randomly assigned to either the positive or negative feedback condition.

After session 1, participants returned a week later to complete the second session. The procedure was exactly the same with one important exception. The changes among event rates were determined by the patterns generated during the first session. Thus, for the second session half of the participants in each feedback group were yoked to either their own pattern of event rate changes (same schedule) or to a pattern generated by another participant in the opposite feedback condition from session 1 (different schedule). Although EEG signals were recorded in session 2, they had no effect on the pattern of event rate changes.

RESULTS

Session 1

Vigilance performance was measured using the nonparametric signal detection indices of sensitivity, A′ (Grier, 1971), and response criterion, B''_D (Donaldson, 1992). The mean A′ scores for each group over the four periods are shown in Figure 1. As can be seen in the figure, better vigilance performance was observed under negative as compared to positive feedback conditions. The A′ scores were analyzed with a 2 feedback (positive, negative) by 4 periods ANOVA.

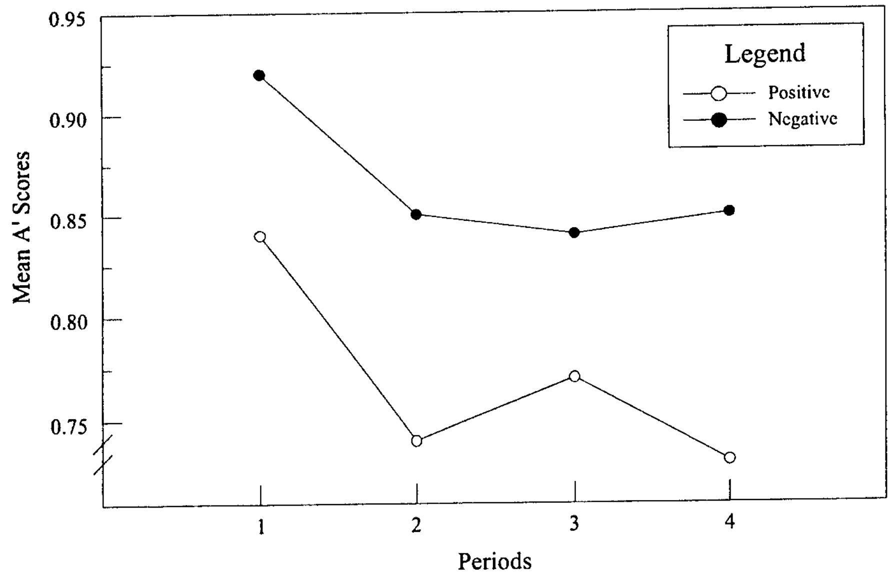


Figure 1. Mean A′ scores for positive and negative feedback groups as a function of time.

Although the results for feedback were in the hypothesized direction, the effect did not reach statistical significance, $F(1, 18) = 3.69$, $p<.08$. A significant effect for periods was observed, $F(3, 54) = 3.15$, $p<.05$, but the interaction between feedback and periods did not reach significance. There were no significant effects of B''_D.

Session 2

The mean A′ scores for each group over the four periods of watch are shown in Figure 2. As can be seen in the figure, the level of performance for the participants in the negative feedback condition who were yoked to their previous pattern of event rate changes was quite good and remained that way across the vigil. Conversely, the level of performance for the participants in the positive feedback condition who were yoked to their previous pattern of event rate changes was initially poor and remained poor throughout the vigil.

The A′ scores for participants who were in the negative feedback condition in session 1, but who were yoked to a participant from the positive feedback condition from session 1, were initially high in session 2, but declined over the course of the vigil. By the last 10 minutes, their performance did not differ from those in the positive-positive feedback group. Those participants who were in the positive feedback condition in session 1, but who were yoked to a participant from the negative feedback condition from session 1 began the second session performing comparably to the participants in the negative-negative feedback group. Although their performance did decrease during the third 10-min period, their performance for the final 10-min period did not differ markedly from the negative-negative feedback group.

A 4 condition (positive-positive, positive-negative, negative-negative, and negative-positive) by 4 periods ANOVA of the A′ scores yielded a significant effect for feedback, $F(3, 16) = 3.38$, $p<.05$ and a marginally significant effect for Periods, $F(3, 48) = 2.73$, $p<.06$.

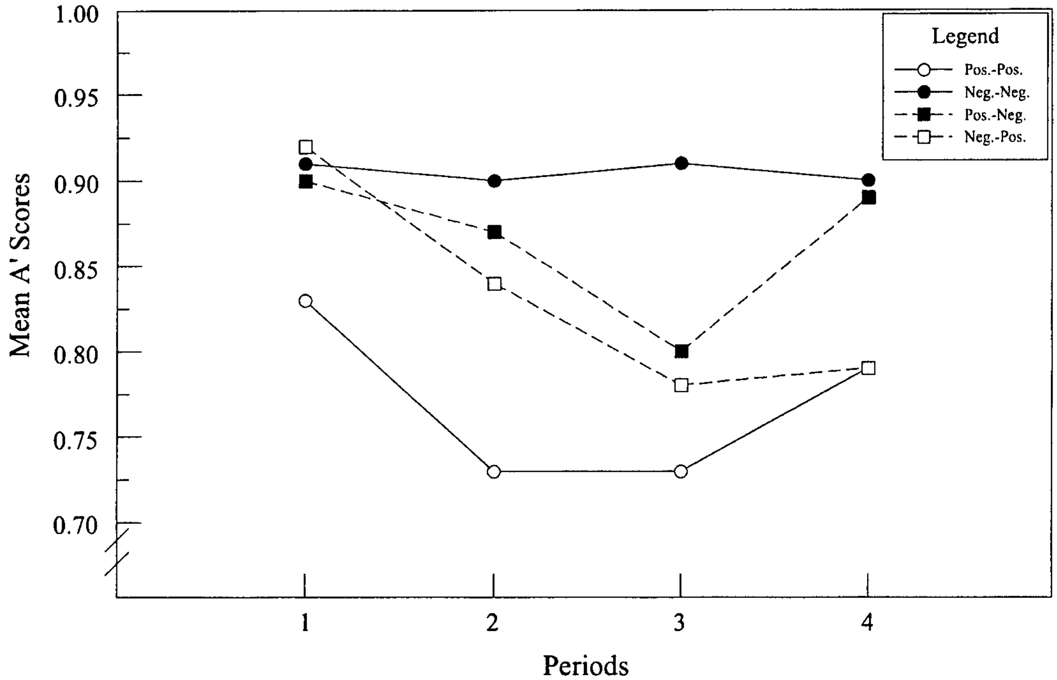


Figure 2. Mean A′ scores for positive-positive, negative-negative, positive-negative, and negative-positive groups as a function of time.

The interaction was not significant. Newman-Keuls comparisons revealed that the negative-negative feedback group performed significantly better than the positive-positive feedback group ($p<.05$). No other differences were significant.

DISCUSSION

The goal of the present study was to examine how different schedules of event rate changes created under positive and negative feedback contingencies with a brain-based, adaptive system would affect vigilance performance. The schedules of event rate changes generated in the first session were used to produce event rate changes in the second session. Half of the participants received the same schedule of event rate changes they generated in their first session and the other half received a schedule generated by someone else in the opposite feedback contingency.

The results from the first session showed an advantage for negative over positive feedback and were consistent with those of Mikulka et al. (2002); however, the effect did not reach significance. Moreover, both groups declined over the course of the vigil.

A different picture emerged from the second session. Although no overall decrement was observed, there were differences between the groups. Specifically, the performance of those individuals operating under the same schedule of event rate changes generated in their first session was dependent upon feedback. The schedule of changes produced under negative as compared to positive feedback in session 1 resulted in better performance in session 2. Thus, the effects of the schedules generated in session 1 transcended the adaptive conditions under which they were created. Moreover, Figure 2 shows that the advantages of the negative feedback schedule and disadvantages of the positive feedback schedule could also be seen for the groups that operated under the opposite feedback contingencies; however, these trends were not statistically significant. This finding suggests that the intra-participant variability in performance was lower than inter-participant variability.

The better performance observed under negative feedback in session 2 is consistent with the observations of Mikulka, et al. (2002). However, it is important to note that in the Mikulka, et al. study, the mean overall event rates for the positive and negative feedback conditions were approximately 26 and 17 events/min, respectively. Those means lie on either side of the 24 events/min value originally proposed by Davies and Parasuraman (1982) to distinguish between slow and fast event rates. According to their taxonomy, the source of the vigilance decrement is perceptual in nature only when observers are required to make an absolute judgment under a high event rate (i.e., 24 events/min or higher). More recently, See, Howe, Warm, and Dember (1995) performed a meta-analysis of perceptual sensitivity decrements in 42 vigilance experiments and reported that the magnitude of the decrement is a function of continuous changes along an event rate continuum.

Thus, it is possible that the results from the present study might also be tied event rate differences. However, an examination of the mean overall event rates generated in session 1 indicated that they were almost identical. Specifically, the mean event rates under positive and negative feedback were 21 and 22, respectively. Thus, the performance differences observed in session 2 could not be attributable to the overall event rate. Instead, the results from this study suggest that the performance differences are related to the timing of shifts to higher and lower event rates dictated by the positive and negative feedback contingencies.

ACKNOWLEDGEMENTS

This research was supported in part by NASA Langley Research C Center, grant NCC-176, to Old Dominion University.

REFERENCES

Byrne, E. A., & Parasuraman, R. (1996). Psychophysiology and adaptive automation. Biological Psychology, 42, 249-268.

Davies, D. R., & Parasuraman, R. (1982). The psychology of vigilance. London: Academic Press.

Donaldson, W. (1992). Measuring recognition memory. Journal of Experimental Psychology: General, 121, 275-277.

Freeman, F., Mikulka, P., Prinzel, L., & Scerbo, M. (1999). Evaluation of an adaptive automation system using three indices with a visual tracking task. Biological Psychology, 50, 61-76.

Freeman, F., Mikulka, P, Scerbo, M, Prinzel, L., & Clouatre, K. (2000). Evaluation of a psychophysiollogically controlled adaptive automation system, using performance on a tracking task. Applied Psychophysiology and Biofeedback, 25, 103-115.

Grier, J. B. (1971). Non-parametric indices for sensitivity and bias: Computing formulas. Psychological Bulletin, 75, 424-429.

Mikulka, P.J., Scerbo, M.W., & Freeman, F.G. (2002). The effects of a biocybernetic system on vigilance performance. Human Factors,44, 654-664.

Parasuraman, R., Bahri, T., Deaton, J.E., Morrison, J.G., & Barnes, M. (1992). Theory and design of adaptive automation in aviation systems (Progress Report No. NAWCADWAR-92033-60). Warminster, PA: Naval Air Warfare Center, Aircraft Division.

Pope, A., Bogart, E., & Bartolome, D. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. Biological Psychology, 40, 187-195.

Prinzel, L.J., Freeman, F.G., Scerbo, M.W., Mikulka, P.J., & Pope, A.T. (2000). A closed-loop system for examining psychophysiological measures for adaptive automation. International Journal of Aviation Psychology, 10, 393-410.

Scerbo, M. W. (1996). Theoretical perspectives on adaptive automation. In R. Parasuraman & M. Mouloua (Eds.), Automation and human performance: Theory and applications (pp. 37-63). Mahwah, NJ: Lawrence Erlbaum Associates.

See, J. E., Howe, S. R., Warm, J. S., & Dember, W. N. (1995). Meta-analysis of the sensitivity decrement in vigilance. Psychological Bulletin, 117, 230-249.

Woods, D.D. (1996). Decomposing automation: Apparent simplicity, real complexity. In R. Parasuraman & M. Mouloua (Eds.), Automation and human performance: Theory and applications (pp. 3-17). Mahwah, NJ: Lawrence Erlbaum Assoc.

TASK DEPENDENCIES IN STAGE-BASED EXAMINATIONS OF THE EFFECTS OF UNRELIABLE AUTOMATION

Scott M. Galster
Air Force Research Laboratory

Raja Parasuraman
The Catholic University of America

ABSTRACT

Following the release of the Parasuraman, Sheridan, and Wickens (2000) model of human-interaction with automation, there have been a number of studies conducted that have examined the effects of unreliable automation by the stage (of information-processing) the automation was present. Overwhelmingly, these studies have indicated that unreliable automation in the decision-aiding stage has contributed to greater performance decrements than unreliable automation present in any of the other stages (information acquisition, information analysis, or action implementation). The present paper will outline the studies that have demonstrated this effect. It will also present data from three recent studies that did not support the general conclusion that the decision-aiding stage produced the greatest performance decrement when the automation was less than perfectly reliable. Further, the paper will outline a plausible explanation for the differences observed based on elements associated with the decision-making stage of the required tasks. In addition, the paper will argue that performance decrements observed due to unreliable automation may be task dependent.

Keywords: Automation, human-interaction with automation, decision-aiding

INTRODUCTION

In an attempt to look at differential performance effects by stage of automation, Crocoll and Coury (1990) examined decision-aiding performance when operators were given status, recommendation, or status *and* recommendation cues in an aircraft identification task. The first two of these conditions can be associated with the information analysis and decision selection stages of automation in the subsequently developed Parasuraman *et al.* (2000) model. Operators were required to visually identify aircraft as being hostile, friendly or unknown and then choose a fire or no fire response in accordance with stated rules of engagement. The "tight" rule of engagement allowed the operator to fire only upon hostile aircraft while the "free" rule of engagement allowed firing upon hostile and unknown aircraft. During the first three sessions, participants learned how to identify 10 friendly and 10 hostile aircraft, identify unknown aircraft types, and apply the rules of engagement criteria. In the fourth session, the data collection session, participants were divided into four groups and tested on their ability to choose the correct engagement decision. The first group was the control group and received no aiding. The second, third, and fourth groups received status only, recommendation only, or status and recommendation aiding, respectively. The decision aiding was reliable 96% of the time when the automation was present. The percent of correct engagement decisions made and the response times were recorded. It was unclear if the trials were time limited or if they continued until the participant responded.

The percent of correct engagement decisions was greater than 96% for all conditions and did not show a significant difference between the automated and control conditions. The response times significantly improved when the automation was present compared to the non-aided control group but there was not a significant difference between the three aided conditions. Crocoll and Coury (1990) decided to examine the performance on the automation-aided trials to see if there was a difference when the aid was unreliable (8 of the 200 trials for each group). They found that the group that received the status only aid responded correctly 95% of the time while the status and recommendation, and the recommendation only groups responded correctly 86% and 80% of the time respectively. The data indicated that there was a greater cost when the recommendation aiding was present compared to the status only or the status and recommendation aiding conditions. Crocoll and Coury surmised that participants who were provided a recommendation decision aid blindly followed that aid compared to the participants who received the status only or status and recommendation decision aiding.

Sarter and Schroeder (2001) conducted a study comparing pilot performance during escalating in-flight icing conditions using two types of decision-aids during simulated flight. The first decision aid in their study presented icing information (status display) and the other decision-aid recommended actions to mediate the icing condition (command display). They demonstrated that imperfect automation led to reduced performance while using the decision aiding (command display) over both the status display and the baseline condition where no automation was present. This result is consistent with the suggestion that the negative effects of unreliable automation in the decision stage may be more pronounced than the information analysis stage (Parasuraman *et al.*, 2000).

Rovira, McGarry, and Parasuraman (2002) also found a greater cost in performance when the decision-aiding automation was unreliable compared to when the information analysis stage was unreliable in a sensor-to-shooter task. These effects generalized across three different forms of decision automation. Furthermore, they found that this performance decrement dropped below manual performance as measured by the percentage of correct detections in a command and control task. In addition, they included varying reliability rates (80% vs. 60%) and noted that there was a greater cost in the decision-aiding stage than in the information analysis stage. This cost was greater in the higher reliability condition compared to the lower reliability condition, consistent with the findings on automation complacency reviewed earlier (Parasuraman *et al.*, 1993). McGarry, Rovira, and Parasuraman (2003) found similar results but also noted that the findings applied to tasks that were longer in duration than the original sensor-to-shooter task that was reported by Rovira, McGarry *et al.* (2002).

A similar pattern of results was obtained in a multi-task environment using the MAT battery (Rovira, Zinni, & Parasuraman, 2002). There was a general decline in performance when the automation was unreliable over when it was reliable. Also, there was a differential performance decrement for the unreliable automation conditions depending on what stage the automation was employed. There was a greater drop in performance when the automation was employed in the decision-aiding stage over the information analysis stage. Further, the results indicated that the higher reliability rate induced a greater cost in detections, again indicating a complacency effect that was similar to that found by Parasuraman *et al.* (1993).

These studies have consistently demonstrated that unreliable automation has a greater detrimental performance effect in the decision-making stage as compared to any other stage that automation may be present. Recently however, results that demonstrated a performance decrement in the information automation stage have been reported (Galster, Bolia, Roe, & Parasuraman, 2001; Galster, Bolia, & Parasuraman, 2002a; Galster, Bolia, & Parasuraman, 2002b). These studies utilized a common simulation environment that required participants to search a display for the presence of a pre-defined target and respond to its presence or absence. The basic visual search task was utilized across the three studies to ensure a common testing environment. To date, a common testing environment has not been used to explore incremental changes in the use of automation by the stage it is implemented. Utilizing this common environment, the first study examined the differences in target detection and response times between manual and automated cueing conditions. The automated cuing condition (IA) represented the fusion of the information acquisition and analysis stages. As pointed out by Parasuraman *et al.* (2000), these stages are commonly combined because they occur prior to the decision-making point and represent information automation. The number of distractors in the search area was manipulated (10 or 20) to represent varying levels of workload. In this and every study that used this task environment, a response was required within 2500ms for the presence or absence of a target among the distractor set. The purpose of the first study was to; (a) evaluate the visual search cueing platform (Yeh & Wickens, 2001); (b) apply a simplified human interaction with automation model (Parasuraman *et al.*, 2002); and (c) use a simple task (Rovira, McGarry *et al.*, 2002) in the evaluation of the benefits of automation in high and low workload conditions (Merlo *et al.*, 2000) under considerable temporal constraints (Muthard & Wickens, 2001). Further, the reliability of the automated cue was manipulated so that cue validity effects could be examined (Wickens Conejo, & Gempler, 1999; Yeh, Wickens, & Seagull, 1999).

The second study included a decision-aiding cue (DA) similar to the one used in the study by Crocoll and Coury (1990). A higher distractor set size (30) was also added to increase the variability of the workload. In addition to the manual, information automation, and decision-aiding automation conditions the latter two were combined and presented either together (co-located) or separately resulting in five automation conditions.

As Wickens and Xu (2002) have noted, automation reliability levels seem to influence human-system performance differently, depending on the stage of automation. The third study varied the reliability level of the automation as a between-groups factor. All other experimental factors from the previous study were unchanged except the condition where the combined information automation and decision-aiding cues that were presented separately was dropped. This study allowed for the examination of human-system performance differences as the reliability level was manipulated between stages, similar to the Crocoll and Coury (1990), Sarter and Schroeder (2001), Rovira, McGarry *et al.* (2002), and Rovira, Zinni *et al.* (2002) studies. These studies did not treat the reliability level of the automation as a between-subjects factor. By including this in the third study the potential human-system performance changes by stage can be examined as a function of the reliability level experienced by the operators.

Visual Search results

The first visual search study had only one stage of automation present and was represented by the (IA) cue. Even though only one stage was present there were differences noted between the automation that was perfectly reliable and the automation that was unreliable. For the percentage of correct responses there was a significant performance decrement between the reliable and unreliable conditions but only for the higher distractor set size. The data for the response times indicated that participants took longer to respond when they made a correct response when the IA cue was unreliable. For this measure, the response times were higher in the larger distractor set size than the smaller distractor set size. A similar pattern of results was obtained for the percentage of trials that ended in a timeout (exceeding the 2500ms threshold).

In the second visual search study, the percent of correct responses in the IA and DA conditions were both above the manual condition when the automation was reliable, as expected. When the automation was unreliable however, the percent of correct responses for both the IA and DA conditions fell below the manual baseline condition. This finding is not consistent with the results of previous studies when the magnitude of the decrement is evaluated. The difference in the IA condition was greater than the difference in the DA condition between reliable and unreliable automation conditions. In other words, unreliable IA cues in the information automation stage created a larger performance cost, in terms of the percentage of correct responses, than the unreliable DA cues in the decision-aiding stage.

The results of the third visual search study were also informative with regard to the reliability level of the automation. In terms of the percentage of correct responses, the IA cue consistently lead to higher performance over the manual condition, regardless of the reliability level of the automation (50%, 70%, or 90%). The DA cueing condition however only surpassed the manual condition when the automation was at the 90% reliability level. Otherwise, the DA conditions were about the same (70% condition) or lower (50% condition) than the manual condition for the percentage of correct responses. Additionally, performance was consistently lower for the DA cueing condition than for the IA cueing condition. This data suggests that there was a performance decrement in the decision-aiding stage for correct detections as compared to the information automation stage. The DA condition performance did not however go below the manual performance until the level of the automation reliability was chance.

The response times to correct responses also revealed a differential effect for the level of reliability by the stage the automation was employed. For the DA cued condition, the response times were consistently close to the response times in the manual condition across all automation reliability levels. The IA cued conditions demonstrated a performance improvement over the manual condition and the DA cued condition as the reliability level of the automation increased.

ANALYSIS

One can postulate that the reason for the inconsistent result is the nature of the task that was being performed. The visual search task was temporally compressed and a decision could not be made until either (a) the target was located, or (b) an exhaustive search was conducted on the entire search field. In contrast, the Sarter and Schroeder (2001) task was based on a decision support system that emphasized the decision-making stage of the information-processing cycle. In addition, the duration of the flight task was much longer than that of the visual search task. The duration of the flight task was often in excess of 65s from the initial onset of the icing condition. The Rovira, McGarry *et al.* (2002) and McGarry *et al.* (2003) sensor-to-shooter task was also focused on decision-support. The trials were also longer (10s) than those in the visual search task. It can be argued that the visual search task is more of a perception task than a decision-making or decision support task. It may be the case that the effects of unreliable automation are task dependent. In higher order, more cognitively demanding tasks, the unreliable automation may have a more detrimental effect in the decision-aiding stage while in lower cognitively demanding tasks the detrimental effect may be tied to the earlier information automation stages. Wickens and Carswell (1997) provide a plausible explanation for the differing decremental effects. They posit that the number of transformations to the raw data that the human needs to make will increase the time and complexity of the overall information-processing cycle.

A similar argument can be made that result differences follow decision-making predictions (Letho, 1997). For example, if a decision tree is utilized to reflect the task structure and decision making process, the visual search task allows for a decision point much sooner (pattern matching) than the task that requires an evaluation of potential decision alternatives. Further, if several decision alternatives are available, the associated risks need to be evaluated for each decision option. This would shift the emphasis within the information-processing cycle from the information stages to the decision-making stages.

The purpose of this paper is to point out that there are inconsistencies in the results of experiments that examine the effects of unreliable automation. Determining the relative costs and benefits of imperfect automation for different stages will lead to the development of more robust automation that supports the human operator.

REFERENCES

Crocoll, W.M., & Coury B.G. (1990). Status or recommendation: Selecting the type of information for decision aiding. In Proceedings of the Human Factors and Ergonomics Society 34th Annual Meeting (pp. 1524-1528). Santa Monica, CA: Human Factors and Ergonomics Society.

Galster, S.M., Bolia, R.S, & Parasuraman, R. (2002a). Effects of information automation and decision-aiding cueing on action implementation in a visual search task. In Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting (pp. 438-442). Santa Monica, CA: Human Factors and Ergonomics Society.

Galster, S.M., Bolia, R.S, & Parasuraman, R. (2002b). The application of a qualitative model of human-interaction with automation: Effects of unreliable automation on performance. In *Proceedings of the RTO HFM Symposium on* "The Role of Humans in Intelligent and Automated Systems" held in Warsaw, Poland, 7-9 October, 2002. (paper 15).

Galster, S.M., Bolia, R.S., Roe, M.M., & Parasuraman, R. (2001). Effects of automated cueing on decision implementation in a visual search task. In Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting (pp. 321-325). Santa Monica, CA: Human Factors and Ergonomics Society.

Lehto, M.R. (1997). Decision Making. In G. Salvendy (Ed.), Handbook of human factors and ergonomics, second edition (pp. 1201-1248). New York, NY: Wiley & Sons.

McGarry, K., Rovira, E. & Parasuraman, R. (2003). Effects of task duration and type of automation support on human performance and stress in a simulated battlefield engagement task. In Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society (pp. 548-552). Santa Monica, CA: Human Factors and Ergonomics Society.

Muthard, E.K., & Wickens, C.D. (2001). Change detection and the confirmation bias in aviation route planning (Technical Report ARL-01-18/NASA-01-9). Savoy, IL: University of Illinois, Aviation Research Lab.

Parasuraman, R., Molloy, R., & Singh, I.L. (1993). Performance consequences of automation induced "complacency". International Journal of Aviation Psychology, 3, 1-23.

Parasuraman, R., Sheridan, T.B., Wickens, C.D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 30, 286-297.

Rovira, E., McGarry, K., & Parasuraman, R. (2002). Effects of unreliable automation on decision making in command and control. In Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society (pp. 428-432). Santa Monica, CA: Human Factors and Ergonomics Society.

Rovira, E., Zinni, M., & Parasuraman, R. (2002). Effects of information and decision automation on multi-task performance. In Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society (pp. 327-331). Santa Monica, CA. Human Factors and Ergonomics Society.

Sarter, N., & Schroeder, B. K. (2001). Supporting decision-making and action selection under time pressure and uncertainty: The case of in-flight icing. Human Factors, 43 (4), 573-583.

Wickens, C.D., & Carswell C.M. (1997). Information processing. In G. Salvendy (Ed.), Handbook of human factors and ergonomics, second edition (pp. 89-129). New York, NY: Wiley & Sons.

Wickens, C.D., Conejo, R., & Gempler, K. (1999). Unreliable automated attention cueing for air-ground targeting and traffic maneuvering. In Proceedings of the Human Factors and Ergonomics Society the 34th Annual Meeting (pp. 21-25). Santa Monica, CA: Human Factors and Ergonomics Society.

Wickens, C.D., & Xu, X. (2002). Automation trust, reliability and attention HMI 02-03 (Technical Report No. AHFD-02-14/MAAD-02-2). Savoy, IL: University of Illinois, Aviation Research Lab.

Yeh, M., & Wickens, C.D. (2001). Display signaling in augmented reality: The effects of cue reliability and image realism on attention allocation and trust calibration. Human Factors, 43(3), 355-365.

Yeh, M., Wickens, C.D., & Seagull, F.J. (1999). Target cueing in visual search: The effects of conformality and display location on the allocation of visual attention. Human Factors, (41), 524-542.

AUTOMATION IN CONSUMER SOFTWARE: NEW DOMAIN, SAME ISSUES

Anthony D. Andre, Jennifer R. Kingsburg and Stephen G. Shelden
Interface Analysis Associates

ABSTRACT

This paper addresses the role of automation in everyday consumer software and demonstrates that many of the lessons learned from the study of automation in complex domains can be directly applied to the personal computer domain. Numerous examples of software automation will be presented, with the end-goal of producing a preliminary taxonomy of software automation purposes, a list of software automation problems and a set of software automation design guidelines.

Keywords: Consumer Software; Software Automation; Human-Computer Interaction

INTRODUCTION

Automation has been a central theme of human factors research for the past twenty years. Scores of articles and books have been published on the topic, and to date, there exists well-established automation taxonomies, descriptions of common problems and design guidelines (e.g., Billings, 1996; Lyall and Funk, 1998; Degani, 2004). As usability consultants we are constantly challenged to design consumer software applications with increasing levels of automation. And as users of personal computers we are equally challenged, on a daily basis, to understand how and why our software behaves the way it does. Yet, there are no published guidelines for addressing automation issues in this context. Books devoted to the topic of automation are typically limited to transportation, process control and medical applications (e.g., Scerbo and Mouloua, 1999). Those that address the human factors of more "everyday" products (e.g., Norman, 1988) are no more concerned with the impact of software automation. Even software manufacturer interface design standards (e.g., Microsoft Corporation, 1995) make no explicit reference to the unique interface design requirements of automated functions. It is interesting to note that there are relatively few accidents in the transportation, process control and medical domains directly attributed to automation compared with the millions of people who everyday experience inconvenience, frustration, lost data and even deceit (Degani, 2004) at the hands of software automation.

A SOFTWARE AUTOMATION TAXONOMY

Automation now pervades personal computer software and operating systems. In fact, both Microsoft's "plug and play" concept and their recent Windows XP™ operating system are founded on advances in software automation. Automation serves many useful purposes in today's personal computers, and on the Web. Table 1 below is our first attempt at a taxonomy of common functions of software automation. For each category of automation, we provide an example or two from typical software and Web applications as well as operating systems.

SOFTWARE AUTOMATION PROBLEMS

As is the case in other domains, automation is not a panacea in consumer software. In fact, one could argue that for every beneficial function of automation the average user is plagued with an equal or greater number of automation surprises (Sarter, Woods and Billings, 1997) and pitfalls.

In the process of cataloging various software automation problems, we were encouraged to see that many of the automation issues and problems that have been identified, described and exemplified in the aviation domain (see Lyall and Funk, 1998) apply directly to the consumer software domain. Below are some example problems, which share many of the same descriptors as found on the Flight Deck Automation Issues Web site (http://www.flightdeckautomation.com/fdai.aspx).

Table 1. Software Automation Taxonomy.

General Function	Examples
Auto Memory	Software can remember and recall information for users based on their previous actions with a system. Examples of automation memory include hyperlinks on a web site, browser history listing of all web sites visited within a period of time, re-launching an application and having it recall the size and/or position of the window, and the ability to retain user preferences within a software application.
Auto Completion	Auto completion occurs when the software completes all or part of the user's required input. Airline reservations web sites are a good example of auto completion. Upon selecting the month and date of one's departure, the software automatically adjusts the return month and date within a logical travel period *after* the selected departure date. This form of automation is also witnessed within desktop software, for example when a word processing application automatically inserts the current date as you attempt to manually type it in.
Auto Format	Another form of automation is in the default formats assumed by most software applications. Users often rely on software automation to choose the best design, layout, arrangement, or configuration, or to apply a particular format based on user's preceding actions.
Auto Decision	Software automation is constantly making decisions for the user. An example of auto termination is found in online banking. When users are logged into their bank account, with no activity for a period of time, the system will recognize the lack of activity and log users out of their account as a safety precaution.
Auto Configuration	Automation configuration occurs when a computer can recognize new components added to the system and seamlessly install required components without user input. The "plug and play" capability of an operating system, such as Windows XP™ exemplifies this category of software automation. Users no longer have to insert a disk and find a driver to set up a new printer, mouse, scanner, etc. This concept extends to software as well, as it is common practice for new programs to be installed and configured with little user involvement.
Auto Process	Here, the software initiates a process automatically rather than requiring the user to manually intervene. Examples of automated processes include the auto run feature used to launch an installation when a CD is inserted into a computer or the automatic virus scanning of a document that has been attached to an email message. Perhaps the most covert of automated processes is the automatic downloading and installing of software updates; these often occur without user involvement or awareness.

Complexity

Some very complex computer processes have been seemingly simplified through the use of wizard interfaces. Hiding these complexities can lead to unexpected behaviors and make the task of manually interacting with these processes more difficult. A good example of this problem stems from the Network Connection wizard found in Windows XP™. If your network setup matches one of the pre-defined configurations then the wizard is likely to successfully automate the process of connecting your computer to the network. On the other hand, if you fall into the "other" category (Figure 1, right image), then the process of manually configuring the network connection is actually much more difficult compared to previous, less-automated systems.

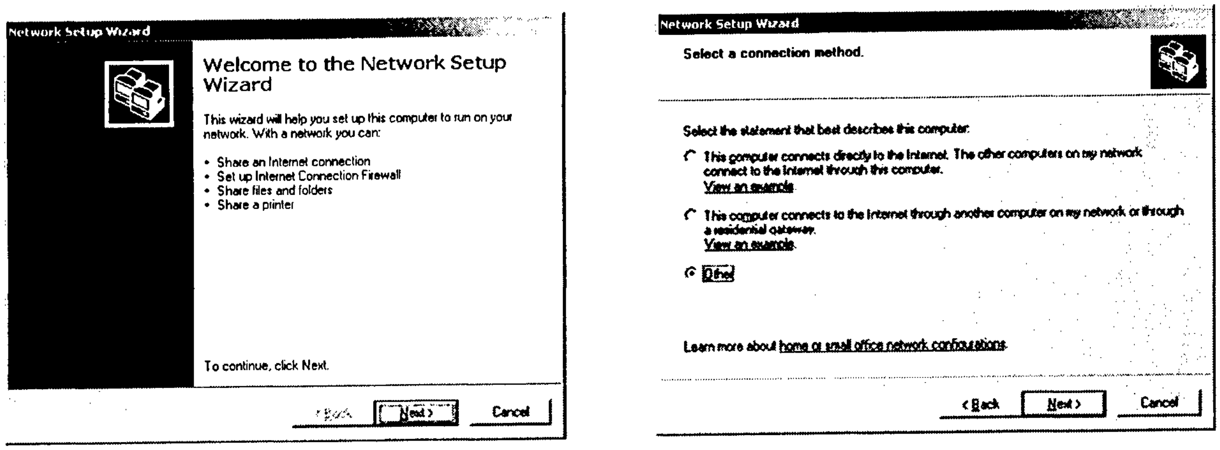


Figure 1. Windows XP™ Network Setup Wizard.

Transparency

The interface for automated functions is often not transparent enough for the user to either find a way to change or optimize the behavior of the automation, or to understand the implications of different automation options and settings. Figure 2 shows a screen from a popular Internet security application. In this example, the application has informed the user of a remote system attempting to access the computer without authorization. The problem lies in the opacity of the options for addressing the situation. How is the average user expected to understand the implication of the suggested action, stated as "Manually configure Internet Access"?

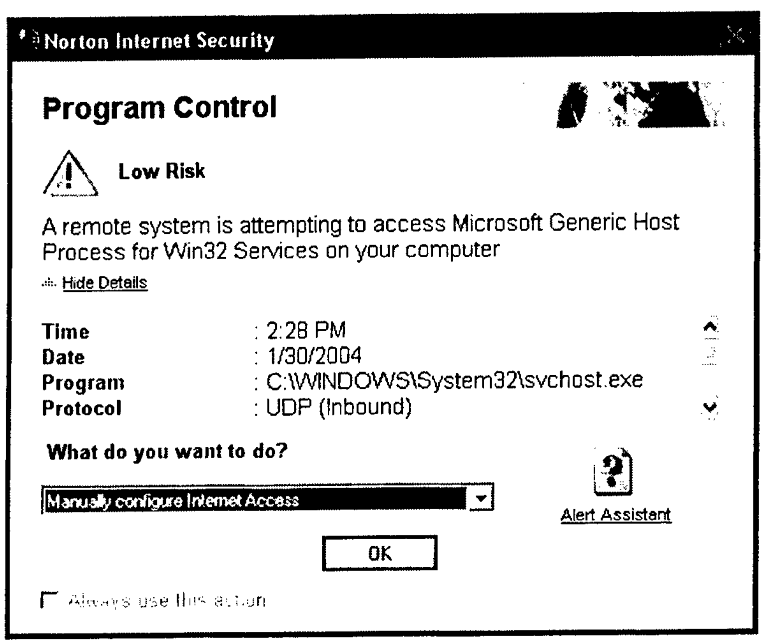


Figure 2. User response options to automated security notice are not transparent.

Under-Trust

We can quickly develop a lack of trust (under-trust) when we don't perceive the benefits of some automated tools. A good example is the fact that many computer users do not employ virus detection and firewall protection applications. This under-trust, coupled with a lack of understanding of how viruses and worms automatically spread across computers, results in millions of dollars of damage and countless hours of lost productivity each year.

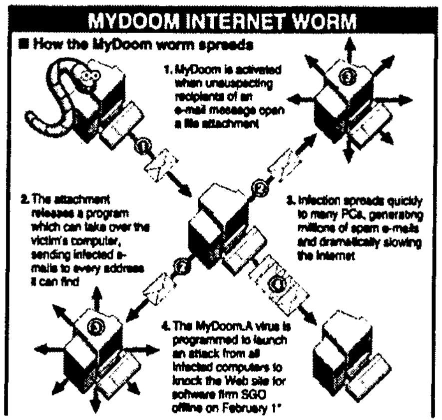


Figure 3. Graphical Depiction of the Computer Worm Spreading Process.

Over-Trust

Sometimes we trust software automation to make intelligent decisions on our behalf. However, this can have drastic consequences. The "chart wizard" in Microsoft Excel™ utilizes default properties that often result in both an unusable and ugly chart, and typically requires the user to manually intervene to change settings, remove unwanted elements, add titles, etc.

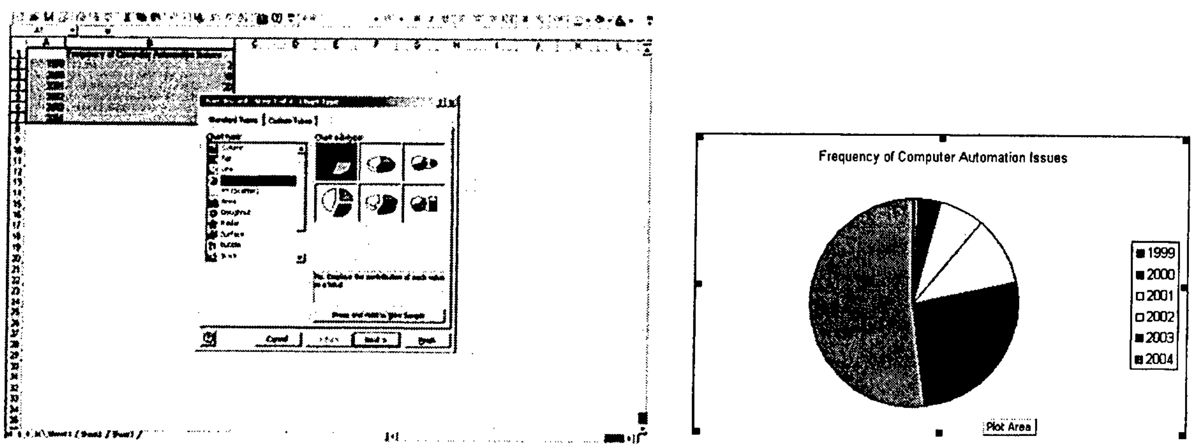


Figure 4. Automated chart wizard produces a poorly designed pie chart.

Mode Awareness

Mode awareness issues (Sarter and Woods, 1995) are perhaps the most common automation problem in consumer software. Mode issues can even surprise the user who carefully takes the time to change application defaults and to configure an application to behave in a specific manner. A great example of this problem can be seen in the automatic font selections in Microsoft Powerpoint™. Assume you bother to change the default font settings in the "slide master" function, from Times New Roman to Arial. If you enter text directly into the slide template it will sure enough be in the set font of Arial. If, however, you add text using the seemingly redundant text tool it appears in the (original) default font of Times New Roman. Unbeknownst to the user, font changes made in one mode have no effect on the other.

Privacy

The other side of the beneficial attribute of automation memory is the problem of privacy. This is especially relevant to households where more than one person use a particular computer. Any user can see what Web sites the previous user has visited, what products they may have shopped for, which documents they recently deleted, and so on.

Deceit

Yes, automation can even be used to deceit computer users! Degani (2004) describes the now ubiquitous banner ads and automatic pop-up windows that capitalize on unassuming Web users. These applications use embedded automation to keep open browser windows, redirect the user to specific Web sites and download dialer programs, among other unsolicited actions.

DESIGN GUIDELINES

General human factors automation guidelines exist for complex systems, many of which apply to the design of software automation interfaces. We recommend the following guidelines, expanded from those provided by Wickens and Hollands (2000), for the design of automation interfaces for consumer software and Web applications:

- Keep the user informed.
- Make the automation logic transparent to the user.
- Introduce automation gracefully.
- Make automation flexible
- Make automation predictable
- Provide direct access to automation settings.
- Allow for quick reversals.
- Inform the user if unsafe modes are manually selected.
- Make automation salient.
- Be consistent with user performance.

SUMMARY

While the topic of automation has been limited to complex systems in the human factors literature, the most common form of automation is exhibited by the ubiquitous personal computer. Many desktop applications use automation to assist users in completing everyday tasks. Businesses are also making an effort to migrate users to automated on-line services for banking, managing investment accounts, and even grocery shopping, touting time and cost savings. Yet, these advances in automation do not come without usability consequences.

Our purpose in writing this paper was to raise awareness of both the promises and pitfalls of consumer software automation and to promote the application of guidelines previously developed for complex systems to this emerging automation domain. We recognize that the relationship between these general guidelines and associated specific interface design techniques can be quite distant and abstract. We therefore encourage future research into automation issues and design strategies unique to the personal computer context.

REFERENCES

Billings, C.E. (1996). Aviation Automation: The search for a human-centered approach. Englewood Cliffs, NJ: Lawrence Erlbaum Associates.

Degani, A. (2004). Taming Hal: Designing Interfaces Beyond 2001. New York: Palgrave Macmillan.

Lyall, B. & Funk, K. (1998). Flight Deck Automation Issues. In M.W. Scerbo & M. Mouloua (Eds.) Proceedings of the Third Conference on Automation Technology and Human Performance. (pp. 288-292). Mahwah, NJ: Lawrence Erlbaum Associates.

Microsoft Corporation (1995). The Windows Interface Guidelines for Software Design: An Application Design Guide. Columbus, OH: Microsoft Press.

Sarter, N.B. & Woods, D.D. (1995). "How in the world did we ever get into that mode?" Mode Error and Awareness in Supervisory Control. Human Factors, 37(1), 5-19.

Sarter, N.B., Woods, D.D., and Billings, C.E. (1997). Automation Surprises. In G. Salvendy (Ed.), Handbook of Human Factors and Ergonomics (2nd edition) (pp. 1926-1943). New York, NY: Wiley.

Scerbo, M.W. & Mouloua, M. (1999). Automation technology and human performance. Mahwah, NJ: Lawrence Erlbaum Associates.

Wickens, C.D. & Hollands, J.G. (1999). Engineering psychology and human performance (3rd Edition). New Jersey: Prentice Hall.

INFORMATION SHARING, TRUST, AND RELIANCE – A DYNAMIC MODEL OF MULTI-OPERATOR MULTI-AUTOMATION INTERACTION

Ji Gao and John D. Lee
The University of Iowa

ABSTRACT

Flow management coordinates and integrates flows between sources, sinks, and reservoirs. It describes domains as diverse as supply chain management and power grid management. It typically involves many operators using many decision aids, linked by various degrees of overlapping information and control. Cooperation between operators and appropriate reliance on automation are critical in flow management. Little research has addressed the factors affecting reliance on automation in the multi-operator multi-automation situation that characterizes flow management. The reliance on automation could influence the balance between cooperative and competitive strategies adopted by the operators. This paper investigates the interaction between the reliance and the cooperation, particularly the role of sharing automation information. We extended Decision Field Theory model (Busemeyer & Townsend, 1993) (EDFT) to investigate how the dynamics of trust and reliance depend on information sharing. We also used a game theoretic perspective to describe a two-supplier one-retailer supply chain that affords cooperation and competition. This game situation is linked with the EDFT model to explore the interaction between reliance on automation and the strategy adoption. Simulation results show that sharing information makes reliance more appropriate and promotes cooperation, compared to the situation with no information sharing. These simulation results help define experimental conditions that can validate and extend the model.

Keywords: Trust, Reliance, Information Sharing, Decision Field Theory, Game Theory, Supply Chain Management, Multi-operator Multi-automation

INTRODUCTION

Inappropriate reliance on automation has contributed to numerous industrial disasters and these disasters will become increasingly costly and catastrophic as automation becomes more prevalent (Lee and See, in press). For a multi-operator multi-automation (MOMA) system, the cooperation between operators is another critical factor for the successful system operation. The interaction of inappropriate reliance on the automation and poor cooperation between operators may be a very important determinant of system performance and has received little attention.

Flow management is a general domain in which MOMA performance is particularly important. Flow management coordinates and integrates flows between sources, sinks, and reservoirs (e.g., materials, information, and power) describing domains as diverse as conventional supply chain management and power grid management. A linked structure of multiple flows and reservoirs defines a network that multiple operators manage with the support of multiple elements of automation (e.g., decision aids). More than single-operator situations, poor coordination between operators and inappropriate reliance on automation can degrade the decision making performance and lead to catastrophes. As an example, the worst power grid failure in the nation's history occurred on August 14, 2003. In this failure, the flow of approximately 61,800 megawatts of electricity was disrupted, leaving 50 million customers from Ohio to New York and parts of Canada without power (Lipton, Pena & Wald, 2003). An important contribution to this event was a lack of cooperation between two regional electrical grid operators that monitor the same region (U.S.-Canada Power System Outage Task Force, 2003). These operators manage flow of the electricity from suppliers to distributors. Poor communication and a failure to exchange detailed information on their operations prevented them from understanding and responding to changes in the power grid. In contrast, cooperation between two operators may improve not only the performance of each but also the successful operation of the whole system. Similar failures in flow management occur in supply chains as well as petrochemical processes where people and automation sometimes fail to coordinate their activities.

Little research has addressed the interaction of operators' reliance on automation and the cooperation between operators that characterizes flow management. For example, in a two-supplier one-retailer supply chain system where both suppliers provide the same products to the retailer. There is a joint production rate that maximizes the joint profit of the suppliers and exceeding this rate will undermine the supplier's profits. Suppliers can either cooperate and coordinate their production rate to maximize their joint profit or they can compete and try to maximize their individual profits. Deciding to cooperate or compete depends on understanding the intent of the other supplier: cooperating when the other competes could greatly undermine the profit of the cooperating supplier. The appropriateness of the supplier's reliance on automation may influence the actual production rate and thereby influence the decision of the other operator to adopt either a strategy to cooperate or to compete. The interaction between the reliance on automation and the cooperate/compete strategy in a MOMA situation is complicated and unexplored. This paper examines the role of information sharing in such an interaction. Computer-based models can help describe the complex interactions between operators as well as between operators and automation. In particular, we use a computational model of reliance on automation coupled with a model of the cooperate/compete relationship to explore factors affecting flow management performance.

A MODEL OF MULTI-OPERATOR MULTI-AUOMATION (MOMA)

Extended Decision Field Theory (EDFT) to describe Operator's Reliance on Automation

Decision Field Theory (DFT) provides a rigorous mathematical framework to understand the motivational and cognitive mechanisms that guide the deliberation process involved in decisions under uncertainty (Busemeyer & Townsend, 1993). DFT differs from most decision-making approaches by being stochastic and dynamic rather than deterministic and static (Townsend & Busemeyer, 1995). However, DFT does not consider the effect of previous decisions in the context of multiple sequential decision process. Moreover, DFT cannot be applied to the multi-person situation directly. Therefore, DFT was extended to consider the multiple sequential decision problems in a MOMA context (Gao & Lee, 2003).

The extended Decision Field Theory (EDFT) links the sequential decision processes by dynamically updating the beliefs of automation or manual capabilities based on the previous experiences to guide the next decision. The belief is updated as:

$$B_C(n) = \begin{cases} B_C(n-1) + 1/b \cdot (C(n-1) - B_C(n-1)) & \textit{if } C(n-1) \textit{ is available} \\ B_C(n-1) & \textit{otherwise} \end{cases}$$

Where B_C represents the belief (estimation) of the automation capability (B_{CA}) or manual capability (B_{CM}), C denotes the true capability and b represents how much the latest experience affects the estimation. The evolution formula of preference in DFT is applied to trust and self-confidence (Busemeyer & Townsend, 1993). The preference towards automatic or manual control is defined as the difference between trust and self-confidence and the decision to rely on automation or intervene is made once the preference evolves beyond a threshold, θ. This dynamic model of trust, self-confidence, and reliance replicates several empirical phenomena including the tendency to adopt an all or none reliance strategy and the tendency of reliance to have inertia (Lee and Moray, 1994). EDFT provides a well-defined computational structure to operationalize the conceptual model of trust, self-confidence, and reliance on automation (Lee & See, 2003). Figure 1 shows how this model describes the dynamic close-loop relationship between the context and operator's decision to rely on automation and Gao and Lee (2003) describe the model behavior and define the model parameters.

Game Theoretic Description of Cooperation in MOMA

A critical element of flow management concerns the cooperative or competitive strategies adopted by the

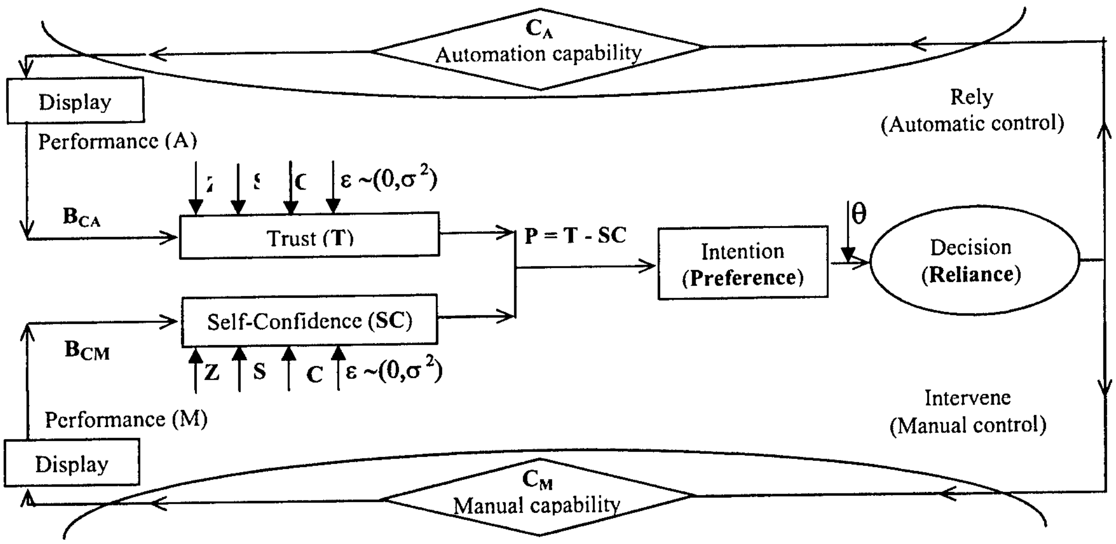


Figure 1. Conceptual model of EDFT for operators' reliance on automation.

operators. Game theory provides a useful formalism to investigate the dynamics of cooperative relationships. As a well-known example of game theory concepts, The Prisoner's Dilemma describes a game situation where two suspects can either confess or not confess, when captured by the police. The optimal decision depends on the decision of the other suspect and can be defined according to payoff matrix (Von Neumann & Morgenstern, 1944). Game theory has become an essential tool in the analysis of supply chain management, which is a system composed of multiple-agents, often with conflicting objectives (Cachon and Netessine, 2003). With respect to the information sharing in supply chain, one firm may have a better forecast of demand than another firm or possess superior information regarding its own costs and operating procedures. The information sharing status often accompanies a game situation that has been described by Cochon and Larivier (1999, 2001) using models of one-supplier one-manufacturer and one-supplier two-retailer supply chains. Although an explosion of game-theoretic papers has been found in the recent supply chain management literature, most only focus on non-cooperative static games (Cachon and Netessine, 2003). Also, these researchers have only considered games with complete information, in which the players' strategies and payoffs are known to all players. In a MOMA situation such as a simple case of two-supplier one-retailer SC system, the decisions of strategy are made over time and the players' strategies and payoffs may not be fully known to all players, therefore it characterizes a dynamic game of incomplete information. The players make decision simultaneously in multiple periods and this type of dynamic game has not been addressed in supply chain management situations (Cachon and Netessine, 2003). No research has addressed the interaction of the game-theoretical description of operators' cooperation strategy and the operators' reliance on automation in a MOMA situation.

The MOMA system used in this paper is a two-supplier one-retailer supply chain system and Table 1 shows a payoff matrix for this situation. The payoff is defined as the product of unit price and the actual production rate and the price is inversely proportional to the joint product rate. Each cell in the matrix shows the payoff for Supplier-1 on the left and Supplier-2 on the right. For example, if both cooperate then both receive a payoff of 50. Based on the payoff matrix, the supplier would choose to compete to maximize individual payoff if the other supplier is assumed to also compete and so both compete and receive a relatively low payoff but not as low as if an individual tries to cooperate when the other competes, which is similar to Prisoner's Dilemma.

Table 1. Two-supplier one-retailer supply chain payoff matrix.

(Supplier-1, Supplier-2)		Supplier-2	
		Cooperate (Target $\dot{P} = \dot{P}_{opt}/2$)	Compete (Target $\dot{P} = \dot{P}_{opt}/2 + d$)
Supplier-1	Cooperate (Target $\dot{P} = \dot{P}_{opt}/2$)	(50, 50)	(35, 63)
	Compete (Target $\dot{P} = \dot{P}_{opt}/2 + d$)	(63, 35)	(40, 40)

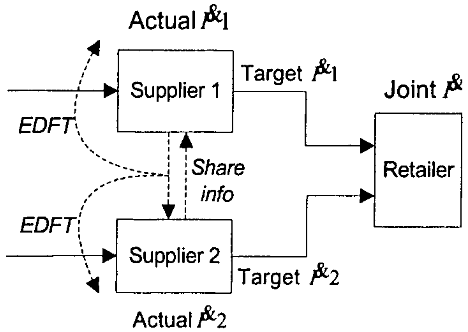


Figure 2. A simple MOMA example.

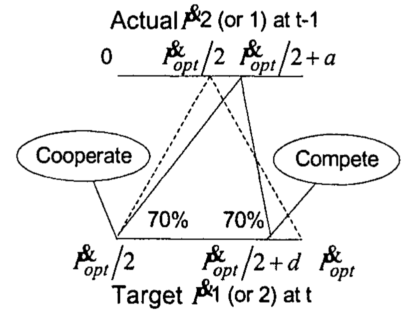


Figure 3. Target determination ($a = 0.1\dot{P}_{opt}$, $d = 0.4\dot{P}_{opt}$).

This two-supplier one-retailer supply chain system is shown in Figure 2. In the game situation given by this structure, two suppliers can choose either to cooperate or to compete and the suppliers make the strategy decision simultaneously without knowing the decision of the other. The strategy to cooperate or to compete is defined as the individual target production rate (Target $\dot{P}$). The optimal joint Target $\dot{P}$ that maximizes the joint profit is denoted by $\dot{P}_{opt}$. Choosing $\dot{P}_{opt}/2$ ($\dot{P}_{opt} = 100$ is used) as Target $\dot{P}$ and intending to maximize joint profit while taking risk being taken advantage of by the other supplier is defined as Cooperate. Choosing a relatively high Target $\dot{P}$ ($\dot{P}_{opt}/2+d$, $d = 0.4\dot{P}_{opt}$ is used) and intending to maximize individual profit by undermining the other supplier's profit is defined as Compete. The supplier's individual Target $\dot{P}$ for the next period is determined by the other supplier's actual production rate (Actual $\dot{P}$) and the correspondence between them is depicted by the solid triangle in Figure 3. The mapping from Actual $\dot{P}$ of the other supplier to Target $\dot{P}$ defines the choice to cooperate or to compete, specifically, when the other supplier's Actual $\dot{P}$ is lower than a cooperation threshold ($\dot{P}_{opt}/2+a$, $a = 0.1\dot{P}_{opt}$ is used), the supplier will cooperate. Otherwise, the supplier has 70% of chance to compete and 30% of chance to cooperate (we assume that the suppliers always tend to cooperate since they realize both cooperating will achieve global optimal). In this may the past behavior of one supplier influence the decision of the other supplier to compete or cooperate.

The choice of Target $\dot{P}$ defines the decision to compete or to cooperate, but it does not completely determine the Actual $\dot{P}$. The Actual $\dot{P}$ depends on the appropriateness of reliance on automation. Inappropriate reliance makes it unlikely to achieve the Target $\dot{P}$. Specifically, the Actual $\dot{P}$ fluctuates around the Target $\dot{P}$ with a variance that depends on the use of automation. In this may inappropriate use of automation break the mutual trust between suppliers. Even when the supplier intends to cooperate, the Actual $\dot{P}$ suggests he is competing to the other supplier. In contrast, appropriate reliance makes it easier to signal cooperation and reach a 'Win-Win' situation because the Actual $\dot{P}$ reflects the supplier's intention correctly. This is why the interaction between the Cooperate/Compete strategy and the appropriateness of reliance on automation becomes important.

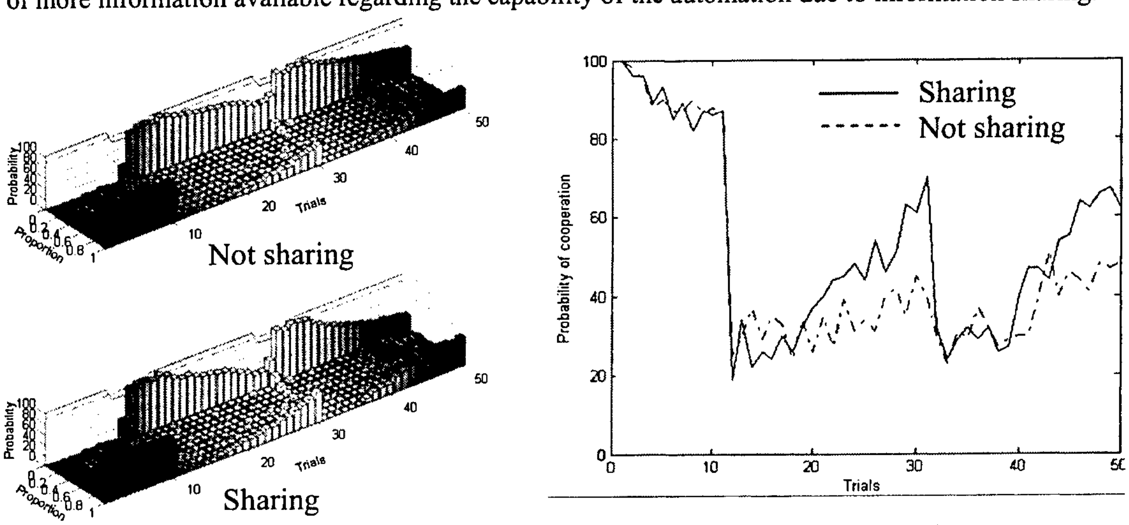
EDFT and Game Theory to Describe MOMA

Sharing information regarding the reliance on automation might have two influences: it might improve reliance on automation and it might help operators understand the intent regarding the other operator to compete or to cooperate. For example, knowing that the other operator was replying on the automation when the actual production rate suggests a competing strategy might lead to a more chartable interpretation of the behavior. Within the scope of this paper, we only examine the influence of improving reliance on automation.

Improving the operators' reliance on automation by sharing the information of the use of automation is implemented in the EDFT model. The information available to an operator regarding the performance of the automation is only available when relying on the automation. Therefore, in the situation where the operator chooses the manual control, the operator is unable to accurately assess the current capability of the automation. In a MOMA system, information regarding the capability of automation might be available if one operator adopts the automatic control and shares his information with other operators. With such information, the operator using manual control can better estimate the capability of automation to rely and intervene more appropriately. The connection between the information sharing, appropriateness of use of automation, and the actual production rate is depicted in Figure 2 by dashed lines and Italic texts to show the role of EDFT model in the MOMA system.

RESULTS AND DISCUSSION

The influence of sharing automation information on the Cooperate/Compete strategy is shown in Figure 4. Figure 4a shows the time-varying distribution of operators' reliance on automation predicted by the EDFT model. Total 50 sequential trials are used and the proportion of reliance represents the amount of time spent in automatic control during each trial (e.g., 0.2 represents 20% of time spent in automatic control during the trial). The vertical coordinate corresponds to the number of operators (total 100) who adopted each the various levels of reliance for each trial. The solid and the dashed curves on the vertical surface represent the capabilities of automatic and manual controls and the drop of the automation capability characterizes the occurrence of automation faults. Automation faults happen during trials 11 to 15, where it returns to normal and then fails again during trials 31 to 35, and then returns to normal afterward. Figure 4a shows that more people return to automatic control when the automation returns to normal after the faults when the information is shared compared to that without information shared. It is reasonable because the system is more transparent in terms of more information available regarding the capability of the automation due to information sharing.



a. Distribution of reliance b. Probability of cooperation

Figure 4. Influences of sharing information on reliance and Cooperate/Compete strategies.

Figure 4b shows the probability of cooperation for one supplier for situations in which information is shared and not. The supplier starts with cooperate strategy, but when the automation faults occurred, the probability of cooperating dropped dramatically. This occurred because the inertia of trust and reliance led the supplier to inappropriately rely on the automation, which leads to competitive actual production rates. The other supplier therefore becomes more likely choose to compete and then both are more likely to compete as a result. After the automation returns to normal, the cooperation increases, but only when the information is shared. One explanation is that the supplier senses the automation capability changes more quickly when information is shared and therefore is more likely to rely on the automation appropriately.

CONCLUSION

Supply chain management and flow management, more generally, represent domains where understanding the factors influencing individual operators to rely on automation is not sufficient to understand the joint behavior of the multi-operator multi-automation system. Computational models using EDFT and game theory offer promising methods to enhance our understanding of these complex systems. The simulation results imply that sharing information regarding the performance of the automation can lead to more cooperation because it promotes more appropriate reliance on automation, which reduces unintentional competitive behavior. Empirical data is needed to assess how well the model represents the MOMA behavior. A supply chain management microworld is under development and the experiments will examine the contribution of information sharing in promoting appropriate reliance and cooperation. These experiments and subsequent model revisions will improve the understanding of the complex dynamics of MOMA systems.

REFERENCES

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic cognition approach to decision making. Psychological Review, 100, 432-459.

Cachon, G., & Lariviere, M. (1999). Capacity choice and allocation: strategic behavior and supply chain performance. Management Science, Vol.45, 1091-1108.

Cachon, G., & Lariviere, M. (2001). Contracting to assure supply: how to share demand forecasts in a supply chain. Management Science, Vol.47, 626-646.

Cachon, G. P., & Netessine, S. (to be published). Game theory in supply chain analysis. In D. Simchi-Levi, S. D. Wu, & Z. J. Shen (Eds.), Supply Chain Analysis in the eBusiness Era (Chapter 1). To be published by Kluwer http://www.ise.ufl.edu/shen/handbook/

Gao, J., & Lee, J. D. (2003). Extend the Decision Field Theory to model operators' reliance on automation for supervisory control. Working document.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. International Journal of Human-Computer Studies, 40, 153-184.

Lee, J. D., & See, K. A. (in press). Trust in computer technology: Design of appropriate reliance. Human Factors.

Lipton, E., Pena, R. P., & Wald, M. L. (2003). Overseers missed big picture as failures led to blackout. New York Times, 9/13/03.

See, K. A. (2002). The effect of sound on performance, reliance, decision making, and trust in semi-automatic process control systems. M.S. Thesis.

Townsend, J. T., & Busemeyer, J. R. (1995). Dynamic Representation of Decision Making. In R. F. Port and T. van Gelder (Eds.) Mind as Motion. MIT press.

U.S.-Canada Power System Outage Task Force. (2003). Interim report: Causes of the August 14^{th} blackout in the United States and Canada.

Von Neumann, J., & Morgenstern, O. (1944). Theory of Games and Economic Behavior. Princeton: Princeton University Press.

ALARM MISTRUST WITH SIGNAL DURATION AS A CUE FOR ALARM VALIDITY

Corey K. Fallon, Nicolae Nica and James P. Bliss
Old Dominion University, Norfolk, Virginia USA

ABSTRACT

The researchers examined the effects of short and long duration alarms and system reliability (60 or 80 percent reliable) on participant response frequency and perception of signal validity. The researchers sampled 45 Old Dominion University psychology students. We predicted that participants would rate long duration alarm signals as more representative of a valid signal. We also believed that participants would respond to significantly more long duration signals regardless of system reliability. The results supported our hypothesis. Participants rated the long duration signals as significantly more representative of a valid signal ($p<.001$). Participants reported that the signal duration influenced their response decision significantly more than the system reliability ($p = .01$). Also, participants responded significantly more often to long duration alarms ($p<.001$). Although further research is needed to support these findings, it appears that designers of complex systems can increase alarm response frequency by designing systems that generate long duration alarm stimuli.

Keywords: Alarm, Duration, Heuristic, False, Trust, Warning, Alert, Reaction

INTRODUCTION

Many of today's alarm systems frequently generate false alarms that often lead to a degradation in responding known as the Cry Wolf Effect (Bliss, 1993; Breznitz,1984). Researchers have begun to examine variables that may moderate this effect (Bliss & Dunn, 2000). One factor that may have an impact is the match between alarm stimuli and users' mental representations of a valid signal. Guillaume, Pellieux, Gastres and Drake (2003) recently suggested that mental representations of alarm signals stored in long-term memory affect people's perceptions of incoming stimuli.

Representativeness Heuristic

The influence of mental representations on alarm reaction decisions is suggested by the representativeness heuristic. According to this heuristic, people often diagnose an event based on the match between perceptual information from the event and their knowledge of similar events from the past (Wickens & Hollands, 2000). For example, people are likely to perceive an alarm as valid if their perception of the signal matches their mental representation of a true alarm constructed from past experiences. Research has shown that the representativeness heuristic is robust to other variables that may affect decisions, including overall probability (Fischhoff & Bar-Hillel, 1984).

Goal of this Study

This study was designed to examine the impact of the representativeness heuristic on responses to alarms of different reliability levels. We wanted to examine how participants respond to alarm stimuli from systems with varying degrees of reliability, when those stimuli may be perceived as representative or not representative of a valid signal. We believed that participants would ignore alarm system reliability levels and base their reactions solely on how well the stimuli matched their mental representation of a valid signal. We predicted that participants would use the duration of the alarm signal as a cue for signal validity. Specifically, the researchers believed that participants would use the representativeness heuristic to make their response decisions and as a result ignore short duration signals and respond to the long duration alarm signals. This hypothesis is consistent with research examining the representativeness heuristic and the impact of mental representations on alarm signal perception (Fischoff & Bar-Hillel, 1984; Guillaume et al., 2003).

METHOD

Participants

A power analysis revealed that 40 participants would yield an experimental power of 0.80 at p = .05. To obtain sufficient power, the researchers collected data from 45 Old Dominion University psychology students. The students were 13 males and 32 females of various ages and ethnic backgrounds. They were randomly assigned to high (80% true alarms) and low (60% true alarms) alarm reliability groups. Twenty-one participants were in the low group and twenty-four were in the high. Participants ranged from 18 to 38 years old, with an average age of approximately 21 years. None of the participants reported suffering from hearing loss.

MATERIALS

The laboratory space used for this study consisted of a workstation with a computer containing the gauge monitoring and tracking sub-task from the Multi-Attribute Task (MAT) Battery program (Comstock & Arnegard, 1992). These two sub-tasks comprised the primary task. A second computer with a secondary alarm response program was placed to the right of the participant at a 90-degree angle to the workstation. The alarm response program generated auditory alarm stimuli reflecting two levels of duration. Both the primary task and the secondary alarm response task have been used in previous research (Bliss, Gilson & Deaton, 1995; Bliss & Kilpatrick, 2000).

The signal was a Boeing 757 overspeed siren presented in two levels of duration (one second and four seconds). The alarm system also had a visual component. When an alarm was sounded the signal word "Warning" flashed on the alarm response computer screen for the entire duration of the auditory signal.

Participants also completed background and opinion questionnaires. The background questionnaire was designed to obtain pertinent background information, such as participants' hearing and computer experience. The opinion questionnaire contained 5 point Likert scale items designed to assess how alarm duration and system reliability affected each participant's perception of alarm signal validity. For example, participants were instructed to rate how much the two independent variables (Duration and Reliability) influenced their alarm reaction decisions. In addition, participants were asked to rate the extent to which they believed the long and short duration sounds matched their perception of how an alarm "should" sound.

PROCEDURE

When the participant arrived, he or she received an informed consent form to read and sign. Next the experimenter administered a participant background questionnaire and randomly assigned the participant to either the 60 or 80 percent reliability group. The random assignment was used to maintain a true experimental design (Tabachnick & Fidell, 2001).

Once the experimenter assigned the participant to a group, the experimenter instructed the participant to sit at the computer workstation. At this point participants were told the reliability of the alarm system, either 60 or 80 percent reliable depending on the participant's assignment. Providing the participant with this information prior to the sessions accelerated the onset of Cry Wolf Effect (Bliss, 1993).

Next, familiarization instructions were presented to the participant, which explained how to perform both the primary MAT task and secondary alarm response task. The experimenter allowed the participants to practice the primary task for five minutes without interruption from the alarms. After the practice session the experimenter explained how to react to the alarms. Participants had to use the mouse from the alarm response computer to click on a box in the lower right hand corner of the alarm response computer screen. The box was labeled "R" for respond. If the participant decided that an alarm was false the correct reaction was to simply ignore the alarm and continue with the primary task. Participants did not receive any feedback regarding the correctness of each alarm reaction decision. The participants were also not provided with any information regarding the validity of each individual alarm. The researchers believed that the providing performance feedback and validity information would overshadow any performance effects due to alarm duration and alarm reliability.

All participants in each group participated in three 10-minute experimental blocks separated by 5-minute rest periods. The alarm system presented 10 alarms in each block with 5 long and 5 short duration alarms randomly generated within each block. After the three experimental blocks, the participants were instructed to complete the opinion questionnaire, which contained questions regarding the alarm system and their response strategy.

RESULTS

Response Frequency

Alarm response performance was measured using participant's response frequency. The researchers measured response frequency by calculating the percentage of responses made by the participant in each experimental block. Data from all three experimental blocks were analyzed using 3x2x2 mixed ANOVA. A significant main effect for alarm duration was found, $F(1,43) = 166.76$, $p<.001$, partial = .80. Participants responded significantly more often to long duration stimuli, regardless of system reliability. The researchers also found a significant interaction between duration and experimental block, $F(1,43) = 4.27$, $p=.025$, partial = .09. Participants responded to more short duration alarms in blocks two and three when compared to block one. Figure 1 illustrates the main effect and interaction.

Subjective Measures

The researchers performed a paired samples T-test to see if there was a statistically significant difference between how participants were influenced by each variable. To be consistent with our hypothesis, we expected participants to rate alarm duration as a more influential variable. The test was significant, $t(44) = 2.67$, $p=.01$. Participants believed that alarm duration influenced their alarm response decisions significantly more than alarm reliability information.

The researchers also performed a 2x2 mixed ANOVA to see if signal duration and reliability group had significant effects on perceived validity. The researchers found a significant main effect for duration, $F(1,43) = 73.16$, $p<.001$, partial = .63 (see Figure 2). Participants believed that the long duration signal was a significantly better match with their perception of a valid signal. These results are also consistent with our hypothesis.

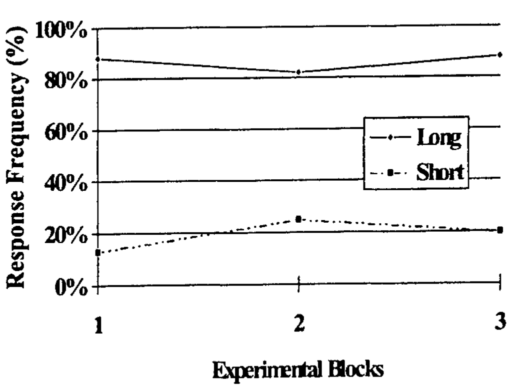


Figure 1. Percentage of Alarm Responses as a Function of Experimental Block and Signal Duration

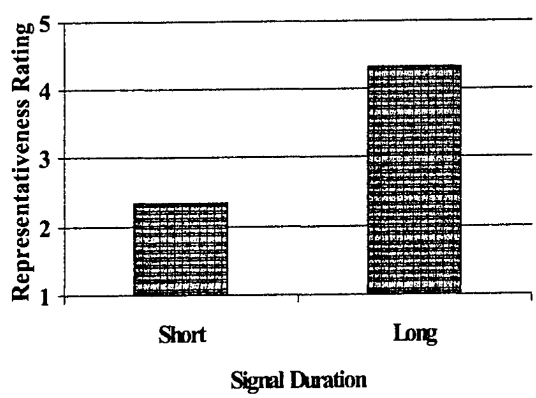


Figure 1. The Extent to Which Signal Duration is Representative of a Valid Signal

DISCUSSION

This study suggests that the duration of the alarm signal is perceived as an important cue for the signal's validity. Specifically, the long duration signal was perceived as more representative of a true alarm. Also, the results provide support for the representativeness heuristic's ability to overpower the Cry Wolf Effect. The response frequency findings and questionnaire data suggest that participants did not incorporate reliability information into their decision making process. The participants based their response decisions almost entirely on the duration of each alarm in all three experimental blocks. These results suggest that the response strategy was not learned over the course of the study. Therefore, this pattern may be based on mental representations of alarm validity stored in long-term memory.

CONCLUSION

This study reveals the representativeness heuristic's power over the decision making process. Rather than incorporate useful reliability knowledge into their decision-making, participants based their decisions on the assumption that signal duration was an indicator of alarm validity. This assumption was made despite the fact that the experimenters never suggested alarm duration as a possible cue for signal validity.

The researchers are currently conducting a follow-up to this study. We will be examining the effects of alarm duration on reaction performance when long and short duration alarms are generated from two separate systems. Comparing the findings from the two studies may help us to better understand the role of alarm duration in the reaction decision-making process. Designers of complex systems can then incorporate these findings to increase the effectiveness of alarm stimuli and overcome signal mistrust by operators.

REFERENCES

Bliss, J. P. (1993). The cry-wolf phenomenon and its effect on operator responses. Unpublished doctoral dissertation, University of Central Florida, Orlando.Bliss, J. P., & Dunn, M. C. (2000). Behavioural implications of alarm mistrust as a function of task workload. Ergonomics, *43,* 1283-1300.

Bliss, J. P., Gilson R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. Ergonomics, *38,* 2300-2312.

Bliss, J.P., & Kilpatrick, F. (2000). The influence of verbal content on alarm mistrust. Proceedings of the 2000 Human Factors and Ergonomics Society Annual Meeting. July 30-August 4: San Diego, CA.

Breznitz, S. (1984). Cry-wolf: The psychology of false alarms, (pp. 1-100) Hillsdale, NJ: Lawrence Erlbaum.

Fischoff, B. & Bar-Hillel, M. (1984). Diagnosticity and the base-rate effect. Memory & Cognition, *12* (4). 402-410.

Guillaume, A., Pellieux, L., Chastres, V. and Drake C. (2003). Judging the urgency of nonvocal auditory warning signals: Perceptual and cognitive processes. Journal of Experimental Psychology: *Applied, 9.* 196-212.

Tabachnick, B. G. & Fidell, L. F., (2001). Computer-Assisted Research Design and Analysis (pp. 1-24) Needham Heights, MA: Allyan & Bacon.

Wickens, C. D. & Hollands, J. G. (2000). Engineering Psychology and Human Performance, (pp. 1-14) Upper Saddle River, NJ: Prentice Hall.

EFFECTS OF VARYING THE THRESHOLD OF ALARM SYSTEMS ON HUMAN PERFORMANCE

Ernesto A. Bustamante, James P. Bliss
Old Dominion University

ABSTRACT

The purpose of this research was to investigate the effects of varying the threshold of alarm systems on human performance. Using Signal Detection Theory, a common Receiver Operating Characteristic (ROC) curve was selected to reflect the sensitivity of the system. The threshold of the system was manipulated by changing the value of beta along the ROC curve. Sixty-six participants performed a compensatory tracking and a monitoring task with or without the aid of the system. Measures of performance included root mean squared error on the tracking task and overall reaction time (ORT) on the monitoring task. Also, alarm reaction time (ART) was calculated for groups using the alarm system. Results indicated greater performance for groups using the system. Furthermore, ART was faster for the group using the system with the highest threshold. Lastly, although differences in ORT between the groups using the system were not statistically significant, a means plot analysis revealed a trend in the shape predicted.

Keywords: Alarm Systems; Signal Detection Theory; Human Performance; Reaction Time

INTRODUCTION

Technological advances have enabled highly sensitive alarm systems to detect the presence of imminent danger. However, the majority of alarm systems are unreliable (Getty, Swets, Pickett, & Gonthier, 1995; Parasuraman & Hancock, 1999). Researchers have tried to determine why this is so. Getty et al. (1995) and Parasuraman and Hancock (1999) analyzed this problem using Signal Detection Theory (SDT). They pointed out that one of the reasons why alarm systems have proven to be so unreliable is because in their effort to detect the occurrence of dangerous events, designers often set the threshold of such systems at a low level. This is what is commonly known as the "engineering fail-safe approach" (Swets, 1992, p. 524). As a consequence, most alarm systems emit a greater number of false alarms than true alarms, which decreases alarm reliability. The major consequence of this decrease in alarm reliability is a loss of trust in alarm signals, a phenomenon commonly known as the "cry-wolf effect" (Breznitz, 1983). This loss of trust, in turn, leads to a reduction in human responsiveness and an increase in reaction time to alarm signals (Bliss, Gilson, & Deaton, 1995; Getty et al., 1995; Parasuraman & Riley, 1997). It seems intuitive to raise the threshold of alarm systems to achieve a lower volume of false alarms and higher reliability. However, raising the threshold of alarm systems increases the chances of not issuing alarms when imminent danger is present. Therefore, the purpose of this study was to examine how changing the threshold of alarm systems affects human performance, taking into account both alarm reliability and probability of missed dangerous events.

In the framework of SDT, the alarm system can be thought of as a detector. Its sensitivity, denoted by d', constitutes its ability to detect the presence of imminent danger. Its threshold, denoted by β, represents the characteristic of its response criterion. A low threshold produces a high number of both hits and false alarms, but a low number of misses. Conversely, a high threshold produces a lower number of both hits and false alarms, but a higher number of misses. It is also necessary to consider the prior probability of imminent danger. For example, a very sensitive alarm system (i.e., $d' = 3.5$) seems very efficient while considering its a priori characteristics (Getty et al., 1995). Given its sensitivity, even if designers set its threshold low enough to achieve approximately 90% probability of a hit, its probability of a false alarm would only be around 1.7%. However, in most cases, the probability of imminent danger is significantly lower than the probability of no danger (Getty et al., 1995; Parasuraman & Hancock, 1999). Getty et al. (1995) argued that a prior probability of 0.1 % is probably realistic for most alarm system situations. Taking this into account, the alarm system mentioned in the previous example would make approximately 1 hit for every 20 false alarms. This means that this alarm system that seemed very effective from the design point of view would only be approximately 5% reliable. This extremely low reliability is what causes humans to decrease their responsiveness to alarm signals. One of the ways to solve this problem is to raise β high enough so that the ratio between true and false alarms will be greater and reliability will be higher. It may seem clear that the more reliable a system is, the more people will be willing to respond to alarm systems. In fact, this phenomenon has become known as probability matching and has been demonstrated by a number of studies (Bliss & Dunn, 2000; Bliss, Gilson, et al. 1995). The problem is that this would greatly increase the probability of misses.

Many researchers have studied the effect of alarm reliability on human trust and responsiveness to alarm signals (Bliss, Dun, & Fuller, 1995; Bliss, Gilson, et al. 1995; Getty et al., 1995). Results from these studies have consistently shown that the low reliability of alarm systems decreases humans' frequency of response to warnings and increases people's reaction time. However, none of these studies has included situations where an alarm system should have issued a signal but failed to. For this reason, the present study examined how alarm reliability affects human performance while taking into account the consequences of varying the threshold of alarm systems. We hypothesized that overall performance on a primary gauge monitoring task would be better for groups using an alarm system than for those that did not. We also hypothesized that setting β high would lead participants to react faster to the gauges than setting β low. Last, we expected that overall gauge reaction time would be fastest for the group using the system with the medium threshold.

METHOD

Participants

Sixty-six (45 females and 21 males) undergraduate and graduate students from Old Dominion University participated in this study. However, one participant was excluded from all analyses because he was an outlier in all measures (more than 3 *SD* from the grand mean). Therefore, data from 65 participants (45 females and 20 males) were analyzed. Participants ranged from 18 to 44 years of age (M = 20.91, SD = 4.48). Experimenters randomly assigned participants to one of four experimental conditions: no alarm system (n = 14), low β (n = 18), medium β (n = 17), and high β (n = 16). Participants received course credit or extra credit as an incentive for their participation.

Materials

Multi-attribute task (MAT). The MAT is a psychomotor task battery that was developed to assess human performance and workload under different conditions (Comstock & Arnegard, 1992). For the present study, only the tracking and gauge monitoring tasks were used. The objective of the tracking task was to keep a ball within a specified rectangular area. Performance on this task was assessed by taking the Root Mean Square (RMS) error of tracking. RMS was measured every second throughout each 20-min experimental session, but only the average of these measures was used for analyses. The objective of the gauge monitoring task was to monitor normal fluctuations of four gauges, two of which indicated temperature changes and two of which indicated pressure changes. When any of these gauges fluctuated out of the normal range, participants had to press the appropriate key to reset it. A total of 1200 normal fluctuations occurred continuously throughout the 20-min session. Twelve out-of-range fluctuations occurred within this period, resulting in a prior probability of .01 for out-of-range fluctuations. Researchers have indicated that this is a realistic value for a number of real-world situations (Getty et al., 1995; Parasuraman & Hancock, 1999). Out-of-range fluctuations occurred randomly throughout each session at a mean rate of 669.17 s and a standard deviation of 379.11 s. Two performance measures were assessed for this task. First, overall reaction time (ORT) was measured in seconds from the onset of the out-of-range fluctuation until participants correctly reset the out-of-range gauge. Since gauges reset automatically after 10 s, this time was assigned to participants who failed to reset a gauge. Second, alarm reaction time (ART) was measured for the groups using the alarm system. This was also measured in seconds from the onset of the out-of-range fluctuation until participants correctly reset the out-of-range gauge, but only for those fluctuations in which an alarm was present.

Alarm system. The alarm system was modeled using SDT. A d' of 3.5 was used to represent the sensitivity of the system. This level of sensitivity was chosen based on previous research by Getty et al. (1995). Three different thresholds were modeled by changing the value of β. The high-β system had a 60% probability of a hit and 0.10% probability of a false alarm, resulting in a total reliability of 88%. The medium-β system had a 75% probability of a hit and 0.20% probability of a false alarm, resulting in a total reliability of 75%. Lastly, the low-β system had a 92% probability of a hit and 1.70% probability of a false alarm, resulting in a total reliability of 35%.

An IBM-compatible computer with an Intel Pentium IV processor and a 17-inch monitor hosted the MAT program. Participants performed the compensatory tracking task with a standard mouse and responded to gauges using a standard QWERTY keyboard. A Macintosh computer running SuperCard 2.5 was used to present the alarm signals. The alarm signals included auditory and visual stimuli presented concurrently at the onset of the out-of-range gauge fluctuations. The auditory stimulus was the overspeed siren of a Boeing 757, presented to participants at 65 dB(A) through a pair of standard speakers for a period of 1.7 s. The visual stimulus consisted of a yellow square with rounded edges with the word "WARNING" written on it. This visual stimulus was presented on a 15-inch monitor for 2 s.

Procedure

Upon arrival, participants first read and completed an informed consent form. Next, they completed a background information form that included demographic and experience items. The experimenters used a standard script to instruct participants how to perform each task. After reading and explaining the instructions, experimenters answered any specific questions that participants had. Participants then completed a one-minute practice session of each individual task. The experimenter then showed participants in the alarm conditions a sample alarm signal. Participants then completed a combined practice session that included the MAT tasks and the alarms (if applicable). During the second practice session, experimenters demonstrated the alarm system's fallibility by pointing out true alarms, false alarms, and misses. After the practice sessions, participants completed their first of two 20-minute sessions separated by a five-minute break. After completing the second session, participants completed an opinion questionnaire, and were debriefed and dismissed.

RESULTS

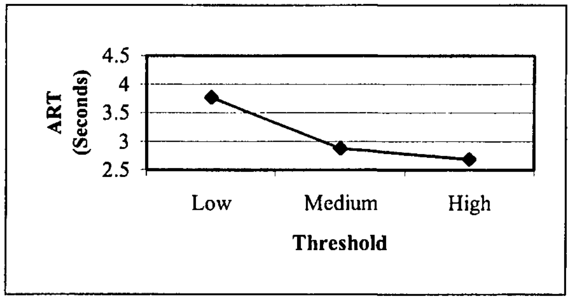
After confirming data normality and covariance matrix equality, a one-way MANOVA was used to test the first hypothesis. Group (no alarm, low-β, medium-β, high-β) was used as the independent variable. Root Mean Squared (RMS) error on the tracking task and overall reaction time (ORT) on the monitoring task were used as the dependent variables. Results indicated that there were non-significant multivariate differences between groups with regard to RMS and ORT. A follow-up one-way ANOVA showed a significant main effect of group on ORT, $F(3,61) = 4.27$, $p < .01$, partial $\eta^2 = .17$, power = .84. Lastly, a Dunnett's post-hoc analysis using the no-alarm group as the contrast group indicated that ORT was slower for the no-alarm group ($M = 5.40$, $SD = 1.65$) than for the low-β ($M = 4.03$, $SD = 1.32$), medium-β ($M = 3.79$, $SD = 1.09$), and high-β ($M = 4.11$, $SD = 1.34$) groups.

A one-way ANOVA was used to test the second hypothesis. Group (low-β, medium-β, high-β) was used as the independent variable, and alarm reaction time (ART) was used as the dependent variable. Results showed a statistically significant main effect of group on ART, $F(2,48) = 4.68$, $p < .05$, partial $\eta^2 = .16$, power = .76. Lastly, a Tukey's HSD post-hoc analysis indicated that ART was faster for the high-β ($M = 2.68$, $SD = .94$) than for the low-β group ($M = 3.77$, $SD = 1.38$). However, there were non-significant differences between the medium-β ($M = 2.88$, $SD = .94$) and any of the other two groups (Fig. 1).

A one-way ANOVA was used to test the third hypothesis. Group (low-β, medium-β, high-β) was used as the independent variable, and overall reaction time (ORT) was used as the dependent variable. Preliminary descriptive analyses indicated that the dependent variable was normally distributed across each group. Results showed a statistically non-significant main effect of group on ART, $F(2,48) = .29$, n.s. Despite this fact, a means plot analysis revealed a trend in the shape predicted (Fig. 2).

DISCUSSION

Results from this study have important implications for the design and implementation of alarm systems. First, consistent with previous findings (Sorkin, Kanowitz, & Kanowitz, 1988), results showed that the use of an alarm system can improve human performance by directing people's attention to a specific task that may require further action. Second, similar to the study by Getty et al. (1995), results from this study suggest that the reliability of alarm systems has a direct effect on the speed with which people react to warning signals. Higher reliability leads to fastest reaction time. This is particularly important in critical areas such as aviation, medicine, and nuclear power, where a difference of a few seconds in human response may have detrimental effects. Lastly, the fact that overall reaction time was not significantly different between the three threshold levels raises an important point to consider while designing alarm systems. Although increasing the reliability of an alarm system by raising its threshold may lead to faster alarm reaction time, this gain may be lost due to the times in which the system fails to draw operators' attention to the presence of imminent danger.



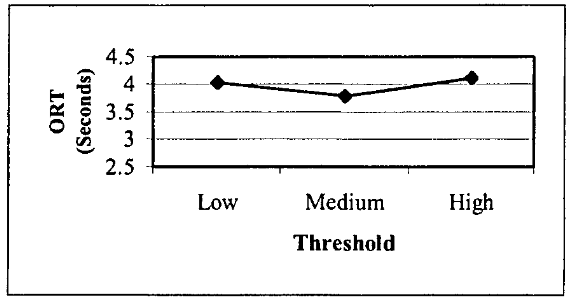


Figure 1. Alarm Reaction Time

Figure 2. Overall Reaction Time

CONCLUSION

When examining the utility of alarm systems, it is necessary to assess the extent to which they aid human performance in complex tasks. Research has shown that performance on such tasks can be improved by the use of alarm systems (Sorkin et al., 1988). Furthermore, this improvement has been greater for more reliable alarm systems (Bliss, Dunn, et al.1995; Bliss, Gilson, et al.1995; Getty et al., 1995). However, as previously pointed out, these studies have not taken into account the effect that missed signals may have on performance. The more reliable alarm systems have a higher probability to fail to issue a warning when danger is present. Therefore, the contribution of high reliability in the form of higher human response frequency and faster reaction time may be counterbalanced by missed dangerous events. Because of this, setting the threshold of alarm systems at extreme levels may not be the best solution. Future research needs to be focused at identifying the optimum alarm system's threshold level to maximize human response efficiency in specific situations.

REFERENCES

Bliss, J.P., & Dunn, M.C. (2000). Behavioural implications of alarm mistrust as a function of task workload. Ergonomics, 43(9), 1283-1300.

Bliss, J.P., Dunn, M., & Fuller, B. (1995). Reversal of the cry-wolf effect: An investigation of two methods to increase alarm response rates. Perceptual and Motor Skills, 80, 1231-1242.

Bliss, J.P., Gilson, R.D., & Deaton, J.E. (1995). Human probability matching behaviour in response to alarms of varying reliability. Ergonomics, 38(11), 2300-2312.

Breznitz, S. (1983). Cry-wolf: the psychology of false alarms. Hillsdale, NJ: Erlbaum.

Comstock, J.R., & Arnegard, R.J. (1992). The multi-attribute task battery for human operator workload and strategic behavior research (NASA Technical Memorandum No. 104174). Hampton, VA: NASA Langley Research Center.

Getty, D., Swets, J., Pickett, R., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. Journal of Experimental Psychology: Applied, *1*, 19-33.

Parasuraman, R., & Hancock, P. A. (1999). Using signal detection theory and Bayesian analysis to design parameters for automated warning systems. In M. Scerbo & M. Mouloua (Eds.), Automation technology and human performance: Current research and trends (pp. 63-67). Mahwah, NJ: Erlbaum.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39, 230-253.

Sorkin, R.D., Kanowitz, B.H., & Kanowitz, S.C. (1988). Likelihood alarm displays. Human Factors, 30(4), 445-459.

Swets, J.A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. American Psychologist, 47(4), 522-532.

THE EFFECTS OF LEVELS OF AUTOMATION IN THE HUMAN CONTROL OF MULTIPLE ROBOTS IN THE ROBOFLAG SIMULATION ENVIRONMENT

Peter N. Squire
Naval Surface Warfare Center

Scott M. Galster
Air Force Research Laboratory

Raja Parasuraman
The Catholic University of America

ABSTRACT

The present study examined the effects of the level of automation available to an operator controlling multiple unmanned vehicles (UVs) while engaging unpredictable opponent postures. Human performance and subjective measures of situation awareness and mental workload were examined using, a simulated multiple UV platform, RoboFlag. There were three automation conditions: manual only, automation only, and a flexible condition in which operators could use both manual control and automated plays. These conditions were factorially combined with three opponent postures, offensive, defensive, or mixed. There were significant effects for performance and subjective measures for both Level of Automation and Opponent Posture, with significant benefits being found for the flexible Playbook condition in comparison to manual only or automation only control. It is concluded that a trade-space exists between mental workload and manual control, and that the Playbook interface provides operators flexibility to adapt to unpredictable situations.

Keywords: Automation, human-robot interaction, mental workload, Playbook, situation awareness, supervisory control, unmanned vehicles

INTRODUCTION

Unmanned vehicles are increasingly being used to support many military and civilian missions involving operations in dangerous or hazardous territory. Having *multiple* UVs under the command of a single human operator may allow mission objectives to be achieved in a cost-efficient manner and also minimize human exposure to threats. Historically, robots and other UVs have conducted work alone in a master-slave relationship, receiving commands directly from operator(s) with little or no autonomous behavior beyond sensing and movement. The lack of greater autonomy in higher-level behaviors can significantly impact an operator's ability to control and monitor more than a single UV. A potential solution to 'single robot parenting' is for UVs to become more autonomous and work in teams (Bruemmer, Dudenhoeffer, & Marble, 2001; Mirmohammad-Sadeghi, Bastani, & Azarnasab, 2003; Ryan, 2003). However, given that completely autonomous operation is not currently technically feasible, human supervision of the robot team is necessary in the face of uncertainty and to allow for the management of unexpected events

Previous publications have discussed different types of control architectures (i.e. teleoperation, trade control, shared control, supervisory control) and design possibilities (Goodrich, Olsen, Crandall, & Palmer 2001; Korenkamp, Bonasso, Ryan, & Schreckenghost, 1997) for human command of autonomous UVs. These discussions have primarily focused on the methods for development of these architectures and their *potential* benefits for human and system performance. Aside from the theoretical nature of these discussions there are few empirical studies of human interaction with multiple UVs, with some exceptions (Crandall & Goodrich, 2003; Dixon & Wickens 2003; Parasuraman, Galster, & Miller, 2003; Ververka & Campbell, 2003). In order to evaluate the various control architectures effectively, the engineering-centered focus needs to be complemented with analysis and modeling of human performance (Adams, 2002; Murphy & Rogers, 2001), so that the probability (rather than the potential) for mission success can be assessed.

A central consideration for control architecture design is determining the appropriate level of flexibility and automation in remote vehicles. The decision on flexibility and level of automation is important because robots can be automated agents with varying levels of autonomy (Parasuraman *et al.*, 2003), and research on human-interaction with automation reveals both benefits and costs associated with particular designs (Parasuraman & Riley, 1997; Sarter, Woods, & Billings, 1997). To understand how the costs and benefits of automation affect a human supervising a team of UVs, empirical data needs to be collected and a framework for interpreting this data needs to be created, allowing designers the ability to identify and select the appropriate flexibility and level of automation to implement in varying control situations. Crandall and Goodrich (2003) examined and evaluated the costs and benefits of various control designs and concluded that a more autonomous interaction scheme (*Scripted)* is more effective than either a *Teleop(Teleoperated)* or a *P2P(Point to Point)* interaction scheme. Because the goal of greater UV autonomy has an upper limit due to technical capability, another possibility is to use a flexible supervisory control architecture in which in which the automation is designed to be *adjustable* or adaptable (Crandall & Goodrich, 2002; Parasuraman, 1993), depending on context. One such architecture is the Playbook delegation concept (Miller, Pelican, & Goldman, 2000), in which human operators can delegate (or not) tasks to automation and autonomous agents at times of their own choosing, and receive feedback on their performance, just as with successful human teams. A playbook interface may allow for effective tasking of robots while keeping the operator in the decision making loop as needed and without increasing mental workload (Miller & Parasuraman, 2002).

Parasuraman *et al.* (2003) examined the effect of environmental uncertainty and unpredictable changes in opponent posture on human-robot performance in the RoboFlag environment. The study demonstrated the effectiveness of the Playbook interface for supervision of multiple UVs and also showed that the RoboFlag simulation environment was a viable platform for gathering empirical evidence related to human supervision of multiple UVs. However, unlike previous research on adaptive or flexible automation (Parasuraman, 1993), a comparison to static delegation was not made in the Parasuraman *et al.* (2003) study. Accordingly, in the present study, we compared Playbook to fixed delegation approaches—either full manual or automation control. We evaluated the effects of these three control types on human-robot team performance under varying adversary "postures" (offensive, defensive, mixed).

We hypothesized that the use of the Playbook interface would afford users maximum flexibility, allowing them to decide when workload was high (and therefore to off-load a task to automation), or when the automation was not effective (and therefore engage in manual control and decrease unpredictability). Additionally, we anticipated that the Playbook interface would allow users the ability to respond more effectively to variable opponent postures than a static control architecture (manual or automated). We tested these hypotheses by measuring overall mission performance indicators (win rate and time to mission completion) and operator mental workload and situational awareness under the different experimental conditions.

METHOD

PARTICIPANTS

Five males and four females between the ages of 19 and 33 (M = 24.00, SE = 1.28 yrs.) served as paid participants. All participants reported normal or corrected to normal vision.

EXPERIMENTAL DESIGN

A within-subjects design was employed, with three Levels of Automation (Manual, Automated, Both) combined factorially with Opponent Posture (Offensive, Defensive, Mixed), yielding nine conditions. Each participant completed five mission trials for each condition, for a total of 45 trials. Level of Automation was treated as a blocked factor while Opponent Posture was randomized within each block. Participants were asked to provide simple mental workload and situation awareness ratings (0, low to 100, high) after each trial, similar to the NASA-TLX (Hart & Staveland, 1988) and the 3-D SART (Taylor, 1990) subjective measure questionnaires.

APPARATUS AND PROCEDURES

Apparatus and procedures were identical to those described in Parasuraman *et al.* (2003) with the exception of the items described below, and a constant robotic visual range. Level of automation was divided into the three most basic control possibilities: manual only, automated plays only, and both (combination of manual and automated plays). In the manual condition, play selection (autonomous robot behavior) was not available to the operator, who had to rely solely on manual (point and click) control. In the automation condition, the operator could select any one of three automated plays available in the Playbook (*circle offense, circle defense, patrol border*) but was unable to use manual control. In the condition where both control options were available, the operator had the ability to choose flexibly between manual and automation control. In addition to varying levels of automation, the opponent's stance/position/configuration varied according to three available scripts: offensive, defensive, or mixed (described in Parasuraman *et al.)*.

Participants were trained by showing them how plays were executed, how robots were selected and moved, as well as how the features of the interfaces showed different robot's status information, fuel, play, and game status. Additionally, they were instructed that the only way a red team red robot could be seen is if they were within the visual range of the blue team robot; otherwise the red team robot was invisible to the blue team operator (see Figure 1). Participants were shown how to retrieve the opponent flag and given a chance to test out RoboFlag without an opponent. Prior to the training trials, participants were given written instructions based on the NASA-TLX and 3-D SART that described how to evaluate and rate their situation awareness and mental workload. Participants completed one trial in each of the nine conditions (with knowledge of the condition) as training prior to the commencement of the data collection trials.

RESULTS

OVERALL PERFORMANCE

The performance data were submitted to a 3 (Level of Automation – manual, autonomous, both) × 3 (Opponent Posture – offense, defense, mixed) analysis of variance (ANOVA). The overall performance metrics included the percentage of games that were won (mission success rate) and the time elapsed for each game (mission completion time). The results of the ANOVA indicated that there was a significant effect of Opponent Posture on the percentage of games won, $F(2,16) = 17.51$, $p < .01$. Expectedly, the participants won 100% of the games when the opponent strategy was defensive and the red team did not make a move to capture the blue team flag. Excluding the defensive condition, there was not a significant difference ($p > .05$) between the offensive and mixed condition where participants won 78% and 79% of the time respectively. No other significant differences were found for the percentage of games won by participants ($p > .05$).

A similar 3 × 3 ANOVA was conducted for the duration of each game (time for mission completion, regardless of win status). Consistent with the results reported by Parasuraman *et al.* (2003), game times were significantly different when participants played against the red team offensive stance ($M = 31.87$s, $SE = 0.51$s) than when they played against the mixed stance ($M = 39.27$s, $SE = 1.79$s) or defensive stance ($M = 103.47$, $SE = 6.44$), $F(2,16) = 56.61$, $p < .01$. The main effect of Level of Automation also showed a significant difference in the amount of time each game took to complete, $F(2,16) = 4.88$, $p < .05$. This difference is illustrated in Figure 2. The longest game times occurred when the participants had only the automation control available ($M = 69.74$s, $SE = 4.72$s) compared to only manual control ($M = 54.06$s, $SE = 5.25$s) and when both type of control were available ($M = 50.81$s, $SE = 4.09$s). These results, coupled with a lack of a significant interaction between the factors, suggest that the participants could complete the mission objective faster when both types of control were available. Moreover, there was a temporal cost associated with having only automation control of the robots without the ability to intervene manually.

STRATEGY USAGE

While each robot state was analyzed, the most interesting results were the differences seen in the experimental conditions where only automation was available compared to the condition where both automation and manual control were available simultaneously. Thus, the percentage of time the robots were commanded to use a particular play (circle offense, circle defense, boarder patrol) was included in a 3 (Opponent Posture) × 3 (Automation play utilized) × 2 (Level of Automation) ANOVA. The results indicate that there was a significant 3-way interaction between these factors (see Figure 3), $F(4,32) = 17.28$, $p < .01$. The interesting finding is the decrease in the use of automated plays between the automation only control condition compared to the both control condition. Further, the pattern of usage was consistent; circle defense was used the most often followed by circle offense and then patrol border. This pattern was true in all cases except the automated only condition when the operator was playing against the red team defensive strategy, in which case, the operator relied more on the use of the circle offense play.

Another indication of strategy utilization is the percentage of time that the robots were under manual control in the manual only condition compared to the condition where both types of control were available. These data were submitted to a 3 (Opponent Posture) × 2 (Level of Automation) ANOVA. The results indicated that there was a significant main effect for the Opponent Posture, $F(2,16) = 38.77$, $p < .01$, and the Level of Automation, $F(1,8) = 8.26$, $p < .05$. Operators used the manual control most often when playing against the red team offensive posture (67.23%) followed by the mixed condition (66.09%) and used manual control least often when playing the defensive red team strategy (53.65%). In the comparison of the percentage of time the robots were under manual control, the operators used manual control 71.67% of the time when only manual control was available compared to 52.98% of the time when both control strategies were available.

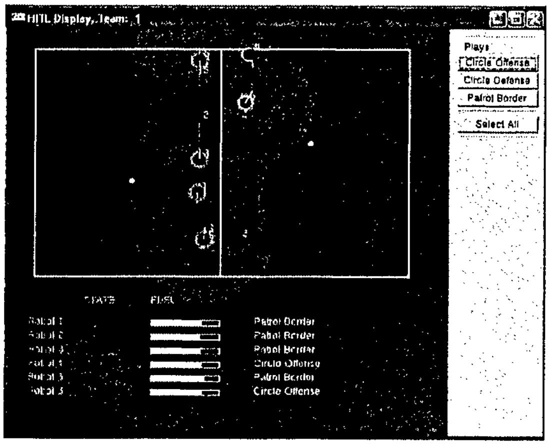


Figure 1. Human-robot Interface (Blue team) (Plays are shown in top right corner)

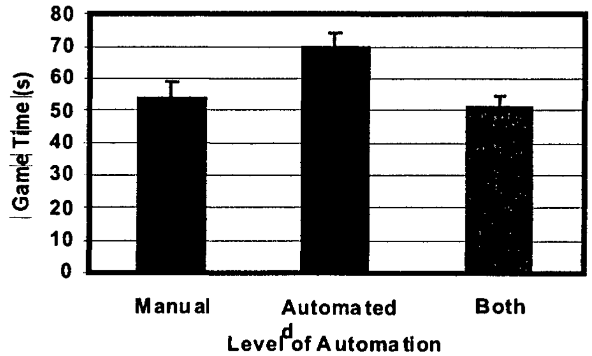


Figure 2. Game Time (s) across Level of Automation

SUBJECTIVE MEASURES

Participants were asked to rate their mental workload and situation awareness after each game (trial) they completed. These ratings were submitted to an analogous 3 × 3 ANOVA as previously described in the overall performance section. For the situation awareness rating, there was a significant main effect for each of the factors – Opponent Posture, $F(2,16) = 5.49, p < .05$, and Level of Automation, $F(2,16) = 7.02, p < .01$. Participants rated their situation awareness highest when the red team status was offensive ($M = 78.30$, $SE = 1.65$) followed by the mixed status ($M = 73.93$, $SE = 1.84$) and reported the lowest rating when playing against the defensive status ($M = 71.19$, $SE = 1.88$). Further, participants reported a higher level of situation awareness for the manual condition ($M = 82.00$, SE 1.27) than automation only ($M= 70.63$, $SE = 2.1$) or both conditions ($M = 70.78$, $SE = 1.78$). For the mental workload rating, there was a significant 2-way interaction between Opponent Posture and Level of Automation, $F(4,32) = 3.92, p < .05$. This interaction, illustrated in Figure 4, indicates that participants rated their mental workload highest when they had both types of controls available. Figure 4 also shows that participants rated their mental workload higher in the manual control over the automated control conditions in all except the mixed red team posture, where the trend was reversed.

DISCUSSION

The shift to a single operator commanding multiple UVs presents a difficult challenge. As technology becomes more capable of robotic teaming but still falls well short of complete autonomy, the *proposed* benefits that automation provides for overcoming 'single robotic parenting' are alluring. However, research in human-interaction with automation presents considerable evidence outlining the costs and benefits of automation. Initial work by Crandall and Goodrich (2003) and Parasuraman *et al.* (2003) have taken the first steps to analyze the affects of controlling multiple robots on human performance. This present study builds on their initial findings around interaction schemes and delegation architectures, while examining the Playbook interface flexibility and how different control architectures influence human performance.

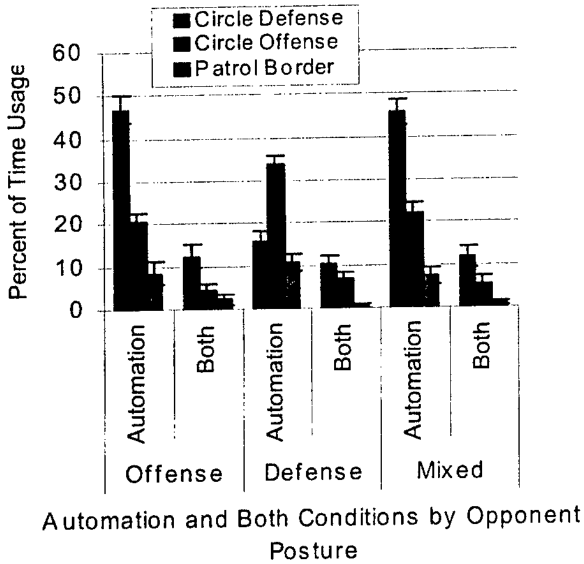


Figure 3. Percentage of Robot Usage for Automation and Both conditions as a function of Opponent Posture

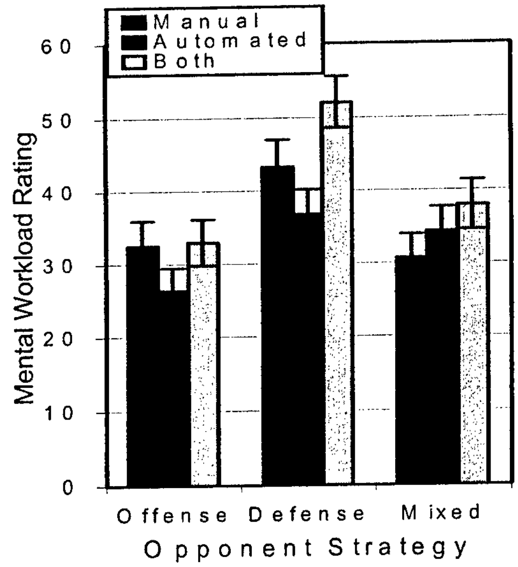


Figure 4. Mental Workload Rating by Opponent Strategy

Several results from the present study are of interest. First, operator usage of the Playbook interface when flexibility was allowed (the "Both" condition) was different than the manual only or automation only condition, as revealed by strategy utilization percentages of manual and automated control. Further, participants were able to effectively use the Playbook interface to adapt to unpredictable opponent postures as revealed by the consistent defensive strategy to oppose forces when they were in an offense or mixed posture, and alternated offensive strategy usage when no opposing forces were sent, defense. Even with the restricted Playbook interface used in this study, participants clearly were able to adapt effectively to the situation, as shown by the high level of competency in game play (win rate > 75%). Manual control allowed participants the ability to overcome ineffective automation movement, decreasing mission completion time. Moreover, participants effectively used the manual control in the "Both" condition, as mission completion time differed from the automation only condition (but not from manual only).

Another proposed benefit to the Playbook interface is the ability to off-load tasks when mental workload is increasing, or increase robotic interaction if the unpredictability of the robots is high. Expectedly, situation awareness was highest in the manual only condition as a result of decreased unpredictability. Increased opponent posture difficulty (indicated by mission completion time) resulted in lower situation awareness for the defensive status condition. Interestingly, in the "Both" condition, participants did not retain the situation awareness benefits of increased robotic interaction, as previously described for mission completion time. This could be due to the increased mental workload seen in this condition, which could have occurred from the cognitive load associated with using the flexibility to decide between when to use automation or manual control.

Our results lead to two important conclusions. Confirmation of a trade-off space between manual control and workload is apparent, as indicated by previous research (Crandall & Goodrich, 2003). In addition, the Playbook interface allows an operator adaptive control and flexibility to determine when automation is ineffective and the ability to switch strategies when needed, as suggested by Parasuraman *et al.* (2003). Although this study provides additional empirical evidence to support previous research, several additional questions are raised that warrant further investigation.

REFERENCES

Adams, J. (2002). Critical considerations for human-robot interface development (Human Robot Interaction Technical Report. FS-02-03). Meno Park, Ca: AAAI Press.

Bruemmer, D. J., Dudenhoeffer, D. D., & Marble, J. (2001). Mixed-initiative remote characterization using a distributed team of small robots, AAAI Mobile Robot Workshop, Seattle, WA, July 2001.

Crandall, J., & Goodrich, M. A. (2002). Principles of adjustable interactions. In Human-Robot Interaction, 2002 AAAI Fall Symposium (A.C. Schultz, Chair). Menlo Park, CA: AAAI Press.

Crandall, J.W., & Goodrich, M.A. (2003). Measuring the intelligence of a robot and its interface. Proceedings of the Performance Metrics for Intelligent Systems Workshop.

Dixon, S.R. & Wickens, C.D. (2003) Control of multiple UAVs: A workload analysis. In Proceedings of the 12th International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Goodrich, M.A., Olsen Jr., D.R., Crandall, J.W., & Palmer, T.J. (2001) Experiments in adjustable autonomy. In Proceedings of the IJCAI-01 Workshop on Autonomy, Delegation, and Control: Interacting with Autonomous Agents. Seattle, WA.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.), Human Mental Workload (pp. 139-183). Amsterdam: Elsevier Science/North Holland.

Korenkamp, D. Bonasso, R.P., Ryan, D., & Schreckenghost, D. (1997). Traded control with autonomous robots as mixed initiative interaction. AAAI Spring Symposium on Mixed Initiative Interaction (pp. 89-94) Menlo Park, Ca: AAAI Press.

Miller, C. & Parasuraman, R. (2002). Designing for flexible human-automation interaction: Playbooks for supervisory control, Technical Report, Minneapolis, MN: Smart Information Flow Technologies.

Miller, C., Pelican, M., & Goldman, R. (2000). "Tasking" interfaces for flexible interaction with automation: Keeping the operator in control. In Proceedings of the Conference on Human Interaction with Complex Systems. UrbanaChampaign, IL.

Mirmohammad-Sadeghi, H., Bastani, H. & Azarnasab, E. (2003). IUTMicrobot Robobup. In Proceedings of the 2003 International Robocup. Padova, Italy : National Institute of Standards and Technology.

Parasuraman, R. (1993). Effects of adaptive function allocation on human performance. In D. J. Garland & J. A. Wise (Eds.) Human Factors and Advanced Aviation Technologies. (pp. 147-157). Daytona Beach: Embry-Riddle Aeronautical University Press.

Parasuraman, R. Galster, S., & Miller C. (2003). Human control of multiple robots in the RoboFlag simulation Environment. In Proceedings of the 2003 IEEE International Conference on Man, Systems and Cybernetics. Washington, D.C.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, Vol. 39, pp. 230-253.

Murphy, R., & Rogers, E. (2001). Human-robot interaction: Final Report forDARPA/NSF study human-robot interaction. Retrieved January 28, 2003, from http://www.csc.calpoly.edu/~erogers/HRI/HRI-report-final.html.

Ryan, P. (2003). Mine Countermeasures a Success [Electronic version], *Naval Institute Proceedings*. Retrieved January 28, 2004, from http://www.usni.org/Proceedings/Articles03/PROryan05.htm.

Sarter, N.B., Woods, D.D., & Billings, C.E. (1997). Automation surprises. In G. Salvendy (Ed.), Handbook of human factors and ergonomics (2nd edition, pp. 1926-1943). New York: Wiley.

Taylor, R.M. (1990). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In Situational Awareness in Aerospace Operations (AGARD-CP-478, pp. 3-1 and 3-17). Neuilly Sur Seine, France: NATO-AGARD.

Ververka, J. & Campbell, M. (2003). Experimental study of information load on operators in semi-autonomous systems. Proceedings of AIAA Guidance, Navigation and Control Conference. Austin, Texas: AIAA.

CONSIDERING SUBJECTIVE TRUST AND MONITORING BEHAVIOR IN ASSESSING AUTOMATION-INDUCED "COMPLACENCY"

Nasrine Bagheri, Greg A. Jamieson
University of Toronto

ABSTRACT

This article presents a study of human monitoring of an automated system considering both sampling behavior and subjective reports of trust and self-confidence. It replicates an experiment by Parasuraman et al. (1993) wherein participants were said to be lulled into complacency by unchanging automation. Results confirmed their finding that automation reliability had a significant effect on automation failures detection. In particular, participants using constant highly reliable automation had the poorest failure detection performance. These participants were also observed to have significantly longer intervals between their samples of the monitoring task. Despite this mean difference, the evolution of the attention allocation patterns for participants in the constant-high condition does not support the attribution of their poor performance to complacency. Results of trust and self-confidence ratings are also discussed. These results provide empirical support for Moray's (2000, 2003) assertion that attention allocation and psychological factors should be considered when evaluating monitoring performance and drawing conclusions about complacency.

Keywords : complacency, sampling rate, trust, monitoring of automated systems.

INTRODUCTION

Difficulties with human-automation interaction in complex systems frequently prevent the full benefits of the automation from being realized. This article focuses on one adverse consequence of automation known as *automation-induced complacency* (Parasuraman et al., 1993). Complacency occurs when the role of the human operator is changed from that of an active manual controller to that of a passive monitor of highly reliable automation, and refers to the ensuing decline of that monitoring performance (Farrell & Lewandowsky, 2000). Although researchers generally agree that complacency is a serious problem, little consensus exists to what complacency is and how it can be measured (Prinzel et al., 2001). Previous research has concluded that operators were complacent based primarily on their automation failure detection performance over time (e.g., Parasuraman et al., 1993). Moray (2000, 2003) questioned whether such evidence adequately supports the existence of complacency. He pointed out that (1) complacency is concerned with attention, and (2) that psychological factors such as trust may influence complacency. Missing signals does not necessarily imply complacency as even optimal sampling behaviour can result in missed signals. Rather, complacency may imply under-sampling and defective monitoring (Moray & Inagaki, 2000).

This paper presents a replication of a study conducted by Parasuraman et al. (1993) in which participants who interacted with a consistent and highly reliable automated system were said to show signs of complacency based on detection performance. Participant eye movements, their trust in the automated system, and their self-confidence were evaluated in addition to detection.

METHOD

The experiment was designed to replicate as accurately as possible Parasuraman et al. (1993).

Participants and apparatus

Based on a power analysis of the data obtained by Parasuraman et al. (1993) 24 participants were recruited. Participants had no prior experience with the simulation used in the study.

The Multi-Attribute Task battery (MAT; Comstock and Arnegard, 1992) was used. The MAT Battery is a multi-task flight simulation that requires participants to perform three equally important tasks: (1) tracking, (2) fuel management, and (3) system-monitoring. The goal of the tracking task was to keep the aircraft within a central rectangular area using a joystick (first-order control). The goal of the fuel management task was to compensate for fuel depletion by pumping fuel from the supply tanks to the main tanks. The system-monitoring task consisted of four engine gauges that participants had to monitor for randomly occurring abnormal values that represented system malfunctions. The monitoring task was automated so that a gauge showing an abnormal value would normally reset itself without participant intervention. However, participants were advised that the automated system would sometimes fail to correct these malfunctions. In such a situation, participants were required to correct malfunctions manually. If they did not detect the automation failure within 10 seconds, the event was scored as a "miss" and the pointer was automatically reset. Participants were not informed that they missed a failure.

An Eye-gaze Response Interface Computer Aid (ERICA) system was also used to track the eye movements of the participants. Gaze location samples were taken 30 times per second.

Procedure

Following a 10-minute training session, participants completed four 30-minute sessions on the MAT battery for a total of 12 10-minute blocks. At the end of each session, participants rated their trust in the automated system and their self-confidence in performing each tasks on a 10-point scale similar to the one used by Lee and Moray (1992, 1994).
A 4 (reliability) by 12 (blocks) mixed factorial design was used. Automation reliability was varied as a between-subjects factor with four levels (see Figure 1). There were 16 malfunctions in each 10-minute block. Automation reliability was defined as the percentage of malfunctions successfully corrected by the automation in each block. Six participants were randomly assigned to each of the four reliability conditions.

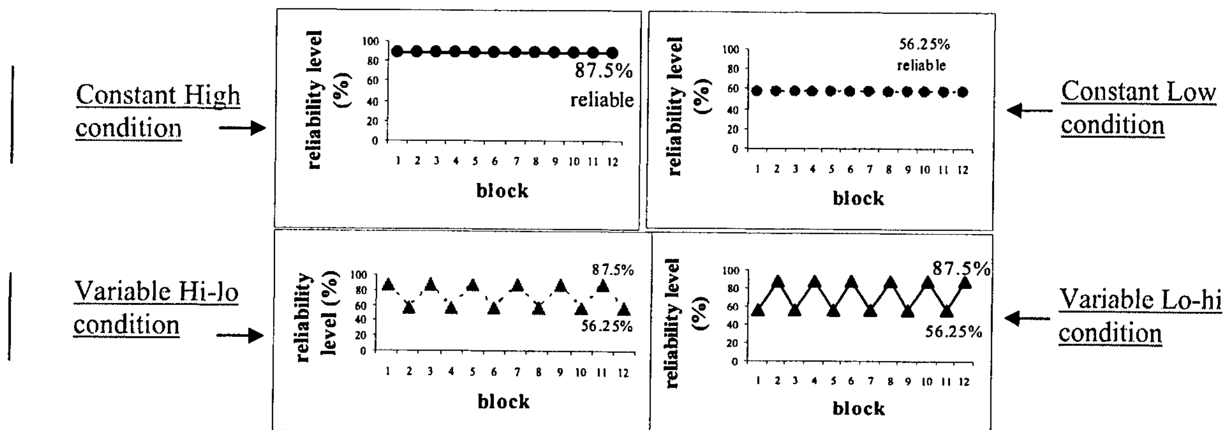


Figure 1. Graphical representation of the automation reliability conditions.

RESULTS

Detection rate of automation failures.

As in Parasuraman et al. (1993), a 4 (reliability) x 12 (Block) ANOVA of the detection rate indicated a significant effect for automation reliability $F(3, 20) = 11.92, p < .001$ (see Figure 2). Post-hoc analysis revealed that the detection performance of the Constant High participants was poorer than that in any other condition. This result differs from that of Parasuraman et al. (1993), who found no significant difference in detection performance between the Constant High and the Constant Low condition. The difference observed in the present study precludes nesting the two reliability groups. As in Parasuraman et al. (1993), the block effect on detection performance was significant $F(11, 220) = 2.23, p < .05$. The interaction was not significant $F(33, 220) = 1.32, p > .05$.

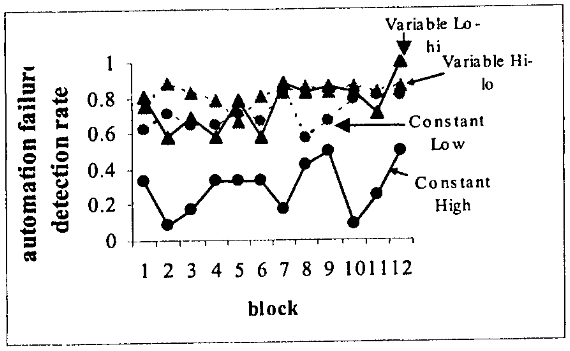


Figure 2A. Effect of automation reliability and block on participants' detection performance in the current study.

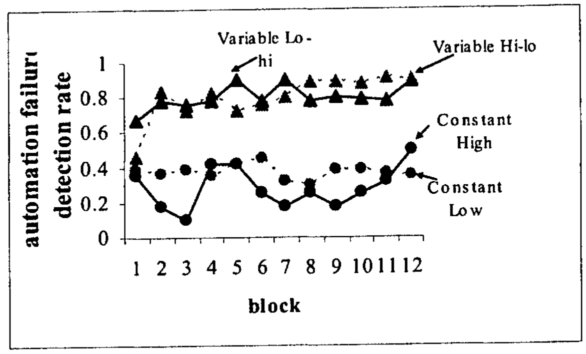


Figure 2B. Effect of automation reliability and blocks on participants' detection performance reconstructed from Parasuraman et al.(1993).

To determine the effect of automation reliability based on equal failure rates, performance of Constant High participants was compared to the performance of those in the variable conditions for the blocks where the reliability was also high. That is, performance in the Constant High condition was compared to the performance in block 1 of the Variable Hi-lo condition and block 2 of the Variable Lo-hi condition, etc. When faced with highly reliable automation, participants in the variable conditions performed significantly better $F(1, 10) = 21.89$, $p < .001$ (Figure 3A). Conversely, for low reliability blocks and conditions, results revealed that whether the reliability was constant or variable did not significantly affect participants' detection rate $F(1,10) = 1.706$, $p > .05$, although Constant Low participants performed poorer in 11 of 12 blocks (see Figure 3B).

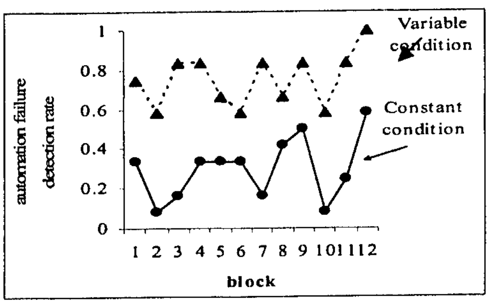


Figure 3A: Comparison of the variable and constant conditions for the high reliability level

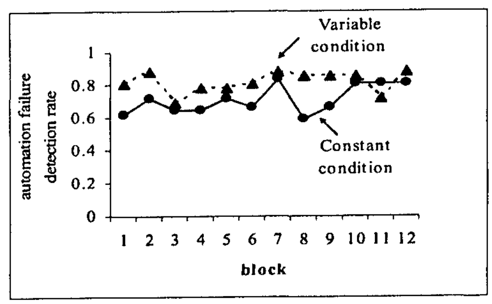


Figure 3B: Comparison of the variable and constant conditions for the low reliability level

There was no significant effect of group difference on either tracking performance $F(3, 20) = 1.27$, $p > .05$ or resource management performance $F(3, 20) = 0.42$, $p > .05$.

Attention and sampling rate.

Parasuraman et al., by informal video observation, did not find any systematic difference in scanning behavior between participants in their constant and variable conditions. In the present study, participants' eye movements were recorded to determine how attention was allocated to the three tasks. The Mean Time Between Fixation (MTBF) for the three lookzones of the MAT battery was measured. The effect of reliability on the log-transformed MTBF of the monitoring task was significant $F(3, 20) = 34.60$, $p < .0001$ (Figure 4), and so was the block effect $F(11, 121) = 2.06$, $p < .05$. The MTBF was transformed to compensate for the skewed variable distribution. The interaction effect was non-significant. Post-hoc analysis further showed that the MTBF of the monitoring lookzone was higher for Constant High participants than for participants in any other condition. Figure 4 shows that the MTBF of Constant High participants gradually increased in the first 3 blocks, but then decreased and converged toward the MTBF of participants in the other three conditions. The detection rate was negatively correlated with the MTBF, $r = -0.57$, n = 189, $p < .01$.

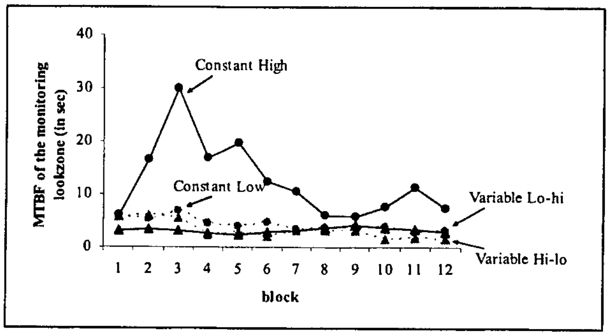


Figure 4: MTBF of the monitoring lookzone

Trust in automation.

Parasuraman et al. (1993) suggested that 'waxing and waning of trust' with the success and failure of the automation could account for part of their detection results. However, the authors did not report any trust measures. In the current study, no significant effect of automation reliability on participants' rating of trust was found $F(3, 20) = 1.19$, $p > .05$ (Figure 5). However, the low power of the test ($1\text{-}\beta = 0.3$) should be noted. The block effect was non-significant $F(3,20) = 0.298$, $p > .05$. Correlation analysis revealed that detection rate was inversely correlated with the level of trust (i.e., the more participants trusted the automation, the lower their detection rate), $r = -0.39$, p <.01. Trust was also positively correlated with the MTBF of the monitoring lookzone $r = 0.34$, $p < .01$.

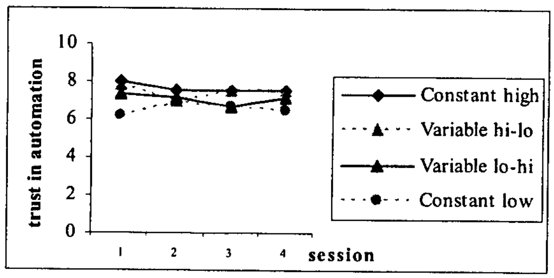


Figure 5. Rating of trust in the automation.

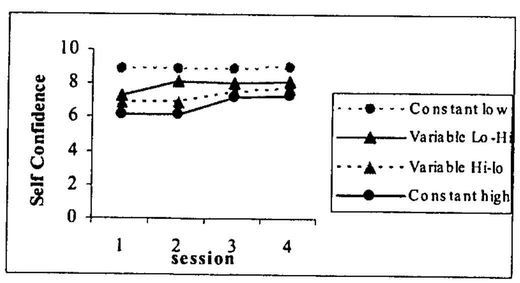


Figure 6. Rating of self-confidence

Self-confidence in performing the monitoring task.

The effect of automation reliability on participants' self-confidence approached significance $F(3, 20) = 2.883$, $p = 0.06$. Post-hoc analysis revealed that the difference between the Constant Low and the Constant High condition approached significance. Constant High participants had the lowest self-confidence in their ability to perform the monitoring task (Figure 6).

DISCUSSION

On the surface, our results replicate much of those found by Parasuraman et al. (1993). The detection of automation failures was significantly worse for participants facing constant, highly reliable automation, which could indicate that these participants showed signs of complacency. Following Parasuraman et al. (1993), several explanations for the observed poor performance should be considered. First, the poor detection performance of Constant High participants could be related to the 'signal rate'. As they faced a low probability of signal occurrence, we might expect their probability of detecting a failure to be lower (Parasuraman, 1986). However, in blocks with equivalent failure rates (i.e., signal rate), we still observed poorer performance for Constant High participants compared to participants in the Variable conditions. Thus, like Parasuraman et al. (1993) we conclude that low signal rates alone do not explain the poor detection performance.

Secondly, Parasuraman et al. (1993) suggested that differences in attention allocation could explain the observed difference. Using informal observations of participants' eye movements, they observed no major differences in scanning behavior between participants in the constant and the variable conditions, although they did not rule out the possibility of small differences. In the present study, eye point of gaze data revealed that Constant High participants had a significantly higher MTBF of the monitoring lookzone. More importantly, the difference in the MTBF between the Constant High condition and the other conditions increased in the first 3 blocks, but then decreased starting in Block 4. This decrease argues against the hypothesis that complacency appeared after a long period in presence of highly reliable automation. This change in attention allocation strategy could not be observed from detection results, which shows the importance of measuring attention in order to accurately evaluate monitoring performance (Moray, 2000, 2003).

Analysis of participants' subjective ratings of trust also forestalls the conclusion that Constant High participants were complacent. Self ratings of trust in the automation revealed no differences between the reliability conditions or across blocks, indicating that poor monitoring performance might not reflect overtrust. This is not to say that trust is not an important factor in monitoring. To the contrary, trust was shown to have a moderate-to-large effect on both monitoring behaviour and detection performance. No trust data were collected by Parasuraman et al. (1993), although the authors cited the 'waxing and waning of trust' as a possible factor in explaining their observations.

Similarly, Constant High participants had the least confidence in their ability to detect failures. Participants with lower self-confidence in their monitoring skills could be expected to be poorer monitors than those with high self-confidence while interacting with a constant highly reliable automation (Prinzel, Pope, and Freeman, 1999). However, it should be noted that Constant High participants were presented with few failures, and did not know if they missed one. Their lower self-confidence might thus be due to a belief that they were missing some signals as they knew that the automation was not 100% reliable. Low self confidence may explain why the sampling patterns of Constant High participants converged towards the level of the other conditions.

The most perplexing observation involves the failure detection performance of Constant Low participants. Parasuraman et al. (1993) observed little difference in detection performance between the Constant Low and the Constant High condition. In contrast, detection performance of Constant Low participants in the present study differed significantly from that of Constant High participants, and was similar to that of participants in the Variable conditions. In the absence of the Parasuraman et al. (1993) results, this observation might indicate that the reliability level was low enough to offset complacency induced by the constant-reliability environment, if complacency there was. Attention data would corroborate this conclusion since the MTBF of the monitoring lookzone of the Constant Low condition was not significantly different from that of the Variable conditions. However, the strong contrast between this observation and that reported by Parasuraman et al. (1993) more likely suggests a discrepancy between the study protocols. All efforts were made to replicate the study as described in the literature, but some details were not readily available.

CONCLUSION

We believe that this study is the first to look at automation-induced complacency based on both sampling behavior and subjective reports of trust and self-confidence. Detection rate results alone might indicate that participants using constant high reliability automation showed signs of complacency. However, assessments of attention allocation, trust, and self-confidence appear to contradict this conclusion. Thus, Moray's (2000, 2003) assertion that investigators must consider allocation of attention and psychological factors when evaluating monitoring performance and drawing conclusions about complacency gains credence from these results. More generalizable conclusions will require that these results be compared against an optimal sampling rate.

ACKNOWLEDGEMENT

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. Thank you to Catherine Burns at the University of Waterloo for the loan of the ERICA system.

REFERENCES

Farrell, S. & Lewandowsky, S. (2000). A connectionnist model of complacency and adaptive recovery under automation. Journal of Experimental Psychology-Learning Memory and Cognition, 26, 395-410.

Lee and Moray (1992). Trust control strategies and allocation of function in human-machine systems. Ergonomics, 35, 1243-1270.

Lee and Moray (1994). Trust, self-confidence, and operators' adaptation to automation. Human Computer Studies, 40, 153-184.

Moray, N. (2003). Monitoring, complacency, scepticism and eutectic behaviour. International Journal of Industrial Ergonomics, 31, 175-178.

Moray, N., T. Inagaki (2000). Attention and complacency. Theoretical Issues in Ergonomics Science, 1, 354-365.

Parasuraman, R. (1986). Vigilance, monitoring and search. In Boff, K., Kaufman, L. & Thomas, J.(Eds), Handbook of perception and human performance: Vol.2 Cognitive processes and performance (43.1-43.39).

Parasuraman, R., Molloy, R. & Singh I. L. (1993). Performance consequences of automation-induced "complacency". The International Journal of Aviation Psychology, 3, 1-23.

Prinzel, L. J., DeVries, H., Freeman, F.G., & Mikulka, P. (2001). Examination of Automation-Induced Complacency and Individual Difference Variates. (Tech. Memorandum No. 211413). Hampton, VA: NASA Langley Research Center.

Prinzel L., Pope, A. & Freeman, F. (2000). The double-edged sword of self-efficacy: Implications in automation-induced complacency. In Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting (124-128). San Diego, 2000.

A "DISTANCE"-BASED CONCEPT OF AUTOMATON FOR HUMAN-ROBOT INTERACTION

Ruiqi Ma, David B. Kaber, Mo-Yuen Chow, Nancy Currie
North Carolina State University

ABSTRACT

Although supervisory control scenarios have attracted significant research attention for nearly half a century, this general type of automation may not be representative of the next generation of automated or semi-autonomous systems (e.g., mobile-telerobotic systems). Such systems will be developed in the context of teams of people working with teams of robots in dynamic and uncertain environments through many different roles, not limited to supervisory control. There is a need to develop a new general concept of automation for contemporary complex automated systems to model such systems, define the roles of human operators and attempt to explain and predict systems performance. We propose a "distance"-based concept of automation that can be used to describe various forms of human-robot interaction. As the physical distance between a human operator and remote robotic work package increases, there is also an increased likelihood of spatial/temporal perturbations influencing system performance; thus, higher levels of automation are required to deal with such disturbances. This concept can be used to develop a hierarchical representation of complex human-robot interaction scenarios, and to classify various forms of automation in existing human-robot interaction applications.

Keywords: human-robot interaction; automation; supervisory control; telerobots; dynamic environments

SUPERVISORY CONTROL AND NEXT GENERATION AUTOMATION

According to Sheridan (2002), "automation" is a term originally used to refer to automatic control in the field of manufacturing, specifically production of a part through a number of successive stages. Today, the term has expanded to encompass any use of electronic or mechanical devices to replace human labor (Parasuraman, Sheridan and Wickens, 2000). From a classical control theory perspective, in many automated systems human operators are relegated to the role of supervisor of over machines that are responsible for the very roles the human once performed. The human operator is typically involved in system control through interaction with automation and by maintaining final decision-making authority. For example, in some supervisory control scenarios (e.g., nuclear power plant control), multiple human operators may intermittently program and receive information from a computer that interacts through sensors and effectors to control the reaction process or core environment. Since fully autonomous operation of many systems is not possible at this point in time, the human remains an integral part of the control loop, as a supervisor of automation or passive decision maker (Endsley and Kiris, 1995).

The study of automation, from a human factors perspective, has historically focused on human-automation interaction and in complex single-user, multiple-machine systems control (Sheridan, 2002). The use of advanced automation, combined with human supervisory control, has found wide spread application and acceptance across various contexts, including aviation, transportation, nuclear power plant process control, hospital systems, and teleoperators (see Sheridan, 2002). The primary concern with these types of systems has been human out-of-the-loop performance problems (Endsley and Kiris, 1995; Parasuraman, et al., 2000; Sheridan, 2002), including operator complacency, vigilance decrements, loss of situation awareness (SA), etc. Since a transformation of system information must occur between the human operator and machine, another major concern has been with the design of the human-machine interface (or interactions).

Although supervisory control scenarios found in many contemporary automated systems have attracted significant research attention for nearly half a century, this general type of automation may not be representative of the next generation of automated, or semi-autonomous, systems, for example, multiple operator control of mobile-telerobotic systems (e.g., the US Air Force Predator Unmanned Aerial Vehicle). Such systems will be developed in the context of teams of people working with robots, in dynamic and uncertain environments, through many different roles not limited to supervisors (Pontbriand, 2003). Mobile robots are being deployed in applications such as search and rescue, first response to chemical/biological incidents, outer space and deep-sea exploration, and tactical military operations. These applications require automation that is intelligent and adaptive in nature. Human interaction with this type of automation necessarily requires a diverse team of users, with different goals and knowledge, acting as operators, teammates, and mechanics to robots, as in the World Tower Center search and rescue operations after September 11, 2001 (Casper, 2002). Consequently, supervisory control, which can be characterized by an operator sensing information from the system, programming or instructing the system, and responding to actions of the system, may be not suitable to describing operations of new coordinated robotic systems in dynamic environments. The demands of such applications, including human-robot coordination cannot be met through the human acting strictly as a supervisor.

APPROACH TO A NEW CONCEPT OF AUTOMATION

As a result of the limitations in the supervisory control concept, there are needs to develop a new general concept of automation for contemporary complex systems to define the roles of humans in such systems, and to serve as a basis for directing enhancements of robotic system capabilities, including displays and controls, to facilitate coordination with humans in jointly carrying-out activities in dynamic environments. Here we propose an approach to a "distance"-based concept of automation to address the first research need. With respect to enhancing mobile-robot system displays and controls, some preliminary research has been done to define a systematic approache to developing effective interface technologies (Kaber and Chow, 2003), and multi-modal interface designs have been developed for specific applications by Estremera, Garcia and Santos (2002).

Many new advances in automated systems can be viewed as means for human perception at a distance or action at distance (Woods, 2003). For example, Rybski, Stoeter, Gini, Hougen and Papanikolopoulos (2002) used roving range and scout robots for surveillance tasks in indoor urban environments, which can be viewed as means for facilitating human perception-at-a-distance. In applications of this technology, the primary objective for human-robot coordination is projecting human intentions into the world at a necessary distance. As the physical distance between the human and robotic work package increases, there is also an increased likelihood of spatial and temporal perturbations influencing system performance. For example, time lags in wireless network-based control of remote rovers may make it difficult for operators to associate control actions with concurrent system states, ultimately degrading performance. This situation typically dictates the need for complex communications technologies, and display and control technologies to account for lag. That is, the level, or degree, of system autonomy is often directly proportional to the physical control distance. As the number of computer systems (acting as information filters) or software ("middleware") applications setup between a telerobot and a human operator increases, the degrees of separation of the human from direct control of the remote manipulator (or the metaphorical "distance" of control) increases. Thus, we have a "distance"-based concept of automation.

In order to achieve performance in a long-distance, telerobot control scenario (e.g., multiple manipulator arm control on the International Space Station (ISS)) comparable to performance in a direct teleoperation scenario (e.g., tele-manipulator control in a nuclear "hot" lab) under no spatial or temporal perturbations, the level of robot system autonomy must be greater. In the ISS manipulator control scenario, there are many different types of local and remote control hardware and software that may need to be implemented to ensure state and efficient operation under time lag. For example, automated manipulator force control may be implemented to constrain user control actions and telerobot motions in Station maintenance tasks to a single axis of translation or rotation at any given time, with the objective of reducing the overall complexity of the control task and promoting system safety (Currie, 2003). Such an algorithm may prevent errors in control and excessive forces on a task object (e.g., station electronic components) causing damage. As another example, automated telerobot control gain adaptation has been implemented in experimental applications in which severe lag conditions exist (Tipsuwan and Chow, 2003). That is, software (or "middleware" applications between the human operator and remote work package) is used to characterize lag conditions (in real time) and the gain of the operator control is adjusted accordingly in order to maintain robot system stability and, at the same time, productivity under human control. Middleware can be loosely defined as a software layer between an application and transport layers in a communication network system. Middleware has been used to make networks transparent in end-to-end user applications. Under lag conditions, operators may execute control actions based on visual feedback that are not appropriate for the actual current state of the remote robot/manipulator. The middleware can be programmed to accurately assess the control lag conditions and alter operator control actions in order to ensure they are safe based on model predictions of actual robot states. These types of methods for dealing with spatial and temporal perturbations in long distance, telerobot control represent forms of system automation that may be necessary to achieve sufficient performance.

Recent empirical research has demonstrated control gain adaptation algorithms in telerobot control to be effective for reducing human control errors under lag conditions yet maintaining task performance efficiency (Sheik-Nainar, Kaber and Chow, 2003). Sheik-Nainar et al. (2003) evaluated the effects of different types of communication networks delays (no-delay, constant and random delays) on operator performance in a telerover navigation task (avoiding obstacles in an outdoor environments). There were two levels of robot control/automation (LOA) investigated in the study, including direct teleoperation (users conveyed discrete movement commands to the telerover) and telerobotic control, in which users and the main robot computer controller jointly defined navigational goals for the rover and formulated potential trajectories. The results revealed significant influences of LOA, delay type and adaptation on the time-to-task completion, and the number of errors (telerover and obstacle collisions). The higher LOA (telerobotic control) produced shorter task times, but significantly more control errors attributable to operator out-of-the-loop performance problem. There were significantly fewer errors (50-60% less), and only slightly longer time-to-task completion (10-20% longer), with gain adaptation, as compared with no adaptation under all network communication conditions.

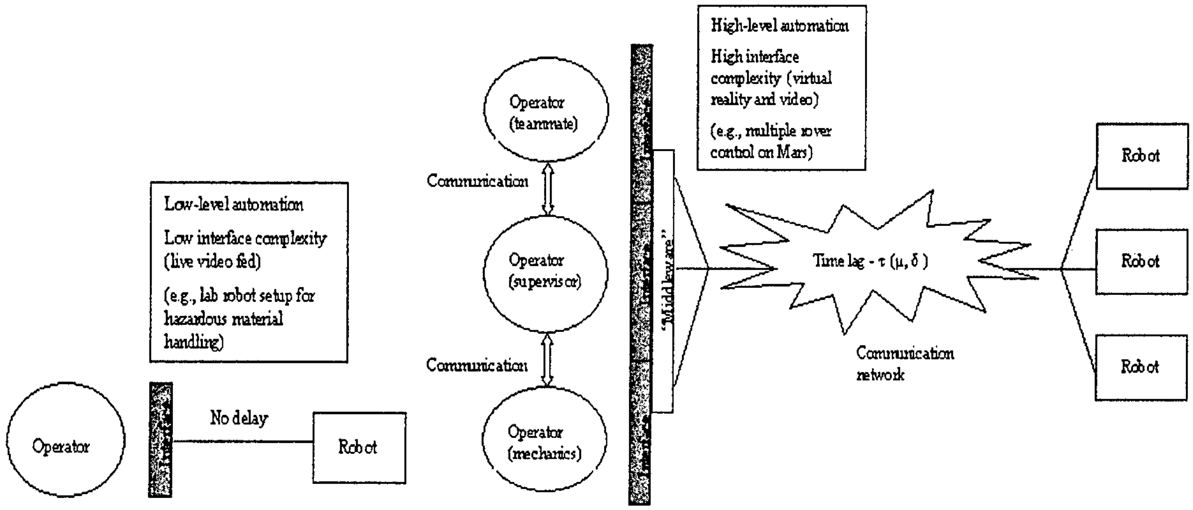
Although technologies such as gain adaptation software may be effective for preserving acceptable levels of human performance in difficult telerobot control situations, they, none-the-less, represent increasing degrees of separation of the human from direct teleoperation or rover control. If the technology fails, the operators' task of diagnosing the problem and recovering the system becomes far more complex than in a direct control scenario. Furthermore, their capability to control the system without automated aids may be fairly limited as a result of becoming accustomed to use of the technology.

A "distance" based concept of automation can also be used to characterize the role of the human operator in a teleoperation scenario. The greater the "distance" of control, the greater the extent to which the operator's role may be limited to monitoring automation and acting as a passive-decision maker (detecting automation errors and intervening for system recovery). This is unlike direct manipulator control, which typically involves the operator planning robot motions, selecting a "best" trajectory and manually implementing the trajectory using a hand controller. This form of teleoperator/control would, on the other hand, be characterized by a short control "distance".

It is also possible to extend the "distance"-based concept of automation to teleoperation scenarios in which multiple operators act to support a single remote system through different roles or multiple operators collaborate with multiple robots. Depending upon the role of the operator (robot teammate, operations supervisor), there may be different control channels and interfaces through which the humans and machines communicate. The control "distance" can be established for each channel and associated with the specific operator roles.

APPLICATION OF THE "DISTANCE"-BASED CONCEPT OF AUTOMATION

The "distance"-based concept of automation could be applicable to describe various forms of human-robot interaction. The concept can be easily quantified by considering the physical distance through control channels between the human and the remote robot task environment. The most direct form of control may be associated with the shortest actual physical distance. We can roughly describe the range of control "distance" as (a) short, (b) medium, and (c) long. Figure 1 presents two types of teleoperation scenarios including either one operator and one robot, or multiple operators and multiple robots collaborating together at short and long control "distances". Under the short control "distance" (Figure 1.a), there may be no temporal perturbation in control communications and, consequently, the degree of automation in the control channel may be limited and a relatively simplistic operator interface can be used. Under the long control "distance" (Figure 1.b), there may be substantial communication delays in controlling remote work packages. The lag may be variable in nature and "middleware" may be required in the communication channel to monitor delay conditions in real-time and act as a fail-safe mechanism when operators are aggressive in their control actions. Beyond this, the lag conditions may dictate that the human interface incorporate a graphical model of the remote systems in order for operators to perform robot programming without having to wait long periods of time for feedback on control actions. Live video may also be provided as a means for verifying the accuracy of programming (essentially creating a predictive display setup). Consequently, both the degree of automation and interface complexity in this scenario may be very high.



a). Short control "distance"   b). Long control "distance"

Figure 1: Teleoperation scenarios with different physical distances and control "distances" (degree of operator-robot separation)

The control "distances" depicted in the figure plates may correspond to general levels of automation already defined in existing taxonomies of automation (Endsley and Kaber, 1999; Parasuraman, et al., 2000; Parasuraman and Byrne, 2003). That is, general units of control "distance" could be defined and specific distances directly related to levels of automation defined in contemporary theories. For example, Endsley and Kaber (1999) developed a 10-level taxonomy of LOAs based on allocations of complex system functions, including systems monitoring, generating processing plans, decision making, and implementing actions to human or computer servers. In their concept of automation, higher LOAs correspond to increasing replacement of functions formerly carried out by the human with machine functions. This is consistent with the control "distance" concept of automation. When the actual distance between the human and robot is great, and the likelihood of spatial and temporal perturbations is high, the robot is given more authority (to accomplish functions independently); thus, leading to higher LOAs in the teleoperation scenario. In fact, Kaber and Endsley (1997) previously related their LOAs to teleoperation scenarios, including equating direct teleoperation to their level of "Action Support". With greater distance between the human and machine, there may also be a need for automated systems or control computers to be more responsible for the robotic system in terms of the four different types of automation functions identified by Parasuraman et al. (2000): information acquisition, information analysis, decision selection, and action implementation. In this way, the "distance"-based concept of LOA may also provide a convenient way of classifying other complex automated systems in terms of existing theories or taxonomies of automation.

Beyond this, the "distance"-based concept of automation may be important, because it could be used to provide a hierarchical representation of multiple, complex human-robot interactions within a single control scenario (multiple operators teaming with multiple robots) or across scenarios. In general, the concept may make it easier to describe various forms of real human-robot interaction that do not fit into the historical concept of supervisory control.

CONCLUSION

With respect to the research need to develop a new general concept of automation for contemporary complex teleoperation systems and to characterize the roles of humans in such systems, we proposed a "distance"-based concept of automation. As the physical distance between the human and robotic work package increases, there is also an increased likelihood of spatial/temporal perturbations influencing system performance; thus, leading to the need for higher LOAs in local and remote control to facilitate stable and safe human-robot interaction. We also compared and linked the new concept of automation to the existing taxonomies of automation presented in the human factors literature (Endsley and Kaber, 1999; Parasuraman, et al., 2000). The potential advantages of this concept for characterizing human-robot interaction include a more objective quantification of LOAs in terms of units of control "distance" between the human operator and point of application, and the capability to relate an operator roles in telerobot control to this theoretical control "distance".

REFERENCES

Casper, J. (2002). Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. Unpublished M.S. thesis, Computer Science and Engineering, University of Southern Florida, Tampa, FL.

Currie, N. J. (2003). Command and Telemetry Latency Effects on Operator Performance of Space Station Robotic Operations. Presentation to Department of Industrial Engineering, NC State University, Raleigh, NC (November 13, 2003).

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics, 42(3), 462–492.

Endsley, M. R. & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automaiton. Human Factors, 37, 381-394.

Estremera, J., Garcia, E., & Santos, G. (2002). A multi-modal and collaborative human-machine interface for a walking robot. Journal of Intelligent and Robotics Systems, 35, 397-425.

Kaber, D. B., & Chow, M. (2003). Human-robot interaction research and an approach to mobile-telerobot interface design. In Proceedings of the XVth Triennial Congress of the International Ergonomics Association and the 7th Joint Conference of Ergonomics Society of Korea.Japan Ergonomics Society (CD-ROM). Seoul, Korea, August 24-29, 2003.

Kaber, D. B., & Endsley, M. R. (1997). The combined effect of level of automation and adaptive automation on human performance with complex, dynamic control systems. In the Proceedings of the 41st Annual Meeting of the Human Factors and Ergonomics Society (pp. 205-209). Santa Monica, CA: Human Factors and Ergonomics Society.

Parasuraman, R., & Byrne, E., A.(2003). Automation and human performance in aviation. In P.Tsang (Ed.), Principles of Aviation Psychology (pp.311-356). Mahwah, NJ: Erlbaum.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model of types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics, 30(3), 286-297.

Pontbriand, R. J. (2003). Human-robot coordination: panel overview. In Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting (CD-ROM), pp 454-457. Denver, Colorado, October 13-17, 2003.

Rybski, P. E., Stoeter, S. A., Gini, M., Hougen, D. F., & Papanikolopoulos, N. (2002). Performance of a distributed robotic system using shared communications channels. IEEE Transactions on Robotics and Automation, special issue on Multi-Robot systems, 18*(5)*, pp. 713-727, October, 2002.

Sheik-Nainar, M., Kaber, D. B., & Chow, M. (2003). The effects of gain adaptation on QoS deterioration in an internet-based teleoperation using a virtual reality interface . In Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting (CD-ROM). Denver, Colorado, October 13-17, 2003.

Sheridan, T. B. (2002). Human and Automation: System Design and Research Issues. A John Wiley & Sons.

Tipsuwan, Y. & Chow, M. (2003) On the Gain Scheduling for Networked PI Controller Over IP Network, In Proceedings of 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, pp. 20-24. Port Island, Kobe: Japan.

Woods, D. (2003). Intent at a distance. In Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting (CD-ROM), pp 455. Denver, Colorado, October 13-17, 2003.

TOWARD AN AUGMENTED COCKPIT

Tom R. Diethe, Blair T. Dickson
QinetiQ Ltd. Centre for Human Sciences

Dylan Schmorrow
DARPA/IPTO

Colby Raley
Strategic Analysis

ABSTRACT

This effort seeks to demonstrate principles of adaptive automation, based on the cognitive–affective status of personnel and the current mission requirements, by combining key enabling technologies, including the support of decision making and the identification of operator status. The fused output from these technologies allows for the adaptive control of interfaces and dynamic function allocation. These systems are integrated using principles of cognitive engineering, and demonstrated in a fast jet simulation environment during a realistic military mission.

These sub-systems are integrated into the augmented cockpit. This provides a test bed for examining the principles and practice for augmenting cognition in the fast jet environment.

Keywords: Augmented Cognition; Adaptive Automation; Dynamic Function Allocation; Cognitive Cockpit

INTRODUCTION

The cockpit environment is changing. Traditionally the major demands placed on a pilot were associated with the task of flying the aircraft; however as levels of cockpit complexity increase, the focus has changed away from skill to knowledge–based tasks, and the role of the pilot is centered on the processing of information. This information may be presented in a number of different formats, in the auditory or visual modality for example, containing either verbal or spatial information, and pilots may interact with cockpit systems from numerous interfaces. The potential for information overload and excessive workload is great. In response to this changing role the Cognitive Cockpit (CogPit) has been developed to support the vision of:

'A Cognitive Cockpit which allows the pilot to concentrate his skills towards the relevant critical mission event, at the appropriate time, to the appropriate level' (Taylor et al, 2000).

The CogPit has been developed by fusing a number of enabling technologies to produce a cockpit that can adapt to the pilots' needs and the mission requirements, in real time. These key technologies comprise the real–time estimation of cognitive-affective status derived from the tracking of physiological and behavioral measures, the implementation of a knowledge-based system designed to provide context-sensitive decision support, and a framework for the implementation of adaptive automation and task scheduling. These are implemented using principles of cognitive engineering through a number of adaptive interfaces. A closed–loop trial has just been completed (November 2003) during which the stability and performance of the system were examined under different levels of threat/workload in a realistic deep-strike mission.

DESIGN AND ARCHITECTURE

The Cognitive Cockpit has been designed to be modular at a functional level enabling the independent development of the core and sub-systems. This has enabled a number of generic principles to be followed, ensuring that the sub-systems may be readily ported to application environments other than a fast-jet cockpit.

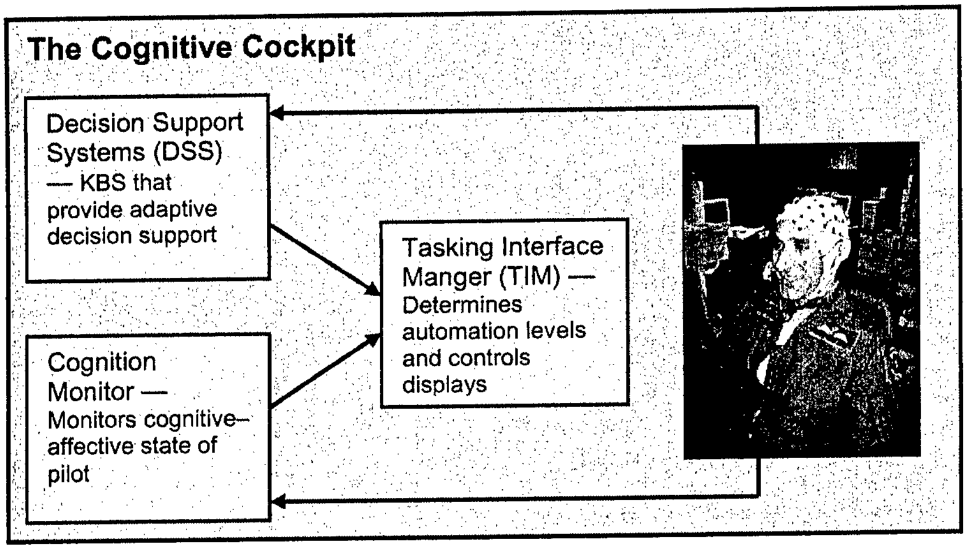


Figure 1 –Top-level design of the Cognitive Cockpit identifying key sub-systems

Figure 1 identifies the three key sub-systems that have enabled a real-time closed-loop platform to be developed and tested. These are the Decision Support Systems (DSS), the Cognition Monitor (CogMon) and the Tasking Interface Manager (TIM). These, along with the simulation test bed into which they are implemented, are characterized in the following sections.

Cognition Monitor

The Cognition Monitor has been developed to provide an on-line analysis of the cognitive–affective status of the pilot. Primary functions of this system include continuous monitoring of workload, and inferences about current attentional focus, ongoing cognition and intentions. Overall, this system provides information about the objective and subjective state of the pilot within a mission context. Inferences about pilot state are derived from four principal sources: behavioral measures, physiological measures, subjective measures, and through a consideration of contextual information (Pleydell-Pearce et al, 2003). These estimations are combined within high-level state descriptors such as levels of stress, alertness and workload, are then provided to the Tasking Interface Manager.

Decision Support Systems

The DSS are knowledge-based systems designed to support decision making and maintain situational awareness based on a dynamic evaluation of the operational context and through the generation of recommendations, or "plans". The DSS monitor the platform and make inferences about the internal and external aircraft environment. The knowledge base of the DSS was derived from RAF tactical manuals and validated through knowledge acquisition with Jaguar and Tornado aircrew.

Tasking Interface Manager

The TIM has been developed to dynamically allocate pilot functions, and to manage cockpit interfaces, mission tasks and timelines, by interpreting inputs from the DSS and the CogMon. These integrative functions enable the TIM to prioritize tasks and to determine the means by which pilot information is communicated. Overall, this system manages the cockpit automation by context-sensitive control over the allocation of tasks to the automated systems. The level of automation can be altered in real time in accordance with mission situation, pilot requirements and/or pilot capabilities. This capability is afforded through the application of a Pilot Authorisation and Control of Tasks (PACT) framework (Bonner et al 2000). PACT allows the pilot to form a *contract,* or set of contracts, with the automation by allocating PACT levels on a task by task basis. During operation, the TIM monitors the output from the DSS. When a plan is developed the TIM examines the PACT levels of each task within the plan and either performs the task automatically or provides assisted decision support, and presents the information in the most appropriate manner. This is derived from an examination of the pilot status gauges identified by CogMon.

Simulation Test bed

A synthetic environment has been developed to demonstrate the principles behind Augmenting Cognition. This test bed integrates the primary functional components of the CogPit as software agents, operating in a synthetic environment with realistic cockpit interfaces and within a representative mission scenario. The simulation environment is a real-time system that enables both the mission and the environment to be simulated, and allows a number of aiding options to be examined in selected mission phases.

In the following section we will discuss the primary method for augmenting cognition, that of mitigation of excessive workload.

MITIGATION STRATEGIES

We take the view that no one mitigation strategy is a panacea, and that only through a thorough examination of a number of strategies will the most effective approach be identified. What follows is a brief discussion of the mitigation strategies that we have taken into consideration, followed by a more detailed discussion of our primary mitigation strategy, namely Adaptive Dynamic Function Allocation (A-DFA).

Temporal aspects of task management have long been recognized as playing a major role in operator workload (Jordan et al 1995). We have therefore identified task scheduling according to resource availability as a possible mitigation strategy. This based on the assumption that additional information load at high workload sections of the mission is likely to compromise the ability of the operator to perform his/her primary task, such as control of the vehicle (if a pilot), or the maintenance of Situation Awareness. Throughput of information (warnings, task-related information, general information) is metered in accordance with available cognitive resources, and as such information of low importance can be discarded during mission-critical events.

A related mitigation strategy that the TIM employs is task queuing and prioritization according to saliency, such that higher saliency information is inserted earlier in the queue than lower saliency information. In addition information of higher saliency is presented in more prominent ways through the use of available interface manipulations. This is based on the assumption that performance is limited when two or more processes compete for a common neural structure – this competition can be removed when task scheduling is employed.

Modality switching is a potentially powerful mitigation strategy based on a model of human cognition that states that information can be more readily assimilated when parallel non-conflicting input channels are employed, and is preferable to loading up a single modality (e.g. Wickens, 1992).

A-DFA is a form of Adaptive Automation in which a negative feedback loop is formed between the operator and the system, such that the system reacts by increasing automation levels in periods of high workload and *vice versa*. This is based on the assumption that additional task load during high workload sections of the mission is likely to impinge on the primary task(s), as stated before, and increased automation of incoming tasks will enable the operator to concentrate on critical mission events. The shifting of task allocation between the operator and the system must be performed through the use of a structured adaptive automation framework, e.g. PACT. The PACT framework (figure 2), is a reduced, practical set of levels, with clear engineering and interface consequences; it is derived from the ten levels of automation for human–computer decision making proposed by Sheridan and VerPlanck (1978), with notable similarities with the levels of control and automation proposed by Endsley and Kiris (1995).

Of the mitigation strategies described above, the CogPit currently has implemented task scheduling, task queuing and prioritization, and A-DFA. In the following section we will describe in more detail how the switching of automation levels occurs.

	PACT LEVEL	Pilot Authority	Computer Autonomy
AUTOMATIC	5	Interrupt	Full
DIRECT SUPPORT	4	Revoke	Action (unless revoked)
IN SUPPORT	3	Authorise	Action (if authorised)
ADVISORY	2	Full	Advisory
AT CALL	1	Full	Advisory (if requested)
COMMANDED	0	Full	None

Assisted modes

Figure 2: PACT framework for A-DFA

CONTROL OF ADAPTIVE AUTOMATION

Prior to the mission a detailed automation analysis is performed on those tasks that may be included in advice from the decision support system. This analysis identifies default, maximum and minimum levels of automation within the PACT framework. Constraints may include individual pilot preferences, rules of engagement and the functionality developed within the test platform. In addition a set of rules that govern the change between automation levels is defined. These 'contracts' are designed to establish trust between operator and automation, and ensure that changes in automation are not surprising.

During mission execution the TIM system is able to alter PACT levels for any given task within the allowable range for that task. During periods of high workload, the system increases the automation levels to effect a reduction in workload, and *vice versa*. It has long been established that effecting changes such as those described shows measurable benefits in performance and workload (e.g. Parasuraman et al, 1995). However, these benefits are dependent on the accuracy of the workload measurement, and on the fidelity of the algorithms used to trigger automation changes.

The general aim when using some index of psychophysiological state as a trigger for adaptation (whether related to A-DFA or to information presentation) is to determine the points at which the workload is sufficiently high or low to initiate changes. The TIM system employs a low-pass filter to ensure that transient peaks and troughs do not cause rapid switching in the system, along with a simple threshold-based algorithm. The smoothing filter currently used by the TIM is the Savitzky–Golay (1964). These time domain filters remove noise while still preserving the true amplitudes and widths of the features. Each data value is replaced by a linear combination of itself and a number of nearby neighbors. The filtered value at each iteration is then passed to the thresholding algorithm. This algorithm takes five parameters:

χ^1	upper threshold
χ^2	lower threshold
φ	refractory period
δ	data window (number of leftward data points)
α	time since last state transition

and

$\{{}^{-\alpha}f, \ldots, {}^{-1}f, {}^{0}f\}$	the filtered data over the time period $[-\alpha, 0]$

After an iteration of the algorithm (with an additional new datum and less the oldest time point) the value of δ is incremented. If δ is less than φ then any possible state transition points will be ignored. This is the "refractory period". The reasoning behind this is that the adaptation is assumed to have an effect on the state being measured (e.g. mitigation reduces workload). Thus if the state met the criterion for being classified as "high", state transition would occur, as a result of which the individual's state might decrease to a level classified as "low", which would trigger a further change of state. Such a cyclic effect would clearly be undesirable and lead to a highly unstable system; hence the refractory period is introduced to determine a minimum time interval between state transitions. In order for the state to be classified as "high", ${}^{0}f$ must exceed χ^1. Conversely, for the state to be classified as "low", the value ^{0}f must be lower than χ^2. This, due to the prior smoothing of the data, ensures that state transitions occur only when data are consistently above the upper or below the lower threshold. It also has the effect of frequency-limiting the state transitions to reduce the switch rate. Given that the data may exhibit highly transient properties, this ensures that "spikes" in the data are filtered out. An example of the effect of this algorithm with state transitions shown as vertical lines can be seen in Figure 3.

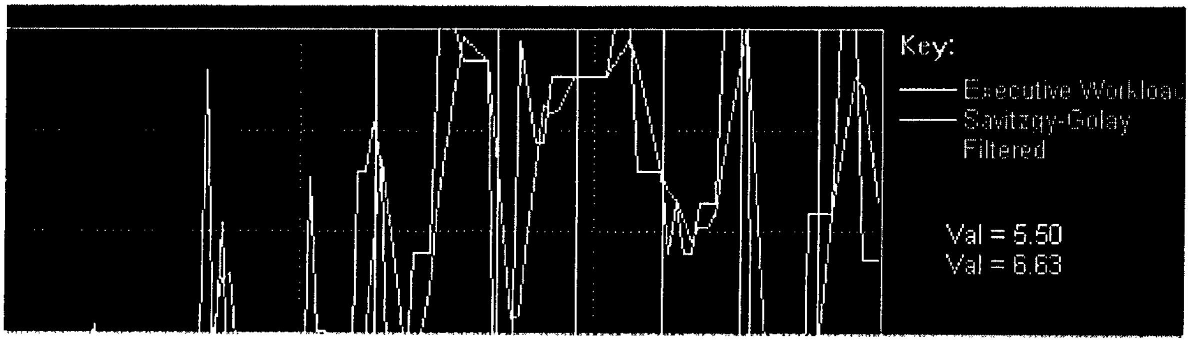


Figure 3: Filtered output from the CogMon with state transitions shown as vertical white lines and the upper threshold shown as a horizontal dashed red line (lower threshold not pictured).

Any tasks that are included in a plan from the DSS during this period are either presented to the operator as advice, or acted upon directly by the cockpit systems. Thus tasks that are critical to the success of the mission are supported during high-workload mission segments, whilst during lower workload segments automation levels are lower enabling the pilot to maintain high levels of situational knowledge and avoid degradation of his/her skill bases.

REFERENCES

Bonner, M.C., Taylor, R.M., Fletcher, K., & Miller, C. (2000). Adaptive Automation and decision aiding in the military fast jet domain. In Proceedings of the Conference on Human Performance, Situation Awareness and Automation: User-Centered Design for the New Millenium.

Endsley, M.R., and Kiris, E.O. (1995). The out-of-the-loop performance problem and level of control in automation. Human Factors, 37*(2)*, 381-394.

Jordan, C.S., Farmer, E.W., & Belyavin, A.J. (1995). The DRA Workload Scales: A validated workload assessment technique. Proceedings – Eighth International Symposium on Aviation Psychology, 2, 1013-1018.

Parasuraman, R., Mouloua, M., & Hilburn, B. (1999). Adaptive aiding and adaptive task allocation enhance human-machine interaction. In M.W. Scerbo, & M. Mouloua (eds.), Automation Technology and Human Performance: Current Research and Trends. Earlbaum, Mahwah, NJ.

Pleydell-Pearce, C.W., Whitecross, S., & Dickson, B.T. (2003). Multivariate analysis of EEG: Predicting cognition on the basis of frequency decomposition, inter-electrode correlation, coherence, cross phase and cross power. Proceedings – 36th Hawaii international conference on system sciences (HICSS).

Savitzky, A., and Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least square procedures, Analytical Chemistry, 36, 1627-1639.

Sheridan, T.B., and VerPlanck, W.L. (1978). Human and Computer Control of Undersea Tele-operators. Technical Report. MIT Man Machine Laboratory, Cambridge MA.

Taylor, R.M., Howells, H., & Watson, D. (2000). The cognitive cockpit: operational requirement and technical challenge. In P.T. McCabe, M.A. Hanson, and S.A. Robertson (eds.), Contemporary Ergonomics 2000, 55-59. London: Taylor and Francis.

Wickens, C.D. (1992). Engineering Psychology and Human Performance. New York: Harper Collins.

THE DAY "GOD" FAILED OR OVERTRUST IN AUTOMATION.
A PORTUGUESE CASE STUDY

José João Martins Sampaio
Sociologist
New University of Lisbon

António Abreu Guerra
Head of Safety and Human Factors Division
NAV - Portugal

ABSTRACT

The increasing development of computer based technologies open new horizons in task automation, helping pilots and air traffic controllers to carry out the analysis and resolution of an increasing number of cognitive tasks, in complex working environments. However, there is a general agreement that cognitive automation may lead to overtrust, complacency and loss of the necessary operational situation feed back, as the basis of the mental model refreshment which, in turn, allows for the maintenance of coherent situation awareness of all the operational processes.

The case study reported suggests there is a dimension to be followed in human machine integration, which is beyond the technological deterministic approach of human machine interface design, and calls for a better human comprehension of system nature. The human comprehension of this dimension, which we introduce as *the technological factor*, represents the basis of systemic self-constructed situation awareness, in a real human centered development.

Keywords: automation; situation awareness; mental model; overtrust in automation

INTRODUCTION

Situation Awareness is one of the most referred concepts, ever since the study of Operational decision Making Processes, in complex working environments, comes to discussion.

From the individual perspective to the team dimension, Situation Awareness evolved throughout many definitions and theories (Dominguez et al., 1994) either supporting the development of sophisticated measurement methods - Query Techniques, Rating Techniques, Performance Based Techniques - or showing the most effective design techniques and rules to integrate Human Factors in System Development.

But, being a complex cognitive process, situation awareness can hardly be disaggregated in a set of simple definitions, as those required to support automation algorithms. On the other hand, there is general agreement that cognitive automation may lead to overtrust, complacency and loss of the necessary operational situation feed back, as the basis of the mental model refreshment which, in turn, allows for the maintenance of a coherent situation awareness of all the operational processes.

Based on a reported incident at Lisbon ACC, this paper intends to discuss the limits of situation awareness in the context of human centred operational decision. Considering the hypothesis that cognitive automation, as an extension of human cognitive capabilities, will lead to the construction of *virtual extensions* (replacing comprehension by information) of human mental models, we introduce the concept of *technological factor* to be balanced against human nature development, as well as human factors are against technological development.

Situation Awareness and trust in Automation

Late 80's and 90's witnessed an enormous development of information technologies, which have been, in the aviation field, the basis for the implementation of new ground and airborne facilities and techniques towards an always greater rational use of the airspace, in response to a continued growing airline industry demand for more processing capacity.

This situation is the basis of a growing development of machine-automated tasks and information processing that has been under air traffic controller's responsibility.

But, automation may lead to data overload (Endsley & Esin, 1995; Grau, Menu & Amalberti, 1995; Woods, Patterson & Roth, 1995), stressing the air traffic controllers to rely on the automated system, as a virtual extension of their own mental models. ATC operators may find themselves in an automation overtrust situation, replacing comprehension by information and loosing control of one of the most important phases of human cognition process: the construction of self mental model on the operational environment (Wickens, 2002; Bonini, Jackson & McDonald, 2001; Dzindolet et al., 2000; Hollnagel, Cacciabue & Bagnara, 2000; Parasuraman 1997; Muir, 1994; Bainbridge, 1982; Hopkin, 1975).

Situation awareness will then tend to be system obtained – figure 1, and not operationally self-constructed. The Human operator may tend to follow and trust unreliable automation, even when there is an evident discrepancy conflict between automation and operational reported or visible evidence (Wickens, 1998).

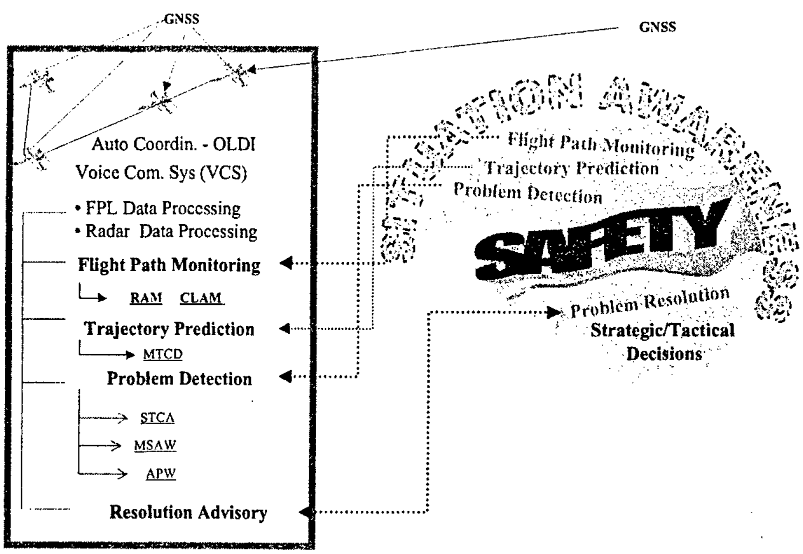


Figure 1– System based decision-making process

This tendency to (over)trust in automation as been well reported by a number of automation research studies (Rahman & Hailes,1995; Bisantz et al., 2000; Dzindolet et al..,2000; Muir, 1987) and it was also the main concern of the US National Research Council Committee on Human Factors study on human factors issues of ATC systems and technology: *efforts to modernize and further automate the air traffic control system should not compromise safety by marginalizing the human controller's ability to effectively monitor the process, intervene as spot failures in the software or environmental disturbances require, or assume manual control if the automation becomes untrustworthy.* (Wickens, Mavor & McGee, 1997, p. ix).

But what if there is no evidence of a system malfunction, while it really exists? What if the system information is so clear and normal, that there is no reason to assume that something is going wrong? How can the air traffic controller spot such an inconsistency of the information presented to him?

The answered is found in the concept of self-constructed situation awareness, as a dynamic/cybernetic cognitive process of checking and validating all the perceived information (mental picture) against cognitive mental model, allowing a coherent planning according to the foreseen future state of the operational environment. Only then, we can say that the air traffic controller may eventually, spot any "invisible" system inconsistencies, although this is virtually impossible in recent air traffic control automated systems, where, as we already said, comprehension is being replaced by more and more information, which has to be processed in real time and in a few seconds. For the air traffic controller, trustful information is fundamental for his job and that is the reason why it is out of the question to even presume that a normal shaped and well-presented automated information should be questioned.

Controllers are system believers. They just need to believe it exists and it's trustful. Like God.

The Day God Failed

Lisbon ATC centre sector was very busy. For that reason, phone coordination between control sectors had been replaced by the "automatic" procedure of assuming the traffic, at the moment it was spotted, on radar display by the next air traffic controller, some five minutes before entering the respective jurisdiction area. While being normal at rush hours, this procedure (resulting from the great knowledge and trust of all air traffic controllers in each other's work) implies that control is essentially radar supported, as no flight progress strips are manually pre-activated at the subsequent control sector.

The facts

At 1640 LMU134 calls for the first time Lisbon control (north sector) and, after squawking 3247, is radar identified.

At 1650, the pilot is told to contact Lisbon centre sector, and the controller of the centre sector asks the pilot to confirm the flight level 370.

At 1657 the air traffic controller had some doubts on the profile and correct position of LMU134, so he asked the pilot to squawk ident. After this new identification, and confirmation of the aircraft's position, the pilot was instructed to turn left, direct to VFA.

Still, three minutes later, the aircraft was showing a different heading that the one it should be flying, if routing direct to Faro. For that reason, the centre sector air traffic controller asked once more the pilot for confirmation, this time on the flying heading. The answered was that LMU134 was flying heading 203. But the radar was showing LMU134, on heading 226… At this time, the controller realised that something was wrong with the radar representation of LMU134.

Searching a reason for the discrepancy between the reported heading, and the one he was spotting in the radar display, the controller assumed the possibility of a mistake of his north sector collage, when assigning the SSR code to the aircraft, i.e., may be the track showing heading 226 would not be the one of LMU134. To verify this possibility, he searched for the LMU134 flight progress strips (remember they were not pre-activated, due to the *automatic* procedure, already mentioned) to confirm the SSR code mentioned there.

Once more, the SSR code allocated to the flight was correct: the flight progress strips showed code 3247, the same code north sector controller gave to the pilot and was displayed in the track's radar label.

After this, the controller thought there was still the possibility of an operator mistake at the flight data section, during the SSR code allocation procedure. So, he called the flight data section for confirmation of the correct SSR code of LMU134. And the answered was 3247…

From this moment on, the air traffic controller lost situation awareness towards LMU134, based on his own comprehension of the operational situation, and decided to adjust his mental picture to a refreshed mental model (after all the radar image was quite clear and trustful, and he had already checked every possible human error - pilot, flight data section and himself) based now on a situation awareness built exclusively on radar processed information.

At this time, DAL693 was also flying FL 370 and, according to the radar information on a parallel track to the LMU134, while XLB566 was flying north at FL 350.

Based on the refreshed mental model, after the checking procedures already mentioned, the position of the three aircraft left no doubt about the good separation between them. That is why, the air traffic controller found no reason for the TCAS advisory reported by the pilot of the LMU134, who requested descent, to avoid a traffic conflict. Anyway, and for the pilot's comfort, the controller decided to clear the descent of LMU134 to FL 350- fig.2 a).

This decision, while absolutely correct in relation to the information showed by the radar, and coherent with the refreshed mental model of the air traffic controller, created an additional air miss conflict between LMU134 (descending to FL 350) and XLB566 (maintaining FL 350) – fig.2 b).

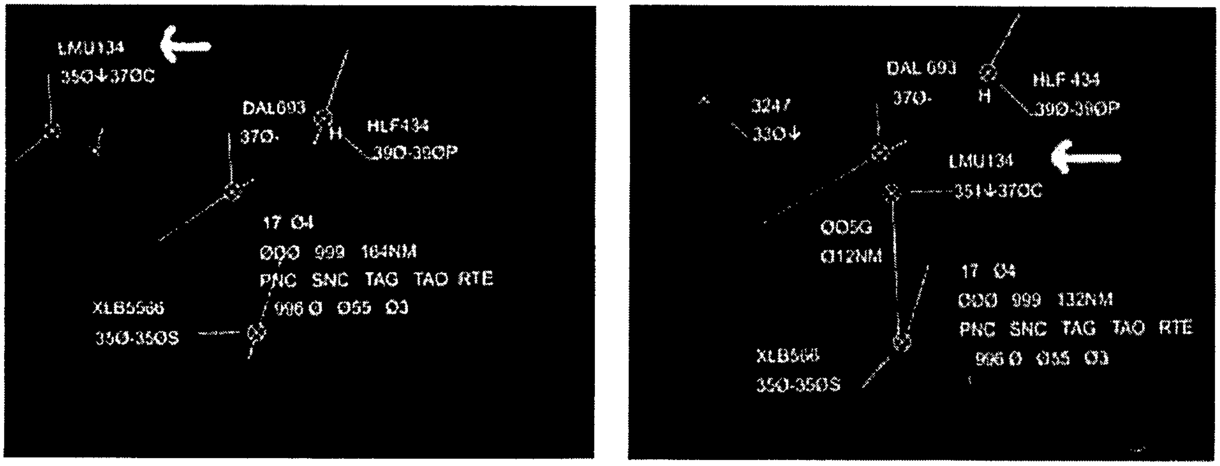


Figure 2. a) – The Radar Image b) – The Real Operational Situation

The Investigation

The investigation, which followed this events, showed that LMU134 has been in conflict with two other aircrafts, while the radar image shown no conflict at all.

The investigation also concluded there has been a real, *trustful*, and almost impossible to detect discrepancy, between the real position of LMU134 and the position processed and displayed by the radar data processing system. This situation lasted for 21 minutes and the real (correct) position of the aircraft could only be spotted in the radar display, for as much as 2 (two) seconds.

The main reason for this abnormal behaviour of the radar processing system, has been found in the incompatibility of the software developed for the recently installed monopulse radar antennas, and the software of the main system, installed in the mid eighties. Yet, there is still a question for which this explanation does not suit:

Why did it **only** happen with LMU134?

Discussion

When analysing this incident, there is a question everybody asks: "How could such a situation last for 21 minutes, without the air traffic controller realise it and find a correct solution?"

In fact, although being aware of the all situation, it took an 18-minute discussion to a group of three incident experts, to find out which kind of action should have been taken by the executive controller, instead of replacing his own constructed and comprehensive situation awareness, by a system processed one. Realising that information is the base of the decision-making process, the group concluded that, for the necessary psychological balance needed for his job, an air traffic controller has to trust the automated system, for the day he doesn't, safe and coherent decision will be replaced by uncertainty and ambiguity.

This incident was only possible because the air traffic controller trusted unconditionally the radar automated processed information. In fact, should he have used a procedural method of identification, for example, VOR/DME readings, he could have realized the correct geographical position of the aircraft.

But procedural control qualification doesn't exist anymore…

Another lesson learned is that in a situational awareness lost situation help is always needed, but no more than one person, preferably the operational supervisor, shall be involved. Otherwise, decisions become incoherent, as the air traffic controller will assume all kind of suggestions he will possible hear from the colleagues, trying to help. To avoid this situation, all air traffic controllers should be acquainted with TRM – Team Resource Management techniques.

NAV has already implemented this training as a routine in normal radar courses, where specific exercises are executed, along with different routine training, according to the specificities of each control unit.

Conclusion

This incident shows that automation needs to be balanced against human nature, but not exclusively in the field of human factors or cognitive ergonomics. Trust and overtrust in automation is an important dimension to be taken into consideration in future human centred technological development (Eurocontrol 2003). This means that, along with the development of error tolerant systems to cope with possible human errors, humans need to be trained in an automation error tolerant perspective, as well, i.e., operational training based on a system nature understanding in a comprehensive way, allowing humans to evolve from system operators to real in-loop system managers.

This approach, including **technological factors** in human training goes beyond user adaptation to automation. It has to be understood in a systemic interaction perspective, where the real interface between humans and machines is each own nature.

While this integrative dimension is not achieved, we will have human error tolerant systems development to be operated by unconditional system believers.

As we said before, that is the case of air traffic controllers. So, what else could have been done, that the controller didn't? One must remember there was no evidence of a system malfunction, whatsoever. "Only" the processed information and the expected one, for that particular flight, didn't match...

Everybody agreed it is not easy, when there is no evidence of a system error, to reject the system automated processed information and assume entire responsibility for that. In these circumstances it is more acceptable, for the air traffic controller, to doubt his own perception and comprehension of the operational situation, than to question the system. After all, "God" doesn't fail!

But, this time "He" did.

REFERENCES

Abdul-Rahman, A. & Hailes, S. (1999). Relying on trust to find reliable information. 1999 International Symposium on Database, Web and Cooperative Systems (DWACOS'99), Baden-Baden, Germany.

Bainbridge, L. (1983). Ironies of automation. http://www.bainbrdg.demon.co.uk/Papers/Ironies.html [13-04-2003 3:10:30].

Bonini, D., Jackson, A. & McDonald, N. (2001). Do I trust thee? An approach to understanding trust in the domain of air traffic control. Proceedings of People in Control, 19-21 June, UMIST Manchester.

Bisantz, A.M., Llinas, J., Seong, Y., Finger, R. & Jian, J.Y. (2000). Empirical investigations of trust-related system vulnerabilities in aided, adversarial decision-making. Center for Multi-Source Information Fusion, Dept. of Industrial Engineering, State University of New York at Buffalo.

Cardosi, K. & Murphy, E. (1995). Human Factors in the Design and Evaluation of Air Traffic Control Systems. Federal Aviation Administration, Office of Aviation Research, DOT/FAA/RD-95/3.

Dominguez et al., 1994. Situation Awareness: Papers and Annotated bibliography. Armstrong Laboratory, Human System Centre, ref. AL/CF-TR-1994-0085.

Dzindolet et al. (2000). Buildings trust in automation. Paper presented at Human Performance, Situation Awareness and Automation: User-Centered Design for the New Millennium, The 4th Conference on Automation Technology and Human Performance and the 3rd Conference on Situation Awareness in Complex Systems, October15-19.

Endsley. Mica & Esin Kiris (1995). Situation Awareness, Global Assessment Technique (SAGAT) TRACON Air traffic Control Version User Guide, TTV-IE-95-02, Texas Tech University.

Eurocontrol (2003). Guidelines for Trust in Future ATM Systems: A Literature Review. EATMP Infocentre, Brussels.

Grau, J. Y., Menu, J. P., Amalbertyi, R. (1995). La Conscience de la Situation en aéronautique de combat. Proceedings of AGARD conference 575, Situation awareness: Limitations and Enhancements in the aviation environment.

Hollnagel, E., Cacciabue, P.C. & Bagnara, S. (1994). The limits of automation in air traffic control. Workshop report, Int. J. Human-Computer Studies, 40, 561-566.

Hopkin, V.D. (1975). The controller versus automation. AGARD AG-209.

Muir, B. (1994). Trust in automation: Part 1. Theoretical issues in the study and human intervention in automated systems. Ergonomics, 37, 1905-1923.

Muir, B. (1987). Trust between humans and machines, and the design of decision aids. Int. J. Man-Machine Studies, 27, 527-539.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230-253.

Wickens, Christopher D. & Xu, Xidong (2002) .Automation Trust, Reliability and Attention. HMI 02-03 Technical Report, AHFD-02-14/MAAD-02-2, October.

Wickens , Christopher D. (1998). Automation in Air Traffic Control: The Human performance issues, in Automation technology and Human Performance: current research and trends, edited by Scerbo & Mouloua.

Wickens, C.D., Mavor, A.S. & McGee, P. (1997). Flight to the future. Human factors in air traffic control. Commission on Behavioral and Social Sciences and Education, National Research Council, National Academy Press.

Woos, David D., Patterson Emily S, & Roth Emilie M. (1998). Can we ever escape to data overload? A Cognitive Systems Diagnosis- Cognitive Systems Engineering Laboratory Institute for Ergonomics, the Ohio State University.

WHEN EVERY MINUTE COUNTS, ALL AUTOMATIC EXTERNAL DEFIBRILLATORS ARE NOT CREATED EQUAL

Anthony D. Andre and Jennifer R. Kingsburg
Interface Analysis Associates
San Jose, CA

ABSTRACT

Sixty-four adults with no prior exposure to, or training with, Automatic External Defibrillators (AEDs) were asked to rush into a room and attempt to use an AED to resuscitate a simulated victim of sudden cardiac arrest. Each of four commercially-available AEDs was used by a different group of sixteen participants. The results demonstrate that not all AEDs are equally usable by untrained laypersons and that while some AEDs are appropriate for use in public settings, other AEDs are not. The results of this study are used to highlight the beneficial use of automation in life-saving products intended for layperson use and the specific interface design attributes that lead to effective user-AED interaction.

Keywords: Automated external defibrillator; AED; Public use; Automation

INTRODUCTION

Sudden cardiac arrest is a leading cause of death in the United States. The American Heart Association (AHA) estimates that about 250,000 people die of coronary heart disease before reaching the hospital each year (AHA, 2002). Unlike many other life-threatening illnesses and conditions, sudden cardiac arrest often occurs outside of a medical setting. In such settings, the victim's only chance for survival rests with the use of a defibrillator, a device that delivers a shock to the heart. During sudden cardiac arrest, every minute counts. In fact, for every minute that goes by without defibrillation, the chance of survival decreases dramatically (AHA, 2002).

The Impact of Automation

The use of automation has made a great impact on the recent design of life-saving devices, such as AEDs, and has contributed significantly to allowing public access to, and effective use of, these devices. In the past, defibrillators were used only by trained medical personnel and required the user to manually determine if defibrillation was necessary, and if so to then manually set various parameters to optimize the defibrillation/shock delivery.

The advent of intelligent analysis algorithms, which rapidly and automatically assess the patient's heart rhythm to ensure that a shock is delivered only if it is appropriate, along with waveform automation, which determines the most effective form of shock to deliver (with the goal of delivering the right amount of electrical current on the first shock), together define the main "automatic" aspect of AEDs. Further, most AEDs use automation logic to perform and interpret periodic self-tests of the battery, electrical components and critical subsystems.

Collectively, these and other forms of automation have allowed for AEDs that are smaller, more reliable, use lower and safer energy levels and provide superior clinical performance relative to their manual predecessors.

The Usability Factor

Recently, there has been a surge of interest in the placement of automated external defibrillators (AEDs) in public environments. For example, AEDs can now be found in airplanes, airports, schools, shopping malls, and various workplaces. In some of these environments, selected individuals (e.g., flight attendants) are trained to use the devices. However, in order for these devices to be practical for broad public use, they must be designed in a way that allows untrained "ordinary" people to use them quickly, easily, and effectively in the context of an unexpected and dramatic emergency medical situation (Caffery, 2002). This premise represents a significant challenge to AED manufacturers, many of whom have historically designed devices to be used by trained medical professionals (e.g., nurses, EMTs) or selected individuals (e.g., lifeguards).

As usability professionals, we make a clear distinction between a product's functionality (what a product can do) and its usability (what users can do with the product). While all AEDs share a common set of functionality and, if used correctly, result in the delivery of a shock to the victim, the objective and subjective experiences of the users are likely to vary based on the presence or absence of critical automation and usability design attributes. To date, there is little if any empirical information on usability differences between AEDs intended for public use. Thus, it is not known if all AEDs can equally support the successful use by untrained persons.

A COMPARATIVE STUDY

To address this concern, we conducted a comprehensive and comparative study of four leading AEDs, all available for public use environments. Each of the four AEDs was used by a different group of sixteen participants. The four devices included in the study were: 1) Cardiac Science Powerheart, 2) Medtronic CRPlus, 3) Philips HeartStart OnSite, and 4) Zoll AED Plus.

Participants

Sixty-four adult participants, ages 35 to 55, representing a variety of occupations, were asked to rush into a room and attempt to use an AED to resuscitate a victim of sudden cardiac arrest. None of the participants worked in medical or related fields, nor did they have any exposure to, prior training, or familiarity with AEDs.

Procedure

The study was conducted in the context of a scenario where AEDs are available in a variety of public settings such as shopping malls, schools, sporting events, etc. The participants were provided only basic information about the main functions of an AED prior to their entering the room, where they found a fully clothed manikin on the floor and one of the four AEDs nearby. The manikin was wired with a simulator that allowed it to transmit signals to the electrode pads of each AED, which prompted the unit to advise a simulated shock to the manikin (under conditions similar to those that would produce a shock command in actual use).

A comprehensive variety of quantitative, behavior, and subjective measures was collected and analyzed. Electrode pad placement measures were reviewed by three members of the research team and later confirmed by an independent reviewer (Dr. Jeanne E. Poole, Associate Professor of Medicine, Acting Director of the Arrhythmia Service and Electrophysiology Laboratory, and Attending Physician, University of Washington Medical Center).

RESULTS

Failure to Deliver Therapy

Clearly, the most important measure was the frequency with which untrained users could deliver a shock with the AED. Nine of the 16 Zoll users (56%) and 4 of the 16 Cardiac Science users (25%) failed to administer a shock to the simulated victim. In contrast, the Philips and Medtronic users were successful in delivering a shock in all completed trials.

It is of interest to note the user behaviors that resulted in the failures to deliver therapy for two of the AEDs (see Figure 1). For example, two of the Zoll users and three of the Cardiac Science users never managed to open the electrode pad package (see Figure 1, left), while another group of five Zoll users placed the electrode pads directly over the victim's clothes (see Figure 1, middle). Still another two Zoll users and four Cardiac Science users failed to remove the liner from one or both electrode pads (see Figure 1, right), though three of the four Cardiac Science users who failed to remove the pad liner still received a shock command, a potential artifact of our simulator.

Time to Deliver Therapy

Managing to get the device to deliver a shock is a necessary but not sufficient goal, as the victim must be shocked within a short period of time from the point of collapse. In our study, the Medtronic and Philips devices were equivalent in the time it took their users to deliver a shock, both averaging well under two minutes at 101.0 and

101.5 seconds, respectively. The other two devices were substantially slower, with the Cardiac Science AED averaging just over 2.5 minutes (151.6 s), and the Zoll AED averaging just under 4 minutes (225.1 s).

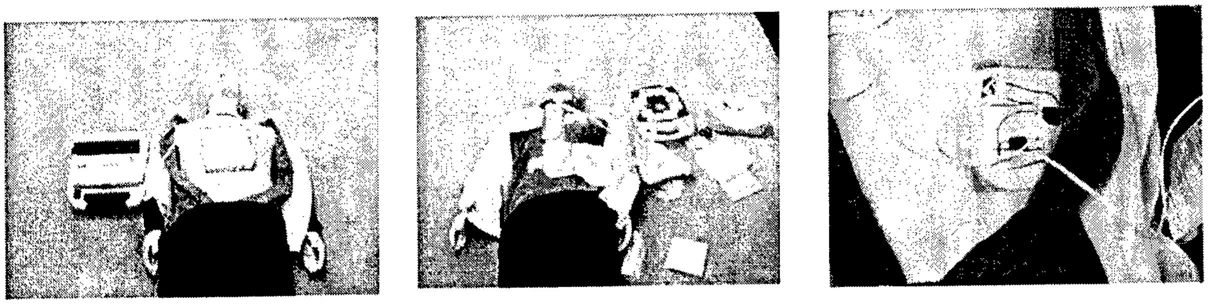


Figure 1. This Cardiac Science user never removed the electrodes from their package (left); This Zoll user placed the electrodes over the victim's clothes (middle); This Cardiac Science user failed to remove the pad liner (right).

Electrode Pad Placement

Pad placement has been well documented as the Achilles heel for lay responders and those with advanced training alike (Mattei, Mackay, Lepper & Soar, 2003; Heames, Sado & Deakin, 2001). Incorrect pad placement results in a decreased percentage of the current passing through the heart, thus reducing the chance of successful defibrillation (Ewy & Bressler, 1982).

As noted earlier, several Zoll and Cardiac Science users demonstrated difficulty in manipulating the electrode pads. For those users who managed to properly place the electrode pads on the victim's bare chest, the quality of the resultant shock was evaluated as a function of the following four parameters: 1) percentage of skin contact, 2) pad location error, 3) inter-pad separation, and 4) inter-pad alignment. Table 1 shows the relative pad placement measures and rankings between the four AEDs.

Table 1. Pad Placement Measures and Rankings.

AED Device	% Skin Contact	Rank	Location Error (avg cm)	Rank	Pad Separation (avg cm)	Rank	% of Pads Adjacent	Rank	Overall Rank
Cardiac Science *Powerheart*	84%	3	7.0	3	10.4	3	0%	1	2 (tied)
Medtronic *CR+*	94%	2	10.4	4	9.0	4	56%	4	3
Philips *HeartStart OnSite*	97%	1	5.4	2	14.7	1	6%	2	1
Zoll *AED Plus*	76%	4	4.9	1	13.9	2	11%	3	2 (tied)

Across the four measures, the Philips device resulted in best pad placement performance, while the Medtronic device yielded the worst pad placement performance. For example, over 50% of the Medtronic AED users placing the pads adjacent to each other (see Figure 2), an arrangement that would often result in shunting between the pads, and a less effective shock.

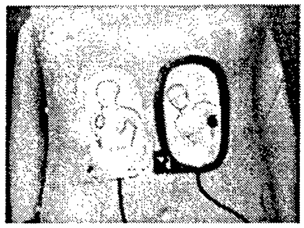


Figure 2. This Medtronic user placed the electrode pads adjacent to each other.

Subjective Data

The Philips and Medtronic devices were consistently rated as easier to use, across a variety of dimensions, relative to the Cardiac Science device and lastly to the Zoll device.

DISCUSSION

We conclude that the probability of lay responders successfully defibrillating a cardiac arrest victim is greater for some AEDs than others. In this study, when combining all measures of performance, behavior and subjective experience, the Philips AED stood out as the most usable device relative to the three other AEDs. These findings were corroborated by a recent independent university study (Eames, Larsen & Galletly, 2003) that also found small differences in time to shock and ease of use ratings between the Philips and Medtronic devices, but large differences in pad positioning accuracy in favor of the Philips AED. Further, and again similar to our findings, they found the Zoll AED to be the most difficult to use across all measures.

It's All About Context

To understand the underlying performance and behavioral differences between the four AEDs it is useful to first discuss what happens to people in an emergency, when they are emotional, scared, time pressured, etc. Lights, sounds, and shapes; things referred to as electrodes, wires, shock buttons—all must be interpreted while the body experiences severe physiological and psychological changes. In this context, users likely operate in a knowledge-based mode with an external locus of control, relying on the product to guide their interaction and responding only to explicitly-provided instructions. In addition, the stressful nature of the situation is likely to induce cognitive tunnel vision whereby users only perceive or process a small sample of the information environment (Stokes and Kite, 1994). Now, let's consider how these devices differently approach supporting users in this context, from the perspective of four design dimensions.

Automation. All of the AEDs, except the Zoll device, automatically turn on, and begin to annunciate the directions, when the unit is opened. This turned out to be a critical feature, as the average time for users to figure out how to manually turn on the Zoll device was nearly equal to the total time needed to shock the victim with the Medtronic and Philips devices. Further, many of the non-optimal behaviors exhibited by Zoll users (e.g., placing pad over clothes; not removing pad liners) can be attributed to their attempt to apply the electrode pads to the victim without having turned the device on, and thereby not receiving the voice instructions.

Explicit guidance. Recall that some of the Cardiac Science users failed to either remove the electrode pads from the package or to remove one of the pad liners (see Figure 1). These errors can be traced to the vague and implicit instruction annunciated when the AED is opened. It says "Place electrodes on patient's bare chest." Note that it says nothing about taking the pads out of the package, or about removing pad liners.

In contrast, an instructional design element that was observed to have helped Philips users to achieve the best pad placement performance was the explicit voice instruction "Look carefully at the pictures on the white adhesive pads… Place pads exactly as shown in the picture." This instruction, unique to the Philips device, often resulted in the users briefly pausing and explicitly reviewing the pad placement graphic before placing the pad on the victim's chest.

Interface design. Taking advantage of the user's attention to the pad graphic, created by the aforementioned auditory instruction in the Philips device, a design feature that aided users in their pad placement accuracy, was the fact that both pads are shown on each pad graphic, giving users a good sense of the relative placement of the two pads (see Figure 3).

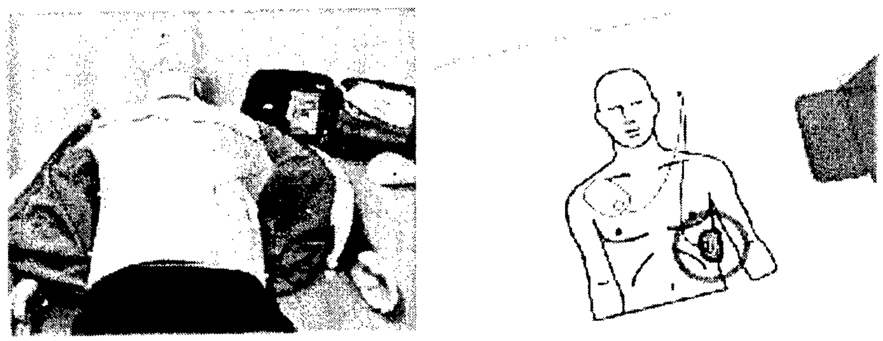


Figure 3. The Philips AED depicts the relative placement of both pads on each pad graphic.

Intelligent pacing. A final explanation for the higher levels of task conformance among Philips users is the device's incorporation of intelligent instruction pacing. This device includes sensor technology that detects the current action of the user and adjusts the instructions to match that action. Indeed, we observed many instances where the Philips users were aided by the intelligent pacing of the device's audio instructions. In contrast, we observed many instances with the other AEDs where the audio instruction and the user's current action were incongruent.

The Devil is in the Details

Many potentially useful device attributes were rendered dysfunctional by the chosen design implementation. The most detrimental example is the design of the pad connector plug on the Medtronic device. An astonishing 31% of the Medtronic users inadvertently pulled the pad connector plug out of its socket while attempting to open the pad package, causing them to spend precious time hunting for the place to put the plug back in. We attribute this frequent problem to both the design of the pad package (which encourages users to grasp a red handle and pull the entire package away from the device) and the ineffectiveness of the design of the cable strain relief.

Another example of a good idea "gone wrong" is the Zoll cover. Users of this device are instructed, via graphics, to use the device cover to help prop up the victim and open their airway. However, this implicit graphic instruction that is too small to clearly differentiate the proper orientation of the cover, resulted in at least one case where the user cut off, rather than opened up, the victim's airway.

CONCLUSION

Defibrillators that are to be used by lay responders should be designed from a human-centered perspective. That is, they should provide explicit, useful and timely guidance, include effective and salient graphics, icons and labels, and induce acceptable levels of workload and stress. This study demonstrates that all automated external defibrillators are not alike. While all AEDs are potentially useful life-saving devices, only some are acceptably usable in the public-use context simulated in this study. We encourage AED manufacturers to consider the unique context of public AED use, and to design future AEDs that address the specific perceptual, information processing and instructional needs of lay responders.

ACKNOWLEDGMENTS

This study was funded by Philips Medical Systems, makers of the HeartStart OnSite AED. The authors, who had sole control over the study design, methodology, analysis and conclusions, made every attempt to conduct an objective and unbiased comparison of the four AEDs. The authors thank David Hird for his valuable contributions to the conduct and analysis of this study.

REFERENCES

American Heart Association (2002). Heart Disease and Stroke Statistics - 2003 Update. Dallas, Texas: American Heart Association.

ANSI/AAMI DF39-93 Standard (1993). Automatic External Defibrillators and Remote Control Defibrillators.

Caffery, S. (2002). Feasibility of public access to defibrillation. Current Opinion in Critical Care, 8, 195-198.

Caffrey, S., Willoughby, P., Pepe, P., & Becker, L. (2002). Public use of automated external defibrillators. New England Journal of Medicine, 347, 1242-1247.

Eames, P., Larsen, P.D., & Galletly, D.C. (2003). Comparison of ease of use of three automated external defibrillators by untrained lay people. Resuscitation, 58, 25-30.

Ewy, G.A., & Bressler, R. (1982). Cardiovascular drugs and the management of heart disease, NY. Raven Press.

Heames, R.M., Sado, D., & Deakin C.D. (2001). Do doctors position defibrillation paddles correctly? Observational study. British Medical Journal, 322, 1393-1394.

Mattei, L.C., Mackay, U., Lepper, M.W., & Soar, J. (2002). Do nurses and physiotherapists require training to use an automated external defibrillator? Resuscitation, 53, 227-280.

Stokes, A., & Kite, K. (1994). Flight Stress. Brookfield, VT: Ashgate.

INDIVIDUAL AND TEAM PERFORMANCE

From the Perceptual to the Organizational: The Science of Expertise and the Practice of Human Performance

Florida Alliance for the Study of Expertise[1]

ABSTRACT

In this paper we describe representative samples of the research efforts being conducted by the "Florida Alliance for the Study of Expertise" (FASE). This is a recently formed organization of scientists whose goal is to advance a science of *Expertise Studies*. FASE focuses on the entire human system and how experience alters this system to produce *meaningful learning* that leads to the highest levels of human performance.

Keywords: Expertise, Perceptual Learning, Stress, Arousal, Problem Solving, Team Cognition, Organizational Dynamics, Organizational Modeling

INTRODUCTION

At its core, Expertise Studies is a science of human learning and performance. Researchers investigating expert performance have developed a strong foundation of knowledge associated with mastery in a variety of domains. This includes a similarly varied set of differing forms of expertise, ranging from perceptual and motor skills to complex conceptual and organizational knowledge. Over the past several years there has been a remarkable convergence in which a considerable number of leaders in the study of expertise have joined the faculties of Florida Universities. In this paper we describe representative samples of the research efforts being conducted by the "Florida Alliance for the Study of Expertise" (FASE). This is a recently formed organization of scientists whose goal is to take advantage of this convergence so as to advance a science of *Expertise Studies*.

FASE focuses on the entire human system and how experience alters this system to produce *meaningful learning* that leads to the highest levels of human performance. FASE considers learning and performance broadly and takes both a componential approach to the science of expertise as well as a representative approach so as to insure fidelity to the contexts in which domain practitioners actually work. In this paper we first briefly describe the historical context of expertise studies and then illustrate how FASE supports research on how the human system achieves levels of exceptional performance in areas ranging from the perceptual to the organizational.

Historical Context of Learning and Expertise Research

Understanding learning and performance at exceptional levels is not a new concept. Hundreds of years ago it was recognized as an important milestone in skill development in the traditional "craft guilds" of the Renaissance. This early thinking gave rise to the notion that learning and education can proceed by understanding and assimilating the skills of experienced practitioners. Indeed, modern studies of expertise still rely on the expert-journeyman-apprentice classification scheme (Hoffman, 1998). The value of the study of expertise was recognized by a number of relatively independent disciplines in the 1970s. For example, psychologists who were interested in human learning began to study the differences between novices and experts in such domains as chess (deGroot, 1965; Chase & Simon, 1973) and physics (e.g., Chi, Feltovich, & Glaser, 1981). Subsequent investigations of expertise found that individuals who have reached the highest levels of performance, in a wide range of domains, have behind them at least ten years of experience (Chi, Glaser, & Farr, 1988; Simon and Chase, 1973). Expertise was similarly recognized by computer scientists in the late 1970s during the development of first-generation "expert systems." Creating these expert systems required computer scientists to interview experts to glean their domain knowledge and

---

[1] FASE represents the collaborative efforts of a number of scientists affiliated with Florida Universities. In alphabetical order, they are Irma Becerra-Fernandez, Jeff Bradshaw, Neil Charness, William Clancey, David Eccles, Anders Ericsson, Paul Feltovich, Stephen Fiore, Peter Hancock, Laura Hassler, Robert Hoffman, Christopher Janelle, Tristan Johnson, Mike Prietula, Eduardo Salas, Jim Szalma, and Gershon Tenenbaum. Writing this paper was partially supported by a National Science Foundation Grant awarded to Eduardo Salas and Stephen M. Fiore. For questions or comments, please contact Stephen M. Fiore (sfiore@ist.ucf.edu).

their reasoning rules. To meet this need, the emerging discipline of cognitive science, which encompasses both human and machine cognition, began to concern itself with the methodology of "expert knowledge elicitation" (see Hoffman, Shadbolt, Burton, & Klein, 1995).

Importantly, the study of expertise forced the research community to broaden its approach in that theories of human learning and performance needed to address how cognition is exercised in the "real-world" by mature, knowledgeable, and highly skilled individuals engaged in complex and difficult task domains. Cognitive scientists came to recognize that theories of cognition have to account for the nature of experts' superior performance, including their impressive knowledge and memory. This meant looking outside the traditional academic laboratory and has required a considerable expansion of the methods and tools that are used, not just by social scientists, but also by scientists in a number of disciplines. With further studies of experts, such as airline pilots, medical doctors, athletes, and chess masters, it became clear that expertise requires more than just knowledge acquisition and simply applying past knowledge (Ericsson, 1996; Salas & Klein, 2001). We turn next to a discussion of representative samples of research by FASE Associates illustrating the far ranging implications for understanding human learning and performance at exceptional levels.

RESEARCH BY FASE ASSOCIATES

Attention and Performance in Expertise Studies

Methodological advances have allowed researchers to broaden their understanding of expertise to include physiological indicators of expert performance (e.g., eye movements and bioelectric signals such as EEG, see Janelle & Hillman, 2003). Innovations linking physiology, basic cognitive processes and performance have illustrated the degree to which these techniques can converge on a finer-grained understanding of factors driving learning and performance. A recent focus of this research has centered on the coupling between visual search patterns and other psychophysiological indices of attention and arousal (such as the spectral characteristics of the electroencephalogram [EEG]), particularly among expert and non-expert performers. For example, under the category of "mind-eye connection," Janelle and colleagues have conducted exploratory investigations among expert and novice small-bore rifle shooters. These studies investigated how pre-shot EEG correlates of arousal and attention (alpha and beta spectral frequencies), relate to gaze behavior characteristics. Eye movements and EEG activity were concurrently measured over the course of a regulation round of shooting. Findings indicated that the two measures were associated with shooting performance and that they accounted for a significant amount of the shooting variability (49%) between expert and novice marksmen (see also Janelle et al., 2000).

Related research investigated expert/novice differences in baseball pitch recognition, in part by examining differences in event-related cortical potentials (ERPs; specifically the P3) in the context of a modified cost-benefit paradigm (Radlo, Janelle, Frehlich, & Barba, 2001). These studies found that intermediate batters exhibited shorter P3 latencies, larger P3 amplitudes, and longer RTs than advanced batters, with the effect more pronounced for curveballs. These findings suggest a comparative ease by which experts are capable of minimizing attentional/anticipatory costs and thus maximizing benefits so as to improve performance.

Stress and Performance in Expertise Studies

Understanding how stress interacts with complex human performance allows us to converge on a deeper understanding of the interaction between exceptional levels of skill and the moderating effects of stress. Within this area, FASE researchers are engaged in investigations that examine the attentional mechanisms underlying human performance under conditions of high stress and workload. One goal for this programmatic research is to develop a comprehensive theory of stress and performance that will underpin the design of training protocols and human-technology interfaces to reduce negative stress effects.

This approach to stress builds primarily upon the extended-U model described by Hancock and Warm (1989). This model specifies two aspects of task-based stress that impact performance: information rate (the speed with which demands are made) and information structure (the complexity of that demand). Information rate represents the temporal component of task demand, while the information structure is often represented in a spatial format. The combined space-time variations in task and environmental demand impose considerable stress on experts, to which they resist via coping efforts. Breakdown of performance under stress and its inverse, behavioral adaptability occurs at both psychological and physiological levels with psychological adaptability failing before comparable physiological adaptability (see Matthews, 2001).

Since time and space are integral dimensions of stress demand, one central facet of this research is exploring disturbance to spatial and temporal features of task performance under stress, often manifested in distortions of spatial and temporal perception resulting from attentional narrowing. It is likely that attentional narrowing along these dimensions have a common resource mechanism (Hancock & Weaver, in press), a proposition currently being tested. Initial results support a common capacity view, but the spatial dimension may be more salient than the temporal dimension. Last, when viewing the Hancock and Warm model in the context of expert performance, it could be hypothesized that the top of the U curve would extend further in experts. Specifically, the threshold for declines in behavioral adaptability would increase, since experts have the skills to more effectively cope with stress, particularly task-based stress. Indeed, such notions support theoretical approaches put forth in analogous domains, a topic we discuss next.

Athletic Expertise

Expertise Studies also encompasses athletic skill, and sports psychologists have studied learning and performance from across the continuum of skill, from novice to expert. As in other domains of expertise, in order to better understand the skill acquisition process, these studies often contrast experts and less skilled performers in terms of the cognitive skills and strategies they bring to bear on their tasks. Within this context, a variety of methods have been employed to investigate the differing skills acquired across sports domains. These include processes tracing measures, such as verbal protocol analysis, eye and head movement tracking, and occluded visual display paradigms, and self-report measures, such as retrospective interviews (e.g., Eccles, Walsh, & Ingledew, 2002a; 2002b; Starkes, & Ericsson, 2003; Tenenbaum, & Elran, 2003; Williams, & Hodges, 2003)

Advances in expertise studies have set the stage for an understanding of the emotional and motivational aspects of expert performance, such as coping strategies that enable experts to sustain a "zone of optimal functioning" in a variety of conditions (Kamata, Tenenbaum, & Hanin, 2002). The regulation of emotions has implications across a broad range of human performance, ranging from the military to artistic to athletic domains. Understanding the complex interplay between stress and performance has long challenged the psychological sciences, and now, with improvements in measurement and in theory, we are converging on a better understanding of the complex interplay between physiological, psychological and cognitive regulation. For example, in athletic domains, technical expertise is a necessary but not sufficient prerequisite to successful performance. Expertise in sports also requires an athlete to effectively regulate their emotional response to a situation. Related studies show that experts can anticipate unfolding events and can reduce uncertainty so that they can prepare for decision-making and action under time pressure (Ericsson & Kintch, 1995; Tenenbaum, in press; 2003; Tenenbaum & Bar-Eli, 1993; 1995; Tenenbaum, Levy-Kolker, Sade, Lieberman, & Lidor, 1996; Tenenbaum & Lidor, in press; Ward & Williams, 2003; Williams, David, & Williams, 1999).

Additionally, studies within expertise in sports show how experts utilize environmental resources so as to distribute "mental workload" across time (Eccles et al., 2002b). FASE researchers are also investigating the potential influence of emotion on attentional processing, specifically with regard to the mechanisms underlying visual selective attention. Janelle and colleagues are examining the search patterns of performers in competitive situations to determine how emotional reactivity might influence eye tracking patterns and potentially, performance. Using a racecar driving simulation these studies show reliable differences in search patterns, such that search strategies are significantly different when under stressful conditions as opposed to relatively benign conditions (Janelle, Singer, & Williams, 1999; Murray & Janelle, 2003). Tenenbaum and colleagues at FSU have similarly worked with athletes to understand how stressors and anxiety alter attentional capacity in complex tasks to predict vulnerability to choking. Through such work greater insight is being gained concerning what experts do to maintain a state of focused attention that permits automated and effective performance.

The aforementioned studies form an important component to our understanding of the complex interplay between stress and performance by evaluating the "how" and the "why" of the mechanisms underlying the efficiency and effectiveness of performance under stress. Understanding such processes in differing domains can inform our understanding of stress response and management in other domains such as military operational environments where the regulation of emotion can be critical to survival.

Complex Problem Solving

Results from studies within athletic tasks requiring not only high levels of motor skill, but also complex cognitive processes (e.g., orienteering) illustrate similar patterns of performance with respect to the differences between experts and novices. For example, experts differ from novices in terms of the knowledge they possess about their domains, and experts develop memory skills that affect the way that this knowledge is stored and accessed during performance. The expert's knowledge affords them cognitive skills and strategies that make the execution of their task highly efficient, such that they can effectively circumvent the natural limitations of visual and neural systems. Furthermore, the experts' memory skills better support the planning, monitoring, and evaluation processes inherent in expert sports performance. For example, Eccles and colleagues (e.g., Eccles et al., 2002a; 2002b) have studied expertise in the sport of orienteering, which requires the performer to navigate, using map and compass, through a series of checkpoints in wild terrain, as fast as possible. A key task constraint in the sport is the requirement to attend to the map and compass, features in the terrain, and to one's running, so as to avoid tripping or colliding with hazards. Attending to each source of information simultaneously is problematic owing to natural human visual and attentional limitations. However, expert orienteers develop attentional scheduling strategies to circumvent this resource limitation, and, in turn, performance is enhanced.

Considering this in the broader context of human performance, these findings suggest that resource limitations can be similarly surmounted. For example, consider methods of augmented cognition where head-mounted displays are providing navigational information to military personnel. Although studies with head-mounted displays are still in their development, findings from sports such as orienteering should be leveraged to show how situation assessment processes can be supported in these forms of augmented cognition and how it is that learning can proceed to support such attentional scheduling strategies.

With regard to problem solving and decision making, research shows how experts are able to rapidly grasp problems, seemingly with little search through a problem space (e.g., Reingold, Charness, Pomplun, & Stampe, 2001; Salas & Klein, 2001). For example, Charness and colleagues suggest that, underlying such behavior, are superior pattern recognition processes that allow the problem solver to rapidly develop effective problem representations (for a discussion, see Charness, 1991). As such, their superior knowledge base allows them to bypass search processes as they engage in problem solving and decision making tasks.

FASE researchers have also been studying the learning, understanding, and application of difficult subject matter, in particular learners' understanding of flow systems. The term *flow systems* encompasses systems at both large scales (e.g., the atmosphere and watersheds) and small scales (e.g., the cardiovascular system). Understanding such complex dynamical constructs represents an important challenge to the welfare of mankind given that misinterpretation of factors within a system or mismanagement of these factors can have devastating consequences. For example, studies show that in South Florida changes in land use due to farming have altered waterflow (e.g., draining wetlands) and consequently, the local atmosphere, to produce a greater number of freezes (Marshall, Pielke & Steyeart, 2003). Given the causal and dynamical complexity of flow systems, they are both very difficult to understand and to manage. Within the field of expertise studies, Feltovich and colleagues have identified characteristics of such subject matter that cause difficulty for learning. This includes dynamics (constant change), high interdependence of multiple variables, and continuity (rather than a step-by-step nature) of processes -- all characteristics of flow systems. Accompanying these characteristics is a pervasive tendency for learners to over-simplify this form of subject matter. This phenomenon, termed "reductive bias," suggests that dynamic factors may be treated as static, continuous factors may be treated as discrete and step-wise, etc. (see Feltovich, Spiro, & Coulson, 1997; Spiro, Coulson, Feltovich, & Anderson, 1994). This human tendency to create initial understandings and explanations that over simplify can lead to misconceptions and errors when applied to complex systems.

Within the field of expertise studies, recent work is concentrating on how it is that experts who understand complex flow systems are able to overcome the reductive bias in their work with systems of flow. In addition to the epistemological gains such research will provide, an important goal is also to develop the capability that will allow educators to determine how to accelerate novices' understanding of such complex systems. Further, to the degree we are able to understand how misconceptions occur when solving complicated problems, the better able we are to train decision makers working in a variety of complex domains where reductive biases may occur (e.g., command and control, see Houghton, Leedom, & Miles, 2002).

Teams and Organizations in Expertise Studies

Studies of expertise also show how exceptional performers utilize environmental resources to distribute workload across other individuals in their teams or collaborative groups (Fiore, Salas, Cuevas, & Bowers, 2003; Salas, Cannon-Bowers, Fiore, & Stout, 2001; Hollan, Hutchins, & Kirsh, 2000). This finding falls within the area of organizational psychology, as the study of team cognition (Salas & Fiore, 2004). Organizational psychology has attempted to understand human performance at the inter-individual level in order to make predictions and improve team processes. Substantial progress has been made in delineating the sub-factors of effective teamwork and researchers are viewing team cognition as a binding mechanism that produces coordinated behavior within experienced teams. Team cognition encompasses an awareness that binds the actions of the expert team as well as the communication (both implicit and explicit) to scaffold coordinated behaviors (Fiore & Salas, 2004). Thus, a team of experts is not necessarily an expert team (Salas, et al., 1997) and team researchers have argued that expert teams maintain high levels of performance via the development and use of shared mental models for their operational environments (e.g., Cannon-Bowers, Salas & Converse, 1993; Klimoski & Mohammed, 1994; Rouse, Cannon-Bowers, & Salas, 1992).

High performing teams are able to coordinate their actions because they possess commonly held knowledge structures with respect to teammate roles (i.e., knowledge pertaining to their individual responsibilities and required actions). They posses a shared understanding of their team task to a level that allows them to integrate actions and they have a common understanding of the potential situations they may encounter. These shared models are the explanatory mechanism behind constructs such as implicit coordination (Entin & Serfaty, 1999) and situation expectations (Cannon-Bowers et al., 1993).

Thus, shared mental models facilitate expert team performance by facilitating accurate expectations of team members (e.g., Fiore, Salas, & Cannon-Bowers, 2001). Furthermore, in expert teams, awareness can be driven by shared situation assessment processes whereby the shared models drive a common explanation of the meaning of task cues with a concomitant assessment of an operational situation (Salas et al., 2001). As such, team cognition encompasses perceptual processes driving pattern recognition of shared cues as well as conceptual processes whereby shared knowledge bases support the development mental models within dynamic environments.

Additionally, at the level of the organization, researchers are using more complex methods to understand how collections of individuals interact synchronously and asynchronously to produce coordinated behaviors. From the organizational sciences an important method is the use of computational modeling, informed by observation, experimentation, or theory (e.g., Zhu, Prietula, & Hsu, 1997). The use of computer models as a form of computational organization theory is an approach that has a relatively long, but shallow, history in organizational science. The task is not to simply model discrete elements of an organization but to craft models that represent and engage legitimate elements of organizational theory (e.g., Prietula, Carley, & Gasser, 1998; Prietula & Watson, 2000). These can range from "bottom-up" approaches modeling interacting individuals in an organization, to agent based models of varying cognitive complexity modeling groups or micro-societies, to "top down" economic formulations of institutions and markets.

Using such methods, research suggests that organizations can be viewed as a collection of deliberating agents that are cognitively restricted and motivated, task-oriented, and socially-situated. Since the study of domain experts reveals much about the task environment, the study of individuals in organizational settings reveals much about the organizational environment. Furthermore, organizational theorists have proposed the *Induced Simplicity Hypothesis* (Prietula, 2002). This states that:

> For many social and organizational settings, much of the available set of decisions (say, the problem space for the task) is relatively restricted and this simplicity is induced by a confluence of the task, the situation, and the individual. These three factors act as constraints that often severely restrict the behavioral options of the individual, such that models of individuals behaving in those contexts can be sufficiently representative to account for parameters underlying most of the variance" in explaining – or modeling – that situated behavior (pp. 7-8).

Consequently, surprisingly "simple" models of individuals can be incorporated for organizational computational modeling to help us understand and predict coordinated behavior on larger scales.

Research advances are also being made concerning the characteristics and dynamics of *expert organizations*. Recent developments in computational modeling have allowed for interesting research on the effects of organizational structure on performance (Prietula et al., 1998). Only within the past few years have researchers begun to carry this work over to the study of "expert organizations" and organizations with non-traditional (i.e., non-hierarchical) structures. For example studies of expert organizations such as NASA can help understand how the context influences the suitability of knowledge management processes (Becerra-Fernandez & Sabherwal, 2001). Research along these lines has brought about the development of expertise locator systems (Becerra-Fernandez, 2000) and may support human performance at the organizational level. Such findings may facilitate linkages between individual and team and inter-team cognition to help our understanding of how group interaction alters cognitive processes at multiple levels.

CONCLUSIONS

As this summary illustrates, the field of Expertise Studies is making strides in the understanding of learning and performance. For such gains to continue, the field must embrace the utility of diverse methods for understanding. One of FASE's core values is that fundamental advances in the science of learning can be made by leveraging both the findings and the methods used in the study of expertise. In particular, the science of Expertise Studies has effectively utilized both laboratory and field studies to examine expertise development and performance (Ericsson & Smith, 1991; Feltovich, Ford, & Hoffman, 1997; Hoffman, 1992). Laboratory studies rely on tasks that can repeatedly reproduce the superior performance of experts under standardized conditions. These require controlled conditions that must also be representative of the contexts in which experts usually perform and their superior performance is consistently demonstrated (e.g., Ericsson & Lehman, 1996). Nonetheless, researchers have also argued that:

> There is no sense in which we can study cognition meaningfully divorced from the task contexts in which it finds itself in the world... the experiment is an essential tool, but it must answer questions raised by nature, and its answers must be tested against nature (Landauer, 1987, pp. 19–20).

By effectively utilizing these methods both theoretical and practical gains have emerged in psychology in general (see Hoffman & Deffenbacher, 1993) and in the understanding of learning and performance at the level of expert.

In sum, this brief review shows how research in expertise has contributed to our understanding of learning and performance across the proficiency continuum. Expertise research has already had a significant impact on domains as diverse as military operations and sports psychology (e.g., Ericsson & Charness, 1994; Ericsson & Lehmann, 1996; Hoffman, 1992; Salas & Klein, 2001). Furthermore, the knowledge of how expert teachers, coaches and mentors support the development of performance is beginning to be adopted to improve training and performance in a variety of domains (e.g., surgeons, meteorologists, managers, sports, see, for example, Starkes & Ericsson, 2003; Hoffman & Markman, 2001). The recognition of the importance of expertise to society at large is among the most significant developments from the last two decades of research. Nonetheless, following this first generation of Expertise Studies is recognition of just how open and broad the horizons are, and how great is the potential for the advancement of scientific knowledge about expertise to improve learning and performance from the perceptual to the organizational.

REFERENCES

Becerra-Fernandez, I. (2000). The role of artificial intelligence technologies in the implementation of people-finder knowledge management systems. *Knowledge Based Systems*, 13(5) 315-320.

Becerra-Fernandez, I., & Sabherwal, R. (2001). Organizational Knowledge Management Processes: A Contingency Perspective. *Journal of Management Information Systems,* 18(1), 23-55.

Cannon-Bowers, J. A., Salas, E., & Converse, S. A. (1993). Shared mental models in expert team decision making. In N. J. Castellan, Jr. (Ed.), *Current issues in individual and group decision making* (pp. 221-246). Hillsdale, NJ: LEA.

Charness, N. (1991). Expertise in chess: the balance between knowledge and search, in Ericsson, K. A. and J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 39-63). Cambridge, United Kingdom, Cambridge University Press.

Chase, W.G. & Simon, A.S. (1973). Perception in chess. *Cognitive Psychology, 4,* 55-81

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Chi, M. T. H., Glaser, R. & M. J. Farr (Eds.). (1988). *The nature of expertise.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

de Groot, A.D. (1965). *Thought and choice in chess.* The Hauge. Mouton.

Eccles, D. W., Walsh, S. E., & Ingledew, D. K. (2002a). A grounded theory of expert cognition in orienteering. *Journal of Sport and Exercise Psychology, 24,* 68-88.

Eccles, D.W., Walsh, S.E., & Ingledew, D.K. (2002b). The use of heuristics during route planning by expert and novice orienteers. *Journal of Sports Sciences, 20,* 327-337.

Entin, E. E. & Serfaty, D. (1999). Adaptive team coordination. *Human Factors, 41,* 312-325.

Ericsson, K. A. (Ed.) (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Ed.) *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports, and Games* (1 – 50). Mahwah, NJ: Erlbaum.

Ericsson, K. A. & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49,* 725-747.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102,* 211-245.

Ericsson, K. A., Krampe, R. & Tesch-Römer, C. (1993), The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363-406.

Ericsson, K. A. & Lehmann, A.C. (1996). Expert and exceptional performance: Evidence on maximal adaptations on task constraints. *Annual Review of Psychology, 47,* 273-305.

Ericsson, K. A., & Smith, J. (Eds). (1991). *Toward a general theory of expertise: Prospects and limits.* Cambridge, England: Cambridge University Press.

Feltovich, P. J., Ford, K. M., & Hoffman, R. R. (Eds.) (1997). *Expertise in context.* Menlo Park, CA: AAAI/MIT Press.Feltovich,P., Spiro, R., & Coulson, R. (1997). Issues of expert flexibility in contexts characterized by complexity and change. In P. J. Feltovich, K. M. Ford, & R. R. Hoffman (Eds.), *Expertise in context.* Menlo Park, CA: AAAI/MIT Press.

Fiore, S. M. & Salas, E. (2004). Why we need *team* cognition. To appear in E. Salas & S.M. Fiore (Editors). *Team Cognition: Understanding the factors that drive process and performance* (pp. 235-248). Washington, DC: American Psychological Association.

Fiore, S. M., Salas, E., & Cannon-Bowers, J. A. (2001). Group dynamics and shared mental model development. In M. London (Ed.), *How people evaluate others in organizations: Person perception and interpersonal judgment in Industrial/Organizational psychology* (pp. 309 – 336). Mahwah, NJ: Lawrence Erlbaum.

Fiore, S. M., Salas, E., Cuevas, H. M. & Bowers, C. A. (2003). Distributed coordination space: Toward a theory of distributed team process and performance. *Theoretical Issues in Ergonomic Science, 4, 3-4,* 340-363.

Hancock, P.A., & Warm, J.S. (1989). A dynamic model of stress and sustained attention. *Human Factors, 31,* 519-537.

Hancock, P. A. & Weaver, J. L. (in press). On distortion under time stress. *Theoretical Issues in Ergonomic Science.*

Hoffman, R. R. (1998). How can expertise be defined?: Implications of research from cognitive psychology. In & R. Williams, W. Faulkner, & J. Fleck (Eds.), *Exploring expertise* (pp. 81-100). New York: Macmillan.

Hoffman, R. R. (Ed.). (1992). *The Psychology of Expertise: Cognitive Research and Empirical AI.* New York: Springer-Verlag.

Hoffman, R. R., & Deffenbacher, K. A. (1993). An analysis of the relations of basic and applied science. *Ecological Psychology, 5,* 315-352.

Hoffman, R. R. & Markman, A. B. (2001). *Interpreting Remote Sensing Imagery: Human factors.* New York: Lewis Publishers & CRC Press.

Hoffman, R. R., Shadbolt, N., Burton, A. M., & Klein, G. A. (1995). Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes, 62,* 129-158.

Hollan, J., Hutchins, E. & Kirsh, D. (2000). *Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research.* ACM Transactions on Computer-Human Interaction (TOCHI), ACM Press.

Houghton, P., Leedom, D. K., Miles, J. (2002). Command information infrastructure and sense-making. Paper presented at the *7th International Command and Control Research and Technology Symposium,* September, Quebec City, Canada.

Janelle, C.M., & Hillman, C.H. (2003). Expert performance in sport: Current perspectives and critical issues. In J. Starkes & K.A. Ericsson (Eds.) *Recent advances in research on sport expertise* (pp. 19-48). Champaign, IL: Human Kinetics.

Janelle, C.M., Hillman, C.H., Apparies, R., Murray, N.P., Meili, L., Fallon, E.A., & Hatfield, B.D. (2000). Expertise differences in cortical activation and gaze behavior during rifle shooting. *Journal of Sport & Exercise Psychology, 22,* 167-182.

Janelle, C.M., Singer, R.N., & Williams A.M. (1999). External distraction and attentional narrowing: Visual search evidence. *Journal of Sport & Exercise Psychology, 21,* 70-91.

Kamata, A., Tenenbaum, G., & Hanin Y., . (2002) Individual Zone of Optimal Functioning (IZOF): A probabilistic estimation. *Journal of Sport & Exercise Psychology, 24(2),* 189-208.

Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or Metaphor? *Journal of Management, 20,* 403-437.

Landauer, T. K. (1987). Relations between cognitive psychology and computer system design. In J. M. Carroll (Ed.), *Interfacing thought: Cognitive aspects of human-computer interaction* (1-25). Cambridge, MA: MIT Press.

Marshall, C.H. Jr., Pielke R. A., & Steyeart, L. T. (2003). Crop freezes and land-use change in Florida. *Nature, 426,* 29-30.

Matthews, G. (2001). Levels of transaction: A cognitive science framework for operator stress. In P.A. Hancock and P.A. Desmond (Eds.), *Stress, workload, and fatigue.* (pp. 5-33). Mahwah, NJ: Erlbaum.

Murray, N.M., & Janelle, C.M. (2003). Anxiety and performance: A visual search examination of the processing efficiency theory. *Journal of Sport & Exercise Psychology, 25,* 171-187.

Prietula, M. J. (2002). *TrustMe 1.0: A social simulation of trust, advice and gossip.* Goizueta Business School, Center for Business Simulations, Emory University.

Prietula, M., Carley, K. & Gasser, L. (Editors.) (1998). *Simulating Organizations: Computational Models of Institutions and Groups*, AAAI/MIT Press, 1998.

Prietula, M. & Watson, H. (2000). A behavioral theory of the firm: Organizational and economic perspectives, *Organizational Science.* 11(5), 565-585.

Radlo, S.J., Janelle, C.M., Barba, D.A., & Frehlich, S.G. (2001). Perceptual decision-making for baseball pitch recognition: Using P300 latency and amplitude to index attentional processing. *Research Quarterly for Exercise and Sport, 72,* 22-31.

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science, 12,* 48-55.

Rouse, W. B., Cannon-Bowers, J. A., & Salas, E. (1992). The role of mental models in team performance in complex systems. *IEEE Transactions on Systems, Man, and Cybernetics, 22,* 1296-1308.

Salas, E., Cannon-Bowers, J. A., Fiore, S. M., & Stout, R. J. (2001). Cue-recognition training to enhance team situation awareness. In M. McNeese, E. Salas, & M. Endsley, (Eds.), *New trends in collaborative activities: Understanding system dynamics in complex environments* (pp. 169-190). Santa Monica, CA: Human Factors and Ergonomics Society.

Salas, E., Cannon-Bowers, J. A., & Johnston, J. H. (1997). How can you turn a team of experts into an expert team?: Emerging training strategies. In C. Zsambok & G. Klein (Eds.), *Naturalistic decision making* (pp. 359-370). Hillsdale, NJ: LEA.

Salas, E., & Fiore, S. M. (Eds.). (2004) *Team Cognition: Process and Performance at the Inter- and Intra-individual Level.* Washington, DC: American Psychological Association.

Salas, E. & Klein, G. (Eds.) (2001). *Linking expertise and naturalistic decision making.* Hillsdale, NJ: Lawrence Earlbaum Associates.

Simon, H.A. & Chase, W. (1973). Skill in chess. *American Scientist, 61,* 394-403.

Spiro, R. J., Coulson, R. L., Feltovich, P. J., & Anderson, D. K. (1994). Cognitive flexibility theory: Complex knowledge acquisition in ill-structured domains. In R. B. Ruddell, M. R. Ruddell, & H. Singer (Eds), *Theoretical models and processes of reading*, 4th ed. (pp. 602-615). Newark, DE: IRA.

Starkes, J, & Ericsson, A. (Eds.). (2003). *Expert Performance in sports: Advances in research on sport expertise.* Champaign, IL: Human Kinetics.

Tenenbaum, G (in press). Decision making in sport. In C. Spielberger (Ed.), *Encyclopaedia of applied psychology.* San Diego, CA: Academic Press.

Tenenbaum, G. (2003). Expert athletes: An integrated approach to decision-making. In J. Starkes & K. A. Ericsson (Eds.), *Expert Performance in sports: Advances in research on sport expertise* (pp. 192-218). Champaign, IL: Human Kinetics.

Tenenbaum, G. & Bar-Eli, M. (1993). Decision making in sport. In R.N. Singer, M. Murphey, and L.K. Tennant (Eds.), *Handbook on Research in Sport Psychology* (pp. 171-192). NY: Macmillan.

Tenenbaum, G. & Bar-Eli, M. (1995). Personality and intellectual capabilities in sport psychology. In D. Saklufske and M. Zeidner (Eds.), *International Handbook on Personality and Intelligence* (pp. 687-710). NY: Plenum.

Tenenbaum, G., & Elran, E. (2003). Congruence Between Actual and Retrospective Reports of Emotions for Pre- and Postcompetition States. *Journal of Sport & Exercise Psychology, 25,* 323-340.

Tenenbaum, G., Levy-Kolker, N. Sade, S., Lieberman, D. G., & Lidor, R. (1996). Anticipation and confidence of decisions related to skilled performance. *International Journal of Sport Psychology, 27,* 293-307.

Tenenbaum, G., & Lidor, R. (in press). Research on decision-making and the use of cognitive strategies in sport settings. In D. Hackfort, J. Duda, & R. Lidor (Eds.), *Handbook of research on applied sport psychology.* Morgantown, WV: Fitness Information Technology.

Ward, P. & Williams, A.M. (2003). Perceptual and cognitive skill development in soccer: The multidimensional nature of expert performance. *Journal of Sport and Exercise Psychology, 25 (1),* 93-111

Williams, A. M, Davids, K, Williams, J. G. (1999). *Visual perception and action in sport.* London: E & FN Spon.

Williams, A. M., & Hodges, N. J. (2003). *Skill acquisition in sport: Research, theory and practice.* London: Routledge.

Zhu, D., Prietula, M. & Hsu, W. (1997). When processes learn: Steps toward crafting an intelligent organization. *Information Systems Research*, 8(3), 302-317.

EFFECTS OF COGNITIVE FUNCTIONING ON STRATEGIC SYSTEM ENGINEERING

Michael Goings, Stephen A. St. Cyr, Steve Hall and Shawn Doherty
Embry-Riddle Aeronautical University

ABSTRACT

Mental workload is an important construct in psychology. Using various methods, researchers have investigated ways to reduce the amount of workload imposed on system operators. Reducing workload through system design might be facilitated by identifying required cognitive resources and designing the system so that tasking does not impose resource conflict which may cause a decrement in performance. Wickens' multiple-resource theory has expanded on the three stages of processing (encoding, central-processing, and responding) to include cognitive resources, such as visual/spatial encoding, spatial/abstract processing, and manual discrete and non-discrete responding resources which are identified in this model. This study represents a first step towards building a research paradigm in which the amount of resource conflict (resulting in performance decrements) is estimated by taxing multiple resources simultaneously.

Keywords: Cognition; Mental Workload; Cognitive Channels; Cognitive Constructs

INTRODUCTION

Mental workload assessment is an important domain in psychology. Workload can be defined by the cost on the operator when a task impresses various variables such as time restraints, number of tasks, and complexity of a task or tasks (Advisory Group for Aerospace Research & Development [AGARD], 1998). This research project is concerned with the effects of simultaneous tasking on mental capacity.

Many methods have been used to investigate and reduce workload. However, when a designer is faced with the decision of what tasks to impose on an operator, some of these methods to assess mental workload may be long and difficult. The present study proposes to examine the impact of simultaneous tasking by the factorial combining of tasks that require specific cognitive resources. The goal is to identify the various combinations of cognitive tasking that result in minimal performance degradation. Wickens' multiple-resource theory provides the groundwork for this study.

The information-processing model describes the three-step path in which information flows. In the input stage, the human must sense, select, and perceive the stimuli. The information processing stage performs the job of encoding, committing to memory, recalling, making decisions and making judgments. Finally, using the processed information the human can react to the stimuli using either a verbal response or execute a physical response depending on what is required (Chapanis, 1996). Wickens' (1992) multiple-resource theory builds on the information-processing model by decomposing the path information takes into a multidimensional model. It is comprised of the visual and auditory modalities, the three stages of processing which are encoding, central-processing, and responding, and the processing codes. Each part of this multi-dimensional model can be considered a distinct cognitive resource, and it has a particular purpose.

The visual and auditory modalities are the channels used for input. Wickens (1992) described a channel as the way information comes into and flows through the stages of processing. Information flows through the visual or auditory channel to the central-processing stage where the information is digested. Lastly, there is the response resource which is dependent upon the output required from the operator. The stages of processing can work with two different perceptual processing codes: spatial and verbal (Wickens, 1992). These resources are useful in researching workload. However, cognitive constructs, which are used by the central-processing resources, may utilize varying levels of mental effort.

The central-processing resource can be divided into simple but intangible functions called cognitive constructs. The cognitive constructs can be tapped into by performing an array of processes. Hyland, Kay, and Deimler (1994) described several cognitive processes and its corresponding construct. The perceptual construct is utilized in the following processes: auditory, visual perception, and visual scanning. Psychomotor is primarily used in tasks involving mind and body coordination, such as a tracking task. Selective/focused, divided, switched, and sustained processes all play a role in the attention construct. Committing information to memory either long-term or short-term is a function of the memory construct. Some tasks require the person to use their information-processing construct to process visual spatial or verbal sequential information. Lastly, a problem solving/decision-making construct can be used in two distinct ways. When a task requires the human to apply rules and draw conclusions based on the given situation, he/she is involved in a domain independent task. Conversely, if the task requires him/her to assess a situation and make a competent decision, the task is domain dependent. One implication of the multiple resource approach is that the combination of these processes can provide an index into the effect of combined processes. The combination of these processes can be outlined in a cognitive matrix.

With consideration to the cognitive resources used in the PUMA (1993) conflict matrix, the proposed cognitive matrix will examine conflicts in the central-processing resources while holding the input and output modalities of the multiple-resource theory model constant. Simultaneous tasking requiring the utilization of different central-processing resources may or may not adversely affect performance depending on how much capacity is used by each resource and potential unique interaction effects. The cognitive matrix may be become a useful tool in identifying tasks that will compete for resources and result in performance decrements. This research project sets out to start a research paradigm to map the likely performance outcomes of simultaneous central-processing tasking via a cognitive matrix.

This study sets up a baseline of single central-processing tasks in order to investigate decrements in performance when additional tasks are added. It is assumed that a decrease in performance will be an outcome in any simultaneous task condition; however, tasks that are orthogonal to each other should produce little if any drop in performance. In terms of Wicken's model, the tasks utilized in this study are presented via the visual/spatial channel, require spatial/abstract central-processing, and completed with a manual response.

METHOD

Participants and Design

Sixteen undergraduate student from a southeastern university participated in the experiment. Most participants were given the option to participate in the experiment in order to receive extra credit in their experimental psychology courses. A smaller number volunteered to participate in the experiment without course benefit. Participants were between the ages of 21 and 27. Six of the sixteen participants were male. All participants had normal or corrected to normal vision.

Apparatus

The experiment used a Dell Dimension XPS R350 with a Pentium 2 Processor with a 15" (~38cm) Dell monitor to run the Multiple Attribute Test (MAT) battery software developed by Comstock & Arnegard (1992). Participants were seated in an open cubicle with minimal background noise.

Procedure

At the beginning of each session, the researcher read a script explaining each task to the participant. The participant was also shown a paper screenshot of the Multi-Attribute Test Battery (MAT) (Comstock & Arnegard, 1992). After the script was read in its entirety and all the participants' questions had been answered, participants were allowed to practice the three tasks for five minutes.

Following the practice session, each participant was presented with one of task conditions. Each condition lasted ten minutes. The six conditions are as follows: system monitoring (M) alone, tracking (T) alone, and resource management (F) alone, system monitoring and tracking (MT), system monitoring and resource management (MF), and tracking and resource management (TF).

In the monitoring task the participants were required to monitor a series of four dials and make corrections based on the position and movement of the dials. Within these four dials there were tic marks and fluctuating pointers. If the pointers began to deviate from their normal fluctuation, either above it or below it, then the subject would respond by striking a corresponding key.

The tracking task required the participants to monitor a scope and cross hairs system and make adjustments as the scope deviated from the crosshairs. Finally, in the fuel resource management task the participant is asked to monitor a fuel tank system and to keep the fuel levels constant. This was done by allocating fuel from a source to specific tanks by using a system of pumps and other tanks.

Variables

The input channel remained constant (i.e. visual) and the independent variable was the type cognitive construct from spatial abstract processing being used. The dependent variable was performance, measured via hit-miss ratio for the monitoring task, root-mean-square error (RMSE) for the tracking task, and tank deviation for the fuel management. The specific tasks presented to the participants (monitoring task, tracking task, and fuel management task) require attention, psychomotor, and domain independent problem solving cognitive constructs, respectively. Though the tasks possibly represented other types of spatial abstract processing, the dominant type was the construct chosen to represent a particular task.

RESULTS

Attention via System Monitoring (M)

Performance in the monitoring-only condition was compared to performance in the monitoring-tracking and monitoring-resource management conditions to determine if there were statistically significant drops in system monitoring performance. In the monitoring-only condition the mean hit-miss ratio measured 96.2% (SD = .074). The monitoring-tracking condition resulted in a mean monitoring hit-miss ratio of 87.9% (SD = .020) and the monitoring-resource management condition resulted in a mean monitoring hit-miss ratio of 85.8% (SD = .137). The results for a paired samples t-test found the performance drop in system monitoring when tracking was added to be non-significant, , $t(15) = 1.502$, *ns*. However, there was a statistically significant drop in monitoring performance when the resource management task was added, $t(15) = 2.676, p < .05$.

Psychomotor via Tracking (T)

Performance in the tracking-only condition was compared to performance in the tracking-monitoring and tracking-resource management conditions to determine if there were statistically significant drops in tracking performance. In the tracking-only condition, mean tracking performance was 46.50 RMSe units (SD = 14.34). When the monitoring task was added, the tracking performance group mean 81.50 RMSe units (SD = 31.91) and when the resource management task was added, mean tracking performance was 86.63 RMSe units (SD = 37.77). Increases in RMSe scores indicate lower levels of performance. The results from a paired samples t-test revealed a statistically significant drop in tracking performance when the monitoring task was added, $t(15) = -5.725, p < .05$, and when the resource management task was added, $t(15) = -4.779, p < .05$.

Domain Independent Problem Solving via Resource Management (F)

Performance in the resource management-only condition was compared to performance in the resource management-monitoring and resource management-tracking conditions to determine if there were statistically significant drops in resource management performance. Group mean performance in the resource management-only condition was 63.17 gallons (SD = 83.26). When the monitoring task was added, mean performance was 66.35 gallons (SD = 55.93) and when the tracking task was added, group mean performance was 60.86 gallons (SD = 37.18). The results from a paired samples t-test failed to find a statistically significant drop in resource management performance when the monitoring task was added, $t(15) = -0.140$, *ns*, or when the tracking task was added, $t(15) = 0.124$, *ns*.

Matrix

Analyses of the scores obtained from each participant were broken down to mean scores and then performance ratios were calculated. These ratios were used to compute the estimated percent decrement in performance ((1 – performance ratio) * 100). Statistically non-significant drops are represented by *ns*.

Table1. Percent-drop in performance for each primary task when a secondary task was added.

Primary Task	Secondary Task (M)	(T)	(F)
System Monitoring (M)	-	.*ns*	11%
Tracking (T)	42.9%	-	46.3%
Resource Mgmt. (F)	.*ns*	*ns*	-

*Note that when one task is matched up with the same task, only one task was performed. The same task was not repeated with itself.

DISCUSSION

This experiment's main goal was to determine and recognize any interference in areas of spatial abstract processing due to simultaneous task loading. While there were no immediate expectations about the outcomes, the original idea was to expose statistically significant drops in primary task performance, if any, and then to apply real world theory to explain those drops. The goal of this research was not to determine *if* there were any drops in performance, but rather *where* those drops might lie.

The results indicate which tasks, and hence cognitive processes, suffer when additional tasks are added. First, when it comes to spatial abstract processing, attention tasks (M) are harder to control when attempting them with another task involving an operator's domain independent problem solving (F), but not when attempting a psychomotor task (T). Second, it also means that when an operator is performing a task involving psychomotor abilities, such as the tracking task (T), his or her performance declines when performed with a domain independent problem solving task (F) or a monitoring task (M), thus indicating that psychomotor activities may require cognitive abilities that are also required by other tasks. Finally, domain independent problem-solving resources may take precedence over other tasks as evidenced by no significant decrease in performance during the introduction of a psychomotor or attention task. Of course, one should be very careful when drawing conclusions on the basis of statistically non-significant findings.

Theoretically, two conclusions could be drawn from these results: (1) any domain independent problem-solving task may be combined with any psychomotor task and attention task without a significant decline in performance, and/or (2) there maybe a tendency for participants to give more attention to a problem-solving task. This theory is largely based on the assumption that each task best represents its dominant function.

Knowles (1963) stated that a system designer should be able to answer questions (1) about the ease of operation, (2) attention required, (3) learning involved, and (4) ability to perform another task. The cognitive can aid system designers by addressing questions about the possibility of two tasks interfering with each other, thereby allowing the designer to predict and avoid unintended decrements in performance. However, in the early stages of building the matrix, it may lack in ecological validity depending on the nature of the tasks that the researcher uses to build it.

Each task in this study represents real tasks that an operator may have to perform while flying an aircraft. However, it may not relate to a different situation that calls upon the same cognitive constructs. To account for this hypothetical situation several dissimilar tasks that use the same cognitive capacities should be explored to strengthen the validity of the matrix. When building the cognitive matrix, researchers should factor in their study the issues set forth by Knowles (1963). Currently, the cognitive matrix can only determine if two tasks being performed simultaneously will cause a decrement in performance. Future research should be directed towards testing the cognitive matrix in real world situations and creating a metric that would measure the level of difficulty.

Only a small portion of the matrix is represented by this study; Wicken's cognitive resource theory includes 11 cognitive constructs. In addition to the results exposing difference in processing constructs, this experiment infers that there is value in completing the matrix. Though it would be a meticulously long task, the portion of the matrix created presently reveals usefulness and importance in completion.

CONCLUSION

This research was only the preliminary step towards creating a cognitive matrix. Future research should be directed towards completing this matrix. The matrix could become a human performance library of workload. It could prove beneficial by simplifying a designer's job in abating that amount of workload impressed upon the operator.

REFERENCES

Advisory Group for Aerospace Research & Development. (1998, December). A Designers Guide to Human Performance Modeling (AGARD AR-356). Quebec, CA: Canada Communication Group Inc.

Chapanis, A. (1996). Human Factors in Systems Engineering. New York: John Wiley & Sons, Inc.

Comstock, R. J., & Arnegard, R. J. (1992) The multi-attribute task battery for human operator workload and strategic behavior research. Norfolk, VA: Old Dominion University. (NASA-TM-104174)

Hyland, D. T., Kay, E. J., Deimler, J. D. (1994). Age 60 Study, Part IV: Experimental Evaluation of Pilot Performance. Oklahoma City, OK, Civil Aeromedical Institute, FAA: (NTIS No. FAA-AM-94-23)

Knowles, W. B. (1963) Standardization of loading tasks, which would obtain more meaningful measures of operator workload in simulation studies. Human Factors, 5, 155-161

Roke Manor Research and British Aerospace Sowerby Research Center. (1993). Perfomance & Usability Modeling and Air Traffic Management. [User Manual].

Wickens, C. D. (1992). Engineering Psychology and Human Performance (2nd ed.). New York: HarperCollins Publishers Inc.

Human Factors in Aviation:
Determining the cause of accidents from a system perspective

Dennis A. Vincenzi
Embry-Riddle Aeronautical University

Todd V. Denning
Northrop Grumman Mission Systems

Abstract

Over the past forty years aircraft accidents continue to occur in spite of efforts by human factors professionals to investigate, determine, and publicize problems encountered by pilots. One problem associated with this phenomenon is that investigating agencies only rarely consider the investigation from a holistic, systemic point of view. Although the literature suggests that the true cause of many accidents, especially those associated with pilot error, may be systemic in nature, many times accident investigators are content with placing the blame solely on an individual (the pilot) or a group of individuals (the flight crew). This practice is detrimental to the industry and misleading, often resulting in superficial conclusions.

Introduction

Over the past forty years aircraft accidents continue to occur in spite of efforts by Human Factors professionals to investigate, determine, and publicize problems encountered by pilots. In the 60's, the predominant cause of accidents appeared to be pilot error mainly attributed to a lack of basic flying skills. In the 70's, the predominant cause of accidents appeared to be pilot error mainly attributed to a lack of technical proficiency. In the 80's, the emphasis shifted from individual pilot error to crew resource management (CRM) problems, and in the 90's, the predominant cause appears to be shifting to a failure of organization and error management among crews (Pariès & Amalberti, 2000). Despite efforts from human factors professionals, accidents due to crew and pilot error still proliferate. Errors of mode confusion based in the flight management system (FMS) and CRM are the main focus of attention.

The development of glass cockpit aircraft in the early 80's and rapid integration of those aircraft in the 90's has led to increasing worries among Human Factors professionals that the cockpit may have become or will become too automated. Human-computer interaction and CRM has come to be the focus of professionals in this field. Automation has changed the nature of the role of the pilot in two major ways. The development and application of highly reliable automated systems in today's world has changed the role of the human from an active system operator to one of a passive system monitor, a role for which humans are not well suited (Parasuraman, 1997). Monitoring of highly automated systems is a major concern for human performance efficiency and system safety in a wide variety of human-machine systems (Parasuraman, 1987; Vincenzi & Mouloua, 1998). Human monitoring of automated systems for malfunctions in the real world can often be poor as a result of low frequency of occurrences of automation failures or automation surprises when dealing with reliable automated systems. Instead of reducing stress and workload in the cockpit, these two quantities may significantly increase resulting in poorer performance and increased possibility of human error. Feedback associated with highly automated systems is often limited. In addition to "flying" the aircraft, the pilot now must understand the actions of the automation. Pilots often find themselves wondering about the automation routines being executed. This can lead to a significant use of available resources as well as loss of situation awareness. Second, rather than flying the aircraft directly, pilots must interact with the FMS and fly indirectly, giving direction to the automation and having it enact the changes (Sarter & Woods, 1992). Human-computer interaction is becoming the focus among professionals as more and more pilots are reporting automation surprises. Since the 80's and the widespread proliferation of flight control automation, the overall accident rate has decreased, but not without problems. There is still a trend of accidents and incidents that may be due to human-computer interaction (Sherman, Helmreich, & Merritt, 1997). As usual with new advances in technology, the new designs have reduced the occurrence and severity of some errors commonly made by crews, but have simultaneously opened the door to new types of errors (Pariès & Amalberti, 2000).

In research conducted by Sarter and Woods, a survey was distributed among pilots that asked them to describe in detail any problems they had experience with the FMS, and specifically if they had ever been surprised by the technology, they were asked to describe the problems they encountered. The results from 135 pilots were broken down into nine major categories including, VNAV modes, data entry, uncommanded mode transitions, and surprising flight director (FD) commands. The report showed that pilots can make the FMS work, however, it is usually by sticking to common operations the pilot uses routinely. In the event of automation surprise, many pilots are caught off-guard, unable to explain the automation, and unable to explain the logic for the FMS action (Sarter & Woods, 1992).

Weiner's concept of clumsy automation can also be useful in explaining the deficiencies in the cockpit. Sarter and Woods (1992) found that individual pilots tend to stick to the automation they know and trust. This can exacerbate bottlenecks in tense, high-pressure emergency situations. Without fully knowing the strategies and automation preferred by different colleagues, situational awareness, full knowledge of pilot-cockpit and pilot-pilot coordination can decrease dramatically. Situation Awareness has recently been accepted as an essential prerequisite for the safe operation of any complex system, including aircraft (Sarter & Woods, 1991).

Mode Awareness

Besides situation awareness, one category from Sarter and Woods (1992) has been the focus of predominately more attention than the others. Mode awareness in flight management systems has plagued pilots and manufacturers alike (Sarter & Woods, 1992, 1994; Hughes & Dornheim, 1995; Hughes, 1995; Phillips, 1995; Sherman, Helmreich, & Ashleigh, 1997; Phillips, 1999; Dornheim, 2000; Dismukes & Tullo, 2000). Results of a study by Sarter and Woods (1994) showed that more than 70% of the pilots surveyed had difficulties 1) aborting a takeoff at 40 knots with autothrottles on, 2) anticipating ADI mode indications in a takeoff roll, 3) anticipating when go-around mode becomes armed throughout landing, 4) disengaging Approach mode after localizer and glide slope capture, 5) explaining speed management, and 6) defining end-of-descent point for VNAV path versus VNAV speed descent. 65% of all pilots in the study could not tell the experimenter how to completely abort a takeoff (Sarter & Woods, 1994). The results showed that the majority of the errors were errors of mode awareness and gaps in the pilots' mental models of the actual function of the automation in the aircraft. They found that for most pilots, it was nearly impossible to navigate the automation when an aborted takeoff had occurred (Sarter & Woods, 1994). These problems indicate a need to develop better interfaces to give the pilots better options and increased awareness during these time-critical situations. In a simple context, mode awareness refers to the ability to have the adequate assessment of the currently active mode (Sarter & Woods, 1994). There is agreement among most professionals that awareness in the cockpit is much more than the basic definition. Pilots need to have a firm grasp on the functions of the FMS; they need to be able to predict what it will do, especially in high-stress situations. It has become clear that this is not the case.

Incidents of mode confusion abound. All aircraft, including those made by Airbus, Boeing and Douglas, suffer from the same plight. Increased automation has confused the pilot. Several crashes, including the Airbus A300-600 at Nagoya and an A310-300 at Orly Airport in France, have revealed pilot interaction with the automation to be a significant factor (Hughes & Dornheim, 1995). How do we get around this factor? The truth is, we can't. While Airbus, Boeing, and Douglas all have different ideas about automation and the role of automation in the cockpit, pilots and crew have to be able to take control of the automation, not the other way around. When crews are not given feedback about a mode transition and are caught off guard, tragedy has been known to happen. Basic communication between pilot and crew are essential, but is being cut off by the automation.

Crew resource management

Crew resource management has seen renewed interest in the 80's and 90's as automation surprises are forcing the pilot and crew to work together. Today, human error is reported to be the most common cause of Naval aviation mishaps (Weigmann & Shappell, 1999). The results of an analysis into the causal factors of Class A Naval aircraft mishaps between 1986 and 1990 showed that aircrew error was the most predominant factor among all human causal factors (59%). Within aircrew error, the most common form of error was lack of communication and coordination between aircrew (Weigmann & Shappell, 1999).

A survey of the next seven years was conducted to see if anything had changed or had been learned from the previous survey. It was found was that 75% of the mishaps were attributable, in at least a small way, to human error. With 56% of the aircrew errors being attributable in part to CRM, it is evident that there are serious human factors implications. Weigmann and Shappell (1999) reported that the most deleterious effects of CRM were during high stress situations. While trying to figure out what one problem is, a crew may miss another problem entirely.

An analysis of 107 reports where crew error was cited claimed that half of those errors were from a crew becoming pre-occupied with one task, and missing another (Dornheim, 2000). Among these distractions, 90% fit into four categories:

1. Communications among the crew or while on the radio was the biggest cause of distraction (68 of 107 incidents).
2. Head down work including programming and scanning the FMS or reviewing approach charts (22 incidents).
3. Response to abnormal situations (19 incidents).
4. Visually searching for traffic (11 incidents).

This first category is the one that is of most concern. Talking to crew members, answering and asking questions and thinking of answers takes valuable time, time that may be used to catch an error somewhere else (Dornheim, 2000). Add the effects of mode confusion somewhere and the culminating effect is disaster.

While the overall accident and incident rate has been reduced compared to previous generation aircraft, new trends in errors are emerging. While technical proficiency was the focus of errors in the 70's, CRM errors were the dominant 80's research, and CRM and error confusion have dominated the research in glass cockpit generation aircraft in the previous decade. It seems that despite best efforts from human factors professionals, accidents continue. Professionals still have not found a way to design a system that perfectly complements the human being. The reality is that the human is not as predictable as the machine, which constitutes the main difference and the challenge for aviation and human professionals throughout the world. The patterns of human error within performance still exist, however, the emphasis, as reported, has shifted to the interface and automation surprises.

The Cause of Accidents

Accident summaries were examined over the past 20 years from 1981 to 2000 for accidents that occurred involving Part 121 and Part 135 operations in the United States. Part 121 applies to air carriers such as major airlines and cargo haulers that fly large transport aircraft. Part 135 applies to commercial air carriers commonly referred to as commuter airlines and air taxis. Some major categories of causes of accidents include pilot error, mechanical failure, and weather. Overwhelmingly, with very few exceptions over the past 20 years, the major cause of aircraft accidents in the United States involving Part 121 and part 135 operations, as determined by the National Transportation Safety Board (NTSB), has been pilot error or some form of pilot related error (Figure 1).

Pilot error, however, is not easily and clearly defined. In fact, the definition of pilot error seems to vary over the years and seems to include, but is not limited to, concepts such as loss of situation awareness, poor CRM, and poor decision making. These concepts, although discussed as individual concepts, are all involved and integrated in the greater overall concept of cognitive information processing. Doesn't loss of situation awareness often lead to poor decision making? Doesn't poor CRM often lead to loss of situation awareness? Other deeper, more probing questions have been conceptually asked throughout the years such as "What causes loss of situation awareness?" and "why do highly trained personnel participate in poor decision making?" These topics and other similar topics have been debated and dissected on a conceptual level quite extensively, however, aircraft accidents involving pilot error still proliferate.

Very rarely are accident cause determinations pursued beyond the point where blame can be placed on an individual (the pilot) or a group of individuals (the flight crew). Once the cause of the accident is determined, the investigation must go further to determine why the problem that ultimately caused the accident was not detected and resolved before the accident occurred.

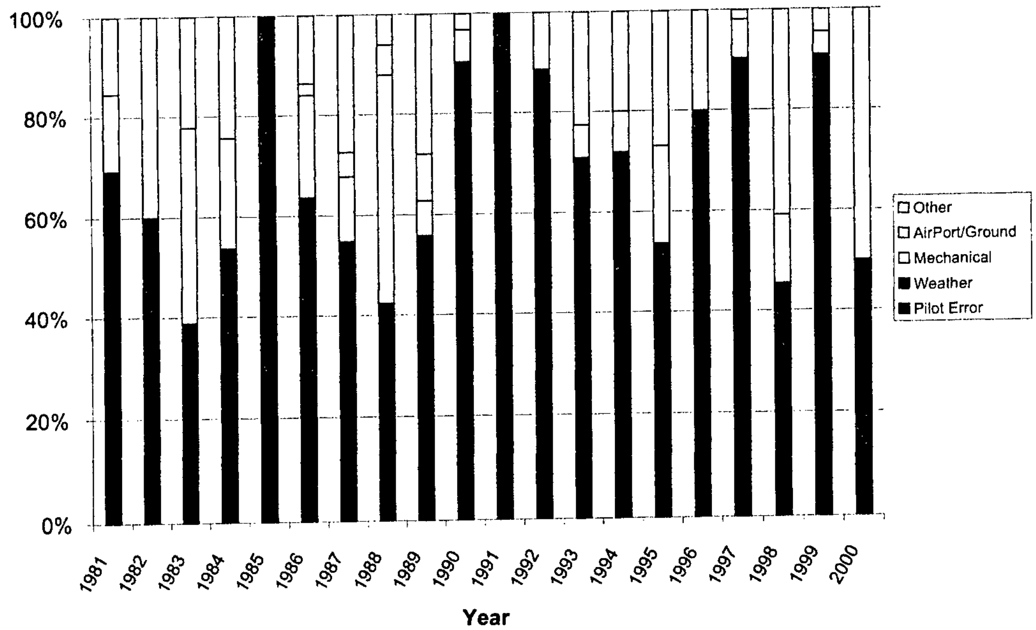


Figure 1. Summary of accident cause percentages by year from 1981 to 2000.

From a systems perspective, determination of the cause of many accidents is not an easy task. The smaller the defined system, the easier it will be to pinpoint the cause. One contributing factor often leading to pilot error is poor decision making. Decision making has long been recognized as a major factor affecting flight safety (Pariès & Amalberti, 2000). Jensen and Benel (1977) found that decision errors contributed to more than one third of all accidents in the United States from 1970 – 1974. They also argued that good decision making skills can be trained. Why do pilots sometimes participate in poor decision making? The answer to this question may branch off into any number of areas that may include but are not limited to broader system aspects such as training, selection, interface design, cultural differences, or organizational considerations. If the root of the problem is determined to be inadequate training, then next step should be why is the training inadequate, in what way is the training inadequate, and what can be done to enhance the training so that poor decision making is not a problem. The same can be said of selection, interface/system design, or cultural/social/economic aspects of the organization. If any of these broader systems aspects, or combination of these aspects are found to be inadequate or deficient in some way, and are determined to be a contributing factor to the problem, blame must be placed accordingly and appropriately, and corrective action must be taken so that the system as a whole can be adequately prepared to deal with the problem.

Interface design may play an important role in decision making. Confusing and cluttered displays may overwhelm an operator, especially in times of high workload and stress, whereas simple displays may not provide adequate information to maintain proper situation awareness and make proper decisions. Highly reliable automated systems are a good example of systems that often provide little feedback as to what is being done and why actions are being taken. In cases such as these, is it still pilot error if a poor decision is made due to lack of information or is the a system design flaw that can be traced back to aspects of the interface design that fails to match and complement the system operator?

Organizational considerations may adversely impact system aspects such as training. Training costs money, and companies do not like to spend money on non-productive, non-revenue generating activities. Training is one such activity. On the surface, training is very expensive, simulators and simulator time is costly, and individual pilots and entire crews must be placed into non-productive, non-revenue generating activities. The natural organizational tendency would be to reduce such activities to the absolute minimum. However, from a system perspective, this may be detrimental.

Conclusion

A major shift in the aviation safety paradigm can be observed in that the focus has moved from reactive to proactive safety, and from individuals to organizations. This paradigm shift is traceable in training and affects the skills and abilities required in a cockpit for more efficient and safer flights (Pariès & Amalberti, 2000). In order to facilitate paradigm shifts of this nature in safety, accident investigation must be pursued from a system perspective and the cause of accidents must be traced back to broader system aspects whenever possible. The human component is still an integral component of the human-machine system. Crews are expected to perceive the environment, to maintain a proper situation awareness, to anticipate the situation and make relevant decision in normal as well as abnormal situations (Pariès & Amalberti, 2000). If accidents occur, the entire system must be scrutinized to determine the cause and the solution to the problem to minimize the possibility of reoccurrence.

References

Dornheim, Michael A. (2000). Crew Distractions Emerge as New Safety Focus. *Aviation Week & Space Technology.* 17 July.

Dismukes, Key, & Tullo, Frank. (2000). Aerospace Forum: Rethinking Crew Error. *Aviation Week & Space Technology.* 17 July.

Hughes, David, & Dornheim, Michael A. (1995). Accidents Direct Focus On Cockpit Automation. *Aviation Week & Space Technology.* 30 January.

Hughes, David. Incidents Reveal Mode Confusion. (1995). *Aviation Week & Space Technology.* 30 January.

Jensen, R.S., & Benel, R. (1977). Judgment evaluation and instruction in civil pilot training (FAA Report No. N-FAA-Rd-78-24). University of Illinois Aviation Research Laboratory.

Parasuraman, R. (1987). Human-computer monitoring. *Human Factors, 29*, 695-706.

Parasuraman, R. (1997). Human use and abuse of automation. In M. Mouloua and J. Koonce (Eds.), *Human-automation interaction: Research and practice* (pp. 42-47). Mahwah, NJ: Erlbaum.

Pariès, Jean, & Amalberti, René. Aviation Safety Paradigms and Training Implications. In Sarter Nadine B., and Amalberti, René (Eds.). *Cognitive Engineering in The Aviation Domain.* Mahwah, NJ. Erlbaum, 2000.

Phillips, Edward H. (1999). Survey Reveals Automated Cockpits Harbor Sarety Pitfalls. *Aviation Week & Space Technology.* 22 November.

Sarter, Nadine B. & Woods, David D. (1991). Situation Awareness: A Critical But Ill-Defined Phenomenon. *The International Journal of Aviation Psychology.* 1, 45-57.

Sarter, Nadine B. & Woods, David D. (1992). Pilot Interaction With Cockpit Automation: Operational Experiences With the Flight management system. *The International Journal of Aviation Psychology.* 4, 303-321.

Sarter, Nadine B. & Woods, David D. (1994). Pilot Interaction With Cockpit Automation II: An Experimental Study of Pilots' Model and Awareness of the Flight management system. *The International Journal of Aviation Psychology.* 4, 1-28.

Sherman, Paul J., Helmreich, Robert L., & Merritt, Ashleigh C. (1997). National Culture and Flight Deck Automation: Results of a Multinational Survey. *The International Journal of Aviation Psychology.* 7, 311-329.

Vincenzi, D. A. and Mouloua, M. (1998). Monitoring Automation Failures: Effects of Age on Performance and Subjective Workload. *Proceedings of the 3rd Annual Automation Technology and Human Performance Conference*, Norfolk, VA: Old Dominion University.

Weigmann, Douglas D., & Shappell, Scott A. (1999). Human Error and Crew resource management Failures in Naval Aviation Mishaps: A Review of U.S. Naval Safety Center Data, 1990-96. *Aviation, Space & Environmental Technology.* 70 (12), 1147-1151.

The Impact of Teamwork Skills Training on Cadet Leadership and Unit Cohesion: An Examination of Military Team Performance

Kari R. Strobel and Robert M. McIntyre
Old Dominion University

Elizabeth C. Stubbs Koman
Human Performance Center, United States Navy

ABSTRACT

This study was an examination of the effects of teamwork skills training on cadet leadership, unit cohesion and performance. Throughout the course of a college semester, ROTC units completed various field tasks and were tracked with regard to levels of cohesion, performance, and effective leadership behaviors.

Results indicate that the teamwork skills training intervention had a significant positive impact on unit cohesion and performance. As predicted, trained unit leaders were successful in completing operational objectives by encouraging and reinforcing correct and effective teamwork behaviors such as communication, monitoring, backup, and feedback.

Keywords: Unit Cohesion; Cadet Leadership; Performance; Teamwork Training

INTRODUCTION

Unit cohesion is recognized as a desirable attribute which characterizes successful teams (Siebold, 1999). Cohesion is a multidimensional construct defined as an attraction to a team in pursuit of either social affiliation or task-related goals. Leaders from many disciplines identify cohesion as a necessary team property, and as such, behavioral scientists have become very interested in developing interventions which foster this team quality (Prapavessis & Albert, 1997).

The military is perhaps the most prominent organization which routinely touts cohesion as necessary for optimal team development and performance (Oliver, Harman, Hoover, Hayes, & Pandhi, 2000). For instance, a review of professional military training curricula for company grade officers emphasizes the need for junior officers to learn how to develop and foster this team trait (Barucky, 1985). These training courses recognize the long held belief that military performance is dependent on personnel coordination and interaction during all operational phases (Orasanu & Backer, 1996). Cohesion among troops facilitates these critical tasks, and it has been found that cohesion also serves a variety of protective functions that are vital to achieving military goals (Zaccaro, Gualtieri, & Minionis, 1995).

This research was an attempt to provide the United States military with an improved training tool for these purposes. Specifically, it was our goal to examine the effects of a teamwork skills training program, based upon the seven dimensions and principles of teamwork (communication, team orientation, team leadership, monitoring, feedback, backup, and coordination) derived from Dickinson, McIntyre, Ruggeberg, Yanushefski, Hamill and Vick (1992), and McIntyre and Salas, (1995) on unit cohesion, leadership, and performance.

It was expected that Reserve Officer Training Corps (ROTC) cadets receiving teamwork skills training would report and maintain increased levels of team cohesion and team performance over time. In addition, research suggests that team leadership may be one of the most critical ingredients in effective team performance, impacting a multitude of teamwork processes including cohesion; therefore, it was hypothesized that teamwork training would facilitate cadet leadership. Through training, leaders would be encouraged to consciously manage the team climate by soliciting and reinforcing correct and effective teamwork behaviors.

METHOD

ROTC units received systematic training on fundamental teamwork components so as to develop leadership and cohesion in the pursuit of a clearly identified and personally salient performance goal. Randomly assigned units made up of four cadets were trained on principle factors of teamwork, and monitored on three occasions over the course of a college semester. Controls in matched groups did not receive the training but were compared on all measures of cohesion, leadership and performance.

PARTICIPANTS

Two ROTC companies, divided into two equal units, each unit consisting of four senior cadets participated in the research. Participants consisted of 10 men (63%) and 6 women (37%). The average age of the participants was 20.27 (SD = 2.87) years. The sample was composed of 69% Caucasian, 25% African American, and 6% Pacific Islander. Their mean cumulative grade point average was 2.74 (SD = .40). All of the cadets were advanced undergraduates students. Eighty-two percent of the participants reported having previous teamwork experience working within the context of teams. These students came from varied academic backgrounds including the college of sciences, liberal arts, and engineering.

PROCEDURE

Following the random assignment of cadets to units; units were then randomly assigned to either the experimental or control condition. Cadets in the experimental group received formal personnel training on teamwork concepts. Cadets in the control group received no training but met and received the same measures at the same measurement intervals.

Once units and conditions were established, cadets were instructed to complete their first field task. At this time, baseline levels of cohesion, performance, and leadership were assessed. In addition, background information on individual unit members was collected. Approximately, one week later, the experimental group was provided the formal teamwork training. Cadets in the control group did not receive training but were required to meet at the same time as the training groups in an alternate location. Unit cohesion, performance and leadership were assessed again, eight weeks later, during mid-term follow-up. Final assessments of each outcome were made at the end of the semester.

The objective of the teamwork skills training program was to have team members identify, define, and demonstrate the seven core components of teamwork as defined by Dickinson et al. (1992). A variety of sources, including previous experiments' methods, team training literature, and books on training, were consulted to select the most appropriate methods for training. A combination of lecture, discussions, games, and behavioral modeling were chosen for the methods. The training program itself was evaluated at the end of the training session by asking participants to complete a post-training evaluation questionnaire requesting participants' reactions to the teamwork skills training.

A variety of activities were included in the training. Blanchard and Thacker (1998) suggested the use of relevant examples, behavioral reproduction (practice), and feedback to maximize trainee learning. These and other learning theories helped guide the development of the training program. Initially, team members were given the Teamwork Skills Knowledge Pre-Test. This test was administered prior to intervention as a way of assessing baseline knowledge of leadership, team orientation, communication, monitoring, feedback, back up, and coordination; components which directly reflect characteristics that were the focus of the training program.

Introductory activities were used to introduce participants to the training topic objectives. Definitions and examples of the seven principles of teamwork were then given via lecture, by an advanced graduate student researcher. Following the lecture, team members viewed portions of popular movies highlighting teams of actors engaging in the seven teamwork behaviors. A team building activity was then used to allow team members to practice the skills in a non-stressful setting while other members observe for the teamwork components. At this time, teams were asked to complete a tower building exercise (Moore, 1992). The Teamwork Skills Knowledge Test was again administered. Finally, participants were asked to evaluate the training session and to assess the perceived effectiveness of the training program. After the intervention, teams were encouraged to track the frequency with which the seven behaviors occurred on a team log. Team logs were given to team members upon completion of the training program. In order to ensure the transfer of the teamwork skills training, teams received weekly "team-o-grams". Team-o-grams were reminder messages sent to team members via electronic messaging, to serve as boosters to the points provided in the training. The entire training program lasted approximately three hours. The training was conducted on the Old Dominion University main campus at the Department of Military Science.

MEASURES

The System for the Multiple Level Observation of Group (SYMLOG; Bales & Cohen, 1980) Adjective Rating Form was administered to all participants as the primary means of assessing cohesion. The SYMLOG is a 26 item self-report measure that utilizes a five-point Likert scale to measure both social and task dimensions: Friendly-Unfriendly (P-N); Task-Oriented-Emotionally Expressive (F-B) of cohesion.

The Leadership Assessment Report (LAR) served as the primary means of assessing cadet leadership. The LAR is an observation tool designed to measure unit leader behaviors such as influencing, operating, planning, and communicating. Unit leaders were assessed in the field by commanding officers using a three-point scale (0 = needs improvement, 1 = satisfactory, and 2 exceptional). Leaders were provided with scores on each behavior of interest. These scores were totaled to provide an overall assessment of unit leadership.

Unit performance was defined in terms of the teams' successful completion of practical field exercises. Units were judged on their successful completion of radio communications drills, safety assessments, movement techniques, planning for team safety/security, and pre-execution techniques. Field observations were made by commanding officers and each team task was scored using a three-point scale (0 = needs improvement, 1 = satisfactory, and 2 = exceptional). Scores on each team task were totaled to provide an overall performance assessment.

The Teamwork Skills Knowledge Test was created on the basis of the teamwork process model described by Dickinson and McIntyre (1997). It assesses knowledge of the Dickinson-McIntyre teamwork components: leadership, team orientation, communication, monitoring, feedback, back up, and coordination. Scores from this scale served as a "manipulation check," by assessing the degree to which training participants acquired knowledge of the teamwork concepts.

RESULTS

An alpha level of .05 was used for all statistical tests. There were no statistically significant differences between groups on demographics.

A manipulation check was performed to gauge the success of the training program (Kazdin, 1998). A paired samples t-test was conducted comparing pre (mean = 31.25, SD = 10.25) and post-intervention (mean = 91.57, SD = 8.10) scores on the Teamwork Skills Knowledge Test. Results indicate they there were significant differences, $t(8)$, 19.38, $p = .00$. Thus, evidence suggests that the cadets successfully learned the teamwork concepts provided in the training program.

Analysis of Variance (ANOVA) was the main statistical technique employed to analyze the effects of teamwork skills training on unit cohesion, performance, and cadet leadership. Results for the P-N social cohesion dimension show that there was not a significant difference between controls and trainees at baseline, $F(1,2) = .47$, $p = .56$. However, there were significant differences between controls and trainees at follow-up, $F(1,2) = 27.56$, $p = .03$, and at final follow-up, $F(1,2) = 25.92$, $p = .03$. In addition, results for the F-B task cohesion dimension show that there was not a significant difference between controls and trainees at baseline, $F(1,2) = 1.00$, $p = .42$, although there were significant differences between controls and trainees at follow-up, $F(1,2) = 17.00$, $p = .05$, and at final follow-up, $F(1,2) = 17.66$, $p = .05$. Means and standard deviations are presented in Table 1.

Table 1
Mean Levels of Unit Task and Social Cohesion

	Baseline	Follow-up	Final Follow-up
	Mean (SD)	Mean (SD)	Mean (SD)
Social Cohesion			
Trainees	7.50(3.53)	27.50 (3.53)	30.50 (3.53)
Controls	9.50(2.12)	7.00 (4.24)	12.50 (3.49)
Task Cohesion			
Trainees	3.50 (2.12)	21.50 (.70)	30.00 (2.82)
Controls	6.00 (2.82)	13.00 (2.82)	6.00 (1.41)

Results suggest that there was not a significant difference in unit performance between controls and trainees at baseline, $F(1,2) = .22$, $p = .69$. However, there were significant differences between trainees and controls at follow-up, $F(1,2) = 32.00$, $p = .03$, and at final follow-up, $F(1,2) = 18.00$, $p = .05$. Means and standard deviations are presented in Table 2.

Table 2
Mean Levels of Unit Performance

	Baseline	Follow-up	Final Follow-up
	Mean (SD)	Mean (SD)	Mean (SD)
Trainees	5.50 (3.53)	9.50 (.71)	9.00 (1.41)
Controls	4.00 (2.82)	5.50 (.67)	3.00 (1.38)

Results suggest that there was not a significant difference in effective cadet leadership behaviors between groups at baseline, $F(1,2) = .24$, $p = .67$. However, following the team training intervention, trained leaders performed significantly better than their control group counterparts at follow-up, $F(1,2) = 21.16$, $p = .04$, and at final follow-up, $F(1,2) = 27.77$, $p = .03$. Means and standard deviations are presented in Table 3.

Table 3
Mean Levels of Effective Leadership Behaviors

	Baseline	Follow-up	Final Follow-up
	Mean (SD)	Mean (SD)	Mean (SD)
Trainees	4.50 (.71)	19.50 (2.12)	20.50 (2.13)
Controls	3.00 (4.24)	8.00 (2.82)	11.00 (1.41)

DISCUSSION

The results of this study were consistent with the principle hypotheses that a training program can be developed to enhance unit cohesion, performance, and cadet leadership. The data demonstrated that for the experimental units involved, training based on the teamwork components model raised cohesion and performance levels above baseline observations. Furthermore, trained cadets out led their control group counterparts. These cadets displayed effective team leadership behaviors including encouraging team members to make appropriate decisions and providing support and direction to unit members during completion of operational objectives.

Despite the favorable results, the study had several limitations. Foremost among these, the study was only concerned with the effects of team training on newly formed teams. Therefore, the utility of the intervention on established teams cannot be determined within the limited framework of this study. Finally, the study examined only a very small number of ROTC units. Additional research is needed examining the effects of training on a greater number of teams. Finally, subsequent research should be conducted to determine the applicability of findings to other settings of interest. Specifically, it would be worthwhile to pursue the effectiveness of the training model in athletic teams, cross-functional and self-managing work teams. More importantly, it would be interesting to analyze the efficacy of the intervention on global teams due to the increasing number of multicultural teamwork within organizations.

Several strengths to the study also deserve to be highlighted. Methodologically, the use of an experimental design and the manipulation of the independent variable lend strong support to the conclusion that the training can produce rapid improvements in cohesion, performance, and leadership. The use of both intellectual and physical team tasks help defend the data against threats to external validity. For example, if leadership, cohesion, and performance were measured solely within the context of accomplishing an intellectual team task, tasks that emphasize physical skill performance may respond differently to the team skills training.

CONCLUSION

Taken as a whole, the implications of this study are potentially far-reaching. Given that the training is empirically derived, behaviorally based, time limited, and financially inexpensive, with a minimum of effort, it is easily translated into usable military training to give military leaders, especially junior officers, the practical tools necessary to effectively lead their units.

REFERENCES

Bales, R. F. & Cohen, S. P. (1979). *Symlog: A System for the Multiple Level Observation of Groups*. New York: Free Press.

Barucky, J.M. (1985). Identification and measurement of the personal leadership development concerns of Air Force officers, cadets, and trainees participating in Air Force professional military education programs. *Dissertation Abstracts International, 46*(2-A), 326-327.

Blanchard, P. N. & Thacker, J. W. (1998). *Effective training: Systems, strategies, and practices*. Englewood Cliffs, NJ: Prentice Hall.

Dickinson, T. L., McIntyre, R. M., Ruggeberg, B. J., Yanushefski, A., Hamill, L. S., & Vick, A. L. (1992). *A conceptual framework for developing team process measures of decision-making performance* (Final Report). Orlando, FL: Naval Training Systems Center, Human Factors Division.

Kazdin, A. E. (1998). *Methodological issues and strategies in clinical research.* Washington, DC: American Psychological Association.

McIntyre, R. M., & Salas, E. (1995). Measuring and managing for team performance: Emerging principles from complex environments. In R. A. Guzzo & E. Salas (Eds.). *Team Effectiveness and Decision-Making in Organizations* (pp. 9-45). San Francisco: Josey-Bass Publishers.

Moore, A. C. (1992). *The game finder: A leader's guide to great activities.* State College, PA: Venture.

Orasanu, J. M., & Backer, P. (1996). Stress and military performance. Driskell, James E. (Ed.), Salas, Eduardo (Ed). (1996). *Stress and human performance: Series in applied psychology* (pp. 89-125). Hillsdale, NJ, US: Lawrence Erlbaum Associates.

Prapavessis, H. & Albert, A. V. (1997). Sacrifice, cohesion, and conformity to norms in sports teams. *Group Dynamics,* 1(3), 231-240.

Siebold, G. L. (1999). The evolution of the measurement of cohesion. *Military Psychology*, 11(1), 5-26.

Zaccaro, S. J., Gualtieri, J., & Minionis, D. (1995). Task cohesion as a facilitator of team decision making under temporal urgency. *Military Psychology, 7(2), 77-93.*

ADAPTIVE MULTIMODAL INTERFACES IN TRAINING AND OPERATIONAL ENVIRONMENTS

David A. Graeber
Boeing Phantom Works

Laura Milham & Kay M. Stanney
Design Interactive

Lt. Joseph Cohn
Naval Research Laboratory

Dylan Schmorrow
DARPA

Colby Raley
Strategic Analysis

ABSTRACT

Virtual environment (VE) systems have advanced into readily available, low-cost, and portable devices that can train personnel on a broad range of skills. Virtual Technologies and Environments (VIRTE) is a program at the Office of Naval Research that aims to leverage VE technology to enhance Navy and Marine Corps expeditionary warfare training. It involves the use of VE technologies for small unit training, mission rehearsal, and other mission critical task training. VIRTE is also involved in the construction of VE training systems, as well as analysis and experimentation for their refinement. This affords the opportunity to incorporate natural, multi-modal communication between humans and machines into VIRTE products based on requirements analyses for both training and operational environments. Presented herein is the application of requirements analysis to the creation of multi-modal interfaces for VIRTE's Virtual Environment Landing Craft Air Cushion (LCAC) training system.

KeyWords: Multimodal, Virtual Environments, VIRTE, Interface Design, Human Performance, Human Systems Integration, Adaptive Interfaces

INTRODUCTION

The U.S. Navy's Office of Naval Research (ONR) Virtual Technologies and Environments (VIRTE) program is conducting research on the application of virtual environment technologies to Naval training problems. The VIRTE program has a number of components. The "Demo I" component is developing networked, interoperable virtual environment training systems for three expeditionary warfare systems, one of which is the Landing Craft Air Cushion (LCAC) whose virtual environment training system is referred to as VELCAC. VIRTE is part of ONR's Capable Manpower Future Naval Capability (FNC), which is tasked with developing technologies that meet a fleet need and can be transitioned to an existing acquisition program. In the case of the VELCAC program, the transition customer is the Naval Sea Systems Command PMS 377J – LCAC Transition and Lifecycle. The overall requirement for the VELCAC system is to provide PMS 377J a prototype and a vision for desirable SLEP (Service Life Extension Plan) LCAC interim training capabilities. A prototype VELCAC system has been developed for PMS 377J, however, the focus herein is to describe how the requirements analyses performed during VELCAC's development was used to derive a notional framework for adaptive multimodal VELCAC operator interfaces. In the following sections the requirements analysis process utilized and its findings are delineated. The data from the requirements analyses are then mapped to a Media Allocation Model (MAM) (Samman & Stanney, 2003) to appropriately apply multimodal information perceptualization (MIP) techniques to VELCAC operator interfaces. This document concludes with a discussion on the potential benefits of instantiating multimodal interfaces to future military systems.

Requirements Analysis Process and Allocation of Complementary Multimodal Interfaces

To ensure an effective and efficient VELCAC design, an iterative user-centric Human-Systems Integration (HSI) effort was conducted. From the outset, this effort was conceived with the goal of complementing conventional Systems Engineering efforts, thereby ensuring design solutions and assessment criteria that adequately meet trainees' needs. The HSI process evolved VELCAC's interface design throughout the development lifecycle, starting with gathering operator knowledge on work tasks and culminating in interface design validation and usability evaluation. It is the initial requirements analysis and gathering of operator knowledge that provides a foundation for the proper allocation of multimodal interface components to a system. This step necessitated analyzing system concepts, requirements, and design documents, as well as knowledge engineering findings on the user population and work practices to both support interface design and ensure training effectiveness. The most critical aspect of this node in the overarching HSI effort is the knowledge engineering findings on work practices because it allows one to decompose their tasks to a level of granularity where the appropriate interface modality can be selected based on task attributes.

A portion of the requirements analysis for VELCAC focused on the components of LCAC operation that are universal across a variety of missions. This analysis looked at task flows, data acquired from cockpit displays, information exchange among crewmembers and other crews, and associated environmental cues that the LCAC 3-person crew (i.e., Craftmaster, Engineer, and Navigator) utilize to perform the following universal tasks: 1) collision avoidance, 2) formation flight, 3) surf zone transition, and 4) reduced visibility conditions. The task of collision avoidance in formation flight is expanded upon below to demonstrate the application of MIP techniques to enhance VELCAC training.

Collision Avoidance in Formation Flight

Collision avoidance is an essential component of safe and effective LCAC operations regardless of the mission objective. The responsibility for detecting contacts (i.e., boats, buoys, etc. that may result in a collision or incursion) in the immediate operational environment primarily falls upon the Navigator due to his control over the RADAR display. However, the Craftmaster and Engineer also aid in collision avoidance by scanning eyes-out and confirming visual recognition of a contact. The current process of collision avoidance is a highly visual task that keeps the Navigator eyes-in scanning the RADAR display and requires rapid, on-the-fly mental calculations of course corrections (heading and speed) to avoid contacts and maintain H-hour (the window of time for crossing the craft penetration point). The Navigator is also saturated with radio communications emanating from crewmembers and other LCACs when flying in formation. Table 1 below sequentially lists the high level tasks involved in formation flight collision avoidance and accompanying cues conventionally communicated during flight. Coupled with each task is a suggested complementary MIP technique that would offload excessive demands on the visual system and facilitate processing of communications. The suggested complementary MIP techniques are based on human information processing and sensory integration capabilities in the context of extending a visually-based task to multiple modalities.

For each task step in Table 1 a complementary adaptive MIP technique has been suggested based upon the Media Allocation Model (Samman & Stanney, 2003), which aims to optimize information processing across the sensory systems. Current systems primarily use visual displays, whose processing is limited by humans' visual capacity. In essence, visual and visuo-spatial processing become bottlenecks when interacting with visually-based displays, leaving other sensory capabilities and cortical processing centers largely untapped (Stanney, Samman, et al., 2003). Complementary MIP techniques make use of these untapped cortical processing centers and relieve the workload on visual processing. The adaptive component refers to MIP techniques that can change dynamically in response to a change in either user or system state. The adaptive component could be controlled by critical system-related events (e.g., low task performance indicator) or via the operator's psychophysiological state (e.g., EEG with brain activity indicating high mental workload), which would trigger mitigation strategies to offload workload and maximize performance overall and within each sensory system.

Table 1. High level task steps and cues associated with formation flight collision avoidance accompanied by complementary multimodal information perceptualization (MIP) techniques.

Task Step	Cue/Data Source	Complementary MIP techniques
Detection of contact on RADAR or out the window (OTW)	Visual indicator on RADAR screen or visual acquisition of contact OTW	Spatialized audio and/or localized tactile cues
Determine range and bearing of the contact	Range and bearing line data on RADAR confirmed by OTW visual estimation	Spatialized audio and/or localized tactile cues; earcons (i.e., non-verbal audio messages often time created via variations in pitch, loudness, timbre)
Visual confirmation of contact by all crewmembers	OTW visual recognition	Spatialized audio and/or localized tactile cues
Collaborative decision making among Navigators in the formation on contact threat level	RADAR display and radio communications	Spatialized audio
Lead Navigator decides course alteration to avoid contact and relays it to the formation	RADAR display and radio communications	Spatialized audio
Craftmaster receives confirmation from the lead Navigator on course correction and maneuvers craft	Radio communications, OTW visual of the contact and other LCACs in the formation	Spatialized audio and/or localized tactile cues
Lead Navigator determines course corrections to maintain H-hour	RADAR display, paper charts, and whiz wheel	Automated decision making
Course corrections relayed to formation and implemented by Craftmaster	Radio communications	Spatialized audio and/or localized tactile cues

Collision avoidance and navigation in an LCAC is a highly spatial task that has traditionally been thought of as being best presented via visual display (Wickens, 1992). However, spatial information can be effectively presented as sound localization, variations in pitch, or localized tactile or kinesthetic cues (Stanney, Samman, et al., 2003). Furthermore, Bach-y-Rita (1999) has demonstrated the ability to substitute spatial information presented visually with tactile "vision". This suggests that the traditional overload on visual processing can be circumvented by instantiating alternate spatial auditory and haptic interfaces. Furthermore, Blauert (1996) has shown that spatialized audio is effective for presenting a multitude of simultaneous sound sources in different locations, thereby aiding comprehension. This suggests that spatialized audio would be an effective aid for monitoring multiple radios on board a LCAC and communications among LCACs in a formation. One can anticipate substantial performance enhancements via such spatialized audio communications (comms). For example, Nelson and Bolia (2003) demonstrated that spatialized audio along the horizontal plane enhanced call sign identification by approximately 50%, as well as speeded reaction time. Thus, for the VELCAC, simply by localizing communications, say placing the Craftmaster comms at +20 degrees, the Engineer comms at -20 degrees, and the Navigator comms at -90 degrees along the horizontal plane, one can anticipate large gains in comms identification and processing.

The two above paragraphs are grounding for why complementary MIP techniques would benefit the task of collision avoidance in formation flight; the paragraphs below further detail the reasoning behind the complementary MIP recommendations in Table 1. The first task step listed in Table 1 is detection of the contact by visual acquisition out the window (OTW) or, more likely, by the Navigator via RADAR. The detection of a contact integrates various task attributes that are well suited to auditory and tactile displays. These task attributes include 3-D localization, detecting objects in the periphery, and expedient reaction to alerts/warnings. When extending a visually based 3-D localization task to multiple modalities it is suggested that spatialized audio and/or localized tactile displays be used to supplement visual 2-D displays because visual displays compress one or more dimensions, whereas audition and/or tactile displays are omnidirectional. In addition, spatialized audio and localized tactile displays afford better performance than visual displays for perceiving absolute and relative locations of objects in 3-D space. The omnidirectional characteristic of spatialized auditory and localized tactile displays also supports their use for directing one's attention to an area in the periphery or outside the visual attention envelope. Furthermore, spatialized auditory and tactile cues can be effectively processed when an operator is in motion (Samman & Stanney, 2003). For the alert/warning aspect of contact detection, spatialized auditory "earcons" and/or tactile inputs provide a redundant cue that, when emanating from the same spatial location, decrease search time, facilitate object detection, enhance attention, and decrease reaction time (Kalawsky, 1993).

The second task in Table 1 pertains to determining the range and bearing of a contact. The combination of spatialized audio and/or localized tactile input(s) with additional auditory earcons (i.e., variations in pitch, loudness, timbre) can effectively supplement a visual range and bearing line creating a more immersive representation of the contact's location. The spatialized component could provide the bearing of the contact, while auditory earcons could convey range.

The remaining task steps for formation flight collision avoidance center on communication within the cockpit and among LCACs in the formation, as well as visual confirmation of a contact in the operational environment. The implementation of spatialized audio and/or localized tactile displays would support the visual acquisition of a contact in the operational environment for reasons discussed above. With respect to the communications component in LCAC operations, it is highly saturated due to the monitoring of up to 5 channels of communications. To facilitate the monitoring of the 5 radios, as well as enhancing recognition of which craft in the formation one is communicating with, spatialized audio can assign each radio a distinct location in the 3-D auditory envelope along the horizontal plane. The use of the horizontal plane is suggested due to its effectiveness as found in Nelson and Bolia (2003), as well as because, in general, the processing of horizontal position is relatively fast (Frens & Van Opstal, 1995), probably due to the use of binaural differential hearing. Communications from the LCACs in the formation could be mapped to their location (to the right or left of ownship) in the operational environment, which could also be redundantly coded by localized tactile cues, to not only facilitate recognition of which craft one is communicating with, but also provide a sense of craft separation to support station keeping.

Benefits of Multimodal Interfaces

An important HSI challenge in current and future military systems is opening new information processing pathways to alleviate bottlenecks in visual and visuo-spatial processing pathways created by heavy reliance on visual display techniques. A key to achieving this is MIP techniques that capitalize on inherent human sensory system characteristics, integration capabilities, and adaptability to optimize cortical processing of extensive sensor data. It has been shown that multimodal interfaces can facilitate cognitive operations by enhancing perception, speeding reaction time, and bolstering memory, thereby yielding effective tools for teaching and learning (Kalawsky, 1993). Studies in object recognition have demonstrated our innate capability for perceptual integration of multiple sensory modalities, which enhances detection of the object via amplifying sensory signals and creating multimodal representations (O'Hare, 1991). It has also been shown that redundant coding of information using multimodal representations hastens reaction time (Miller, 1982) and improves memory performance (Sulzen, 2001). Multimodal displays also aid conceptualization of a problem space by employing visual, auditory, and haptic techniques to assist users in finding relevant data, visualizing domain semantics, and restructuring their view of a problem (Woods & Roth, 1988). Iterations between these display mediums and fusing them together results in an information processing – problem solving feedback loop that affords querying and refinement of hypotheses about data from both unique and fused perspectives (Ware, 2000).

Leveraging cross-modal effects in an adaptive feedback display is a powerful technique for expanding operator information processing capacity and mitigating information processing bottlenecks. Humans constantly experience and correlate parallel stimulation of various sense modalities from external events or objects in our daily interactions. The brain combines these inputs to forge multimodal determined percepts (Driver & Spence, 2000) that lead to marked improvements in the detection, localization, and discrimination of external stimuli and quicken reaction, assuming the correct task-relative cross-modal synthesis occurs (King & Calvert, 2001). Taken together, the aforementioned benefits coupled with advancement in multimodal interface technology (particularly auditory and haptic interfaces) implore a transformation from visually burdened user interfaces into next generation MIP displays that capitalize on humans' innate multi-sensory integration capabilities.

REFERENCES

Bach-y-Rita, P. (1999). Theoretical aspects of sensory substitution and of neurotransmitter-related organization in spinal cord injury. Spinal Cord, 37, 465-474.

Blauert, J. (1996). Spatial Hearing - The Psychophysics of Human Sound Localization (revised edition). The MIT-Press, Cambridge MA.

Driver, J.S., & Spence, C.J. (2000). Multisensory perception: Beyond modularity and convergence. Current Biology, 10, 731-735.

Frens, M.A., & Van Opstal, A.J. (1995). A quantitative study of auditory-evoked saccadic eye movements in two dimensions. Experimental Brain Research, 107, 103-117.

Kalawsky, R.S. (1993). The science of virtual reality and virtual environments. Wokingham, England: Addison-Wesley.

King, A.J. & Calvert, G.A. (2001). Multisensory integration: Perceptual grouping by eye and ear. Current Biology, 11, R322-R325.

Miller, J.O. (1982). Divided attention: Evidence for coactivation with redundant signals. Cognitive Psychology, 14, 247-279.

Nelson, T., & Bolia, R. (2003). Evaluating the effectiveness of spatial audio displays in a simulated airborne command and control task. Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting (pp. 202-206). Denver, CO, October 13-17.

O'Hare, J.J. (1991). Perceptual integration Journal of Washington Academy of Sciences, 81, 44-59.

Samman, S., & Stanney, K.M. (2003). Media Allocation Model. Design Interactive, Inc., Oviedo, FL.

Stanney, K.M., Samman, S., Reeves, L.M., Hale, K., Buff, W., Bowers, C., et al. (2003). A paradigm shift in interactive computing: Deriving multimodal design principles from behavioral and neurological foundations. Manuscript in preparation.

Sulzen, J. (2001). Modality based working memory. Retrieved, January 14, 2004 from http://ldt.stanford.edu/~jsulzen/james-sulzen-portfolio/classes/PSY205/modality-project/paper/modality-expt-paper.PDF.

Ware, C. (2000). Information visualization: Perception for design. San Diego, CA: Academic Press.

Wickens, C.D. (1992). Engineering psychology and human performance (2nd ed.). New York: Harper Collins.

Woods, D.D., & Roth, E.M. (1988). Cognitive engineering: Human problem solving with tools. Human Factors, 30(4), 415-430.

THE EFFECTS OF VIEWING MEDIUM ON DEPTH PERCEPTION IN HUMAN PERFORMANCE OF A TELEROBOTIC MANIPULATION TASK

Clayton J. Hutto, Dennis A. Vincenzi, Steve Hall and Sathya Gangadharan
Embry-Riddle Aeronautical University

ABSTRACT

Telerobotics are being used in several domains (such as space, undersea, medicine and surgery, bomb disposal, or toxic material clean-up) as a means of extending human abilities to remote environments. Some of these tasks may be performed by an operator who has an unobstructed direct stereoscopic view of the environment. Unfortunately, many of these environments require fine manipulation of objects that are outside of the operator's field of vision, and visual information of the environment must be relayed via remote video. This study addresses the performance differences for teleoperators who attempt a robotic manipulation in either direct stereoscopic viewing conditions, or while viewing the task environment with a monoscopic video monitor. Participants performed ten telerobotic placement attempts, and were judged for performance based on the average time to complete the placement attempts, as well as their placement accuracy for each attempt. Results of this study suggest that telerobotic operators rely heavily on the stereoscopic depth cues that are available in binocular vision, and that viewing medium should be considered a relevant factor for operators when performing telerobotic tasks.

Keywords: Telerobotics; Depth Cues; Binocular, Stereoscopic Vision; Monocular, Monoscopic Vision

INTRODUCTION

Due to the hazardous nature of some environments, such as explosive ordinance disposal and toxic material removal, remote robotic manipulation is quickly becoming an ideal method of increasing human safety by removing the human operator from the dangerous setting (Sheridan, 1992). The use of telerobotic systems also gainfully extends human capabilities into unstructured environments (e.g., space and undersea), and can be used to reduce human performance limitations such as lack of strength and resistance to fatigue (Sheridan, 1992; Bounds, Schroer, & Schroer, 1990; Pepper & Hightower, 1984).

Since the visual system is the primary means by which most humans gain spatial information of objects in their environment (Chapanis, 1996), it follows that accurate visual sensory input is vital while conducting telerobotic manipulations. Unfortunately, telerobotic tasks such as space station construction or undersea exploration are often conducted in environments and locations that extend beyond the operator's direct field of vision. As a result, a relay of visual information from the environment to the human operator is necessary. This is accomplished most often via video monitoring (Horikawa & Nagatomo, 1998; Park & Woldstad, 2000; Sheridan, 1992; Yeh & Silverstein, 1992; McGovern, 1991; Bounds, Schroer, & Schroer, 1990). Consequently, telerobotic systems used for spatial manipulation tasks in remote environments require a visual display system that adequately accounts for human visual perception limitations. More specifically, careful consideration of the effects of a human's attempt to perceive three-dimensional information from two-dimensional video monitors is critical (Kim, Tendick, & Stark, 1991), and an understanding of the human performance differences between three-dimensional stereoscopic viewing conditions and two-dimensional monoscopic viewing conditions is necessary. This study examines the effects of viewing conditions on human performance for teleoperators who perform a simple robotic placing task while viewing the environment either directly with stereoscopic vision, or indirectly via two-dimensional monoscopic video monitoring.

Background

A typical telerobotic task involves the collection, manipulation, and accurate placement of objects within a remote environment (often referred to as pick-and-place tasks). To accommodate these tasks, an operator needs visual information from the environment, which is usually provided via video monitoring. However, when three-dimensional spatial information is displayed on a two-dimensional monitor, the operator must mentally interpret the information in order to translate the 2D scene into an accurate representation of the remote 3D environment. Regardless of the graphical accuracy of the display, human interpretation may result in misrepresentation. Detrimental consequences of inadequate teleoperator interpretations of the visual information that is relayed from the remote environment may result. A teleoperator may mishandle the remote robot, producing unintentional collisions of objects, thereby causing damage to the objects, the telerobotic equipment, or both. It is therefore important to understand human performance differences within the context of viewing medium for telerobotic systems.

Due to the combined advantages of economy of cost, easy availability, and suitability to visual information transmission, conventional video communication systems in teleoperation often consist of monoscopic cameras and two-dimensional monitors. Unfortunately, however, standard monoscopic video systems cannot match the level of visual and depth acuity of the human visual system. Monoscopic video is very capable of relaying limited depth information through a variety of pictorial cues such as interposition (occlusion), lighting effects such as shading and shadows, linear and geometric perspectives, texture gradients, and size constancy of familiar objects. The human visual system, however, is much more adept at picking out depth information due to stereopsis. Stereopsis is the ability to extract depth information from binocular cues (Coren, Ward, & Enns, 1999). Binocular and oculomotor cues such as retinal disparity (angular offset between retinal images in the left and right eyes), vergence movements (rotation of the eyes to a point in space), and accommodation (compression or expansion of the lens to focus at a particular distance) all combine to produce human stereopsis (Coren, Ward, & Enns, 1999).

Numerous studies have investigated the effects of stereoscopic viewing and monoscopic viewing conditions within the context of aviation (Haskell & Wickens, 1993; Ellis, McGreevy & Hitchcock, 1987), for scientific visualization (Wickens, Merwin, & Lin, 1994; Sollenberger & Milgrim, 1993), and for remote operations (Massimino & Sheridan, 1994; Pepper & Hightower 1984; Drascic, 1991; Drascic, Milgrim, & Grodski, 1989; Lumelsky, 1991). Many of these studies have ambiguous, conflicting, or sometimes intuitively contradicting results. For example, while measuring mean task times for teleoperator performance, Massimino & Sheridan (1994) did not find significant differences in direct versus video viewing conditions. Park & Woldstad (2000) found that in the absence of visual enhancement depth cues, teleoperators performed better with multiple two-dimensional monoscopic video displays than with either monoscopic or stereoscopic three-dimensional displays. Also, Bejczy (1976) reported significantly poorer performance for pick-and-place tasks with stereoscopic displays than with monoscopic displays. On the other hand, there is an abundance of contradicting studies that show superior teleoperation performance with stereoscopic displays (Drascic, 1991; Pepper, Smith, & Cole, 1981; Kim, Ellis, Tyler, Hannaford, & Stark, 1987).

In the more elaborate video display systems using stereoscopic cameras (which generally combine the images from two offset cameras through various processes of multiplexing), the risk for damage to the video system in unstructured environments such as space, undersea, or bomb disposal is much more considerable than one might expect from teleoperation in a standard manufacturing setting. Damage to one camera or the other in these stereoscopic systems may leave a teleoperator in a monoscopic viewing condition. Additionally, just because a video system is able to render spatial information does not necessarily mean that an operator will accurately interpret the spatial information (McGreevy & Ellis, 1986). It is therefore important to understand what differences or relationships, if any, exist for performance of teleoperation tasks in different viewing conditions. In this study, the authors attempt to answer the following question: to what extent will the viewing medium used by telerobotic operators affect their performance of a manipulative task? It is hypothesized that operators will perform significantly better in the direct (stereoscopic) viewing condition due to the additional binocular depth cue advantages afforded to them. The superior performance will be evidenced both by increased placement accuracy and by decreased average time-to-completion for the telerobotic manipulation task.

METHOD

Participants

A total of 180 naïve participants (131 male and 49 female) volunteered for this study. Ages ranged from 18 to 47 (mean = 21.58, SD = 3.57). None of the participants had previous experience with telerobotics. Three participants were replaced after reporting having visual acuity worse than 20/20 or known depth perception problems; all other participants reported normal or corrected to normal vision, and no problems with depth perception.

Apparatus

The monoscopic video camera used in this study was a Sony with model number CCD-TR87. A fifteen-inch Panasonic color monitor, model number CT13R14V, was used for the two-dimensional video display of the remote environment. The telerobotic system was a Questech Robot Manipulator Arm model number TCM, which was modified to allow the remote control to reach to a distance greater than 15 feet. A plastic ring measuring 3.81 cm in total diameter (center aperture measuring 2.7 cm), and a wooden dowel rod post with a diameter of 2.25 cm and a length of 40.5 cm, which was vertically fixed within the telerobotic arm work space, were used to evaluate performance of a remote manipulation of the telerobotic system.

Design

This study examined the differences in performance of a telerobotic-placing task based on the type of viewing medium afforded to the operator at a distances ranging from 20 cm to 250 cm. Viewing conditions were of two types: a) direct stereoscopic viewing, and b) indirect monoscopic viewing. Dependent variables included a) the accuracy of the placing task, measured by the number of times out of ten attempts that a participant successfully dropped a ring completely to the bottom of a dowel post, and b) the time to completion for the task, measured in seconds for each attempted drop, from the first motion of the robotic arm to the release of the ring. The study was conducted as a fully between-subjects experimental design.

Procedure

Each participant was shown the experimental apparatus with the telerobotic manipulator arm holding the ring, and was then instructed in the use of the manipulator arm. Upon the completion of the instructions for the telerobotic manipulator arm, the participants were asked to drop a plastic ring measuring 3.81 cm in total diameter (center aperture measuring 2.7 cm) over a wooden dowel rod post with a diameter of 2.25 cm and a length of 40.5 cm, which was vertically fixed within the telerobotic arm work space. The manipulator arm was reset to the same start position for each trial with the plastic ring being held in the arm's gripper. Participants were not allowed to view the work area while the manipulator arm was being reset to the start position by the researcher.

Accuracy was judged by the amount of times that the plastic ring fell to the bottom of the dowel rod out of ten drops (successful drops counted as "hits"). Rings that did not fall completely to the bottom of the dowel (i.e., rings that were hung up on the top of the dowel) or rings that missed the dowel were considered errors and were not counted as hits. Time was measured in seconds beginning from the first movement of the telerobotic manipulator arm and ending with the release of the ring.

In the direct stereoscopic view condition, a chinrest was used to ensure that all participants were at eye level with the top of the dowel, and to ensure each participant viewed the apparatus from the same distance. For the indirect view, a monoscopic video camera was leveled with the top of the dowel at the appropriate distance, and adjusted to approximate the same field of view and visual angel as the direct viewing condition.

RESULTS

This study examines the effects of viewing medium on human performance for teleoperators who performed a simple robotic placing task while viewing the environment either directly with stereoscopic vision, or indirectly via two-dimensional monoscopic video monitoring. *Table 1* summarizes performance data for each of the experimental groups. Figures 1 and 2 are graphical representations of the group means presented in Table 1 for task completion time and task accuracy.

Table 1: Group Means, Standard Deviations and Standard Errors

Measure	**Viewing Medium**	**Mean**	**SD**	*Std. Error*	*N*
Time	Direct (stereoscopic)	33.64	10.36	1.09	90
	Indirect (monoscopic)	55.33	19.53	2.06	90
Accuracy (out of 10)	Direct (stereoscopic)	6.66	2.37	.250	90
	Indirect (monoscopic)	2.58	1.87	.197	90

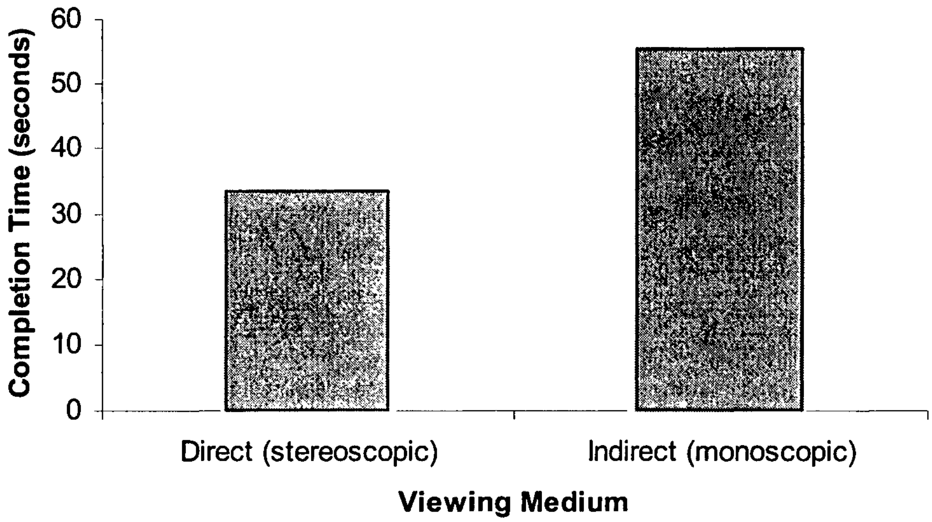


Figure 1: Average time to task completion

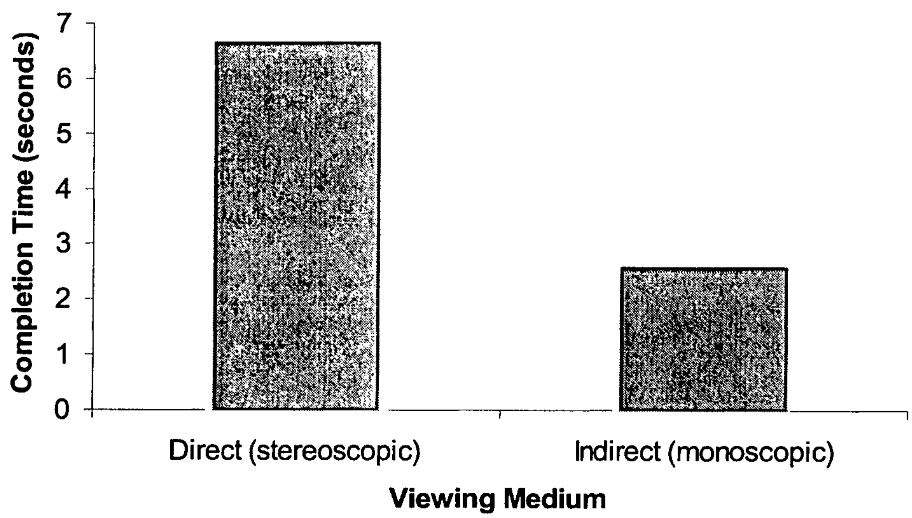


Figure 2: Drop accuracy for the assigned task

The data were analyzed using an independent samples two-tailed t-test. Results for the average time to completion of the task indicate that operators who performed with a direct (stereoscopic) viewing medium performed significantly better (M = 33.64, SD = 10.36) than operators who performed the task with an indirect (monoscopic video) viewing medium (M = 55.33, SD = 19.53), t (178) = -9.30, p < .001. Additionally, the results for drop accuracy also indicate superior performance for operators who performed with a direct (stereoscopic) viewing medium (M = 6.66, SD = 2.37) than for operators who performed the task with an indirect (monoscopic video) viewing medium (M = 2.58, SD = 1.87), t (178) = 12.80, p < .001. *Table 2* presents additional information regarding the t-test results, including Levine's test for equality of variance as well as differences between the means and standard errors.

Table 2: Independent Samples t-test

	Levine's Test for Equality of Variance		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95 % Confidence Interval of the Difference	
								Lower	Upper
Time	10.97	.001	-9.30	178	.000	-21.68	2.33	-26.29	-17.09
Accuracy	7.76	.006	12.80	178	.000	4.08	.318	3.45	4.71

DISCUSSION

In this study of 180 naïve participants, there was a significant difference in performance of a telerobotic manipulation across viewing medium conditions. That is to say that those operators completing the manipulation task under the direct viewing condition performed better than those under the indirect viewing condition. The magnitude of the differences between the means for each dependent measure suggests that telerobotic operators rely heavily on the stereoscopic cues that are available in binocular vision. This also suggests that viewing medium should be considered an extremely relevant factor for operators when performing telerobotic tasks.

These results support the research hypothesis that operators will perform significantly better in a direct (stereoscopic) viewing condition due to the additional binocular depth cue advantages afforded to them, and is consistent with previous research that reports advantages of stereoscopic viewing over monoscopic viewing (Barfield & Rosenberg, 1995; McLean, Prescott, & Podhorodeski, 1994; Yeh & Silverstein, 1992). Stereoscopic viewing increases a human's awareness of the spatial relationship between objects in an environment by increasing the amount of depth information relayed from the environment. The results of this study demonstrate how the increase in the awareness of depth information translates directly into better human performance of a remote telerobotic manipulation task for distances less than 250 cm.

In unstructured environments such as space, undersea, or remote bomb disposal, the risk of damage to a stereoscopic video system exists. Damage to the stereoscopic video system may leave a telerobotic operator in a monoscopic viewing condition, and therefore with severely degraded depth information about the remote environment. The results of this study suggest a need to train telerobotic operators in the differences they may expect as a result of reduced depth information when operating in a three-dimensional remote environment with information that is visually displayed in two-dimensions. If time and accuracy are critical factors in the remote action being performed, it will be essential that teleoperators understand how those factors will be effected as a result of the differing viewing mediums. When performing a remote telerobotic manipulation task attempting to reconstruct 3D information from 2D displays, operators should expect an increase in the time needed to make accurate placements, and decrease in the accuracy of the manipulation tasks.

REFERENCES

Barfield, W., and Rosenberg, C. (1995). Judgments of Azimuth and Elevation as a Function of Monoscopic and Binocular Depth Cues Using a Perspective Display. Human Factors, 37*(1)*, 173-181.

Bejczy, A.J. (1976). Performance evaluation of computer aided manipulator control. IEEE International Conference on Systems, Man and Cybernetics – *1976*.

Bounds, S.R., Schroer, S. E., and Schroer, B.J. (1990). Experimentation with telerobotics for use in space systems. Journal of Intelligent and Robotic Systems, 3, 365-378.

Chapanis, A. (1996). Human Factors in Systems Engineering. New York: John Wiley & Sons.

Coren, S., Ward, L.M., & Enns, J.T. (1999). Sensation and Perception (5th ed., pp. 253-267). Fort Worth: Harcourt College Publishers.

Drascic, D.(1991). Skill acquisition and task performance in teleoperation using monoscopic and stereoscopic video remote viewing. Proceedings of the Human Factors Society 35th Annual Meeting – 1991. 1367-1371.

Drascic, D., Milgrim, P., & Grodski, J. (1989). Learning effects in telemanipulation with monoscopic versus stereoscopic remote viewing. IEEE International Conference on Systems, Man and Cybernetics, 3, 1244-1249.

Ellis, S.R., McGreevy, M.W., & Hitchcock, R.J. (1987). Perspective traffic display format and airline pilot traffic avoidance. Human Factors, 29, 371-382.

Haskell, I.D., & Wickens, C.D. (1993). A theoretical and empirical comparison of two- and three- dimensional displays for aviation. International Journal of Aviation Psychology, *3*, 87-109.

Massimino, M.J., & Sheridan, T.B. (1994). Teleoperator performance with varying force and visual feedback. Human Factors, 36*(1)*, 145-157.

McLean, G. F., Prescott, B., and Podhorodeski, R. P. (1994). Teleoperated System PerformanceEvaluation. IEEE Transactions on Systems, Man, and Cybernetics, 24*(5)*, 796 - 804.

McGovern, D.E. (1991). Experience and results in teleoperation of land vehicles. In S.R. Ellis, M. Kaiser, & J. Grunwald (Eds.) Pictorial communication in virtual and real environments (pp. 182-195). London: Taylor & Francis.

McGreevy, M., & Ellis, S.(1986). The Effect of Perspective Geometry on Judged Direction in Spatial Information Instruments. Human Factors, 28, 439-456.

Lumelsky, V. (1991). On human performance in telerobotics. IEEE Transactions on Systems, Man & Cybernetics, 21, 971-981.

Kim, W.S., Ellis, S.R., Tyler, M.E., Hannaford, B., & Stark, L.W. (1987). Quantitative evaluation of perspective and stereoscopic displays in three-axis manual tracking tasks. IEEE Transactions on Systems, Man & Cybernetics, 17, 61-72.

Kim, W.S., Tendick, F., & Stark, L. (1991). Visual enhancements in pick-and-place tasks: Human operators controlling a simulated cylindrical manipulator. In S.R. Ellis, M. Kaiser, & J. Grunwald (Eds.) Pictorial communication in virtual and real environments (pp. 265-282). London: Taylor & Francis.

Park, S. & Woldstad, J. (2000). Multiple Two-Dimensional Displays as an Alternative to Three-Dimensional Displays in Telerobotic Tasks. Human Factors 42, 592-603.

Pepper, R.L., & Hightower, J.D. (1984). Research Issues in Teleoperator Systems. Proceedings of the Human Factors Society 28th Annual Meeting – 1984, 803-807.

Pepper, R.L., Smith, D.C., Cole, R.E. (1981). Stereo TV improves operator performance under degraded visibility conditions. Optical Engineering, 20, 579-585.

Sheridan, T.B. (1992). Telerobotics, automation, and human supervisory control. Cambridge, MA: The MIT Press.

Sollenberger, R.J., & Milgrim, P. (1993). Effects of stereoscopic and rotational displays n a three-dimensional path-tracing task. Human Factors, 35, 483-501.

Horikawa, Y., & Nagatomo, M. (1998). On the results of the manipulator flight demonstration for the JEM. 49th International Astronautical Congress – 1998.

Wickens, C.D., Merwin, D., & Lin, E. (1994). Implications of graphics enhancements for the visualization of scientific data: Dimensional integrality, stereopsis, motion, and mesh. Human Factors, 36, 44-61.

Yeh, Y. Y., and Silverstein, L. D. (1992). Spatial Judgments with Monoscopic and Stereoscopic Presentation of Perspective Displays. Human Factors, 34, 583-600.

HUMAN PERFORMANCE IN EXTREME ENVIRONMENTS: FROM THE BATTLEFIELD TO THE FINAL FRONTIER

Jason P. Kring (Panel Chair), Haydee M. Cuevas
University of Central Florida

John Barnett (Panel Co-Chair), Donald Lampton
U.S. Army Research Institute for the Behavioral and Social Sciences

James F. Fletcher
Embry-Riddle Aeronautical University

ABSTRACT

Humans operate in an increasingly diverse assortment of extreme environments. From deep sea divers supporting offshore drilling operations, military exercises in the desert, and space crews aboard the International Space Station (ISS), personnel must perform under physically and psychologically challenging conditions. Humans, however, are not naturally suited to endure such environments and are therefore reliant on technology and training for safety, mission success, and in many cases, survival. The goals of this panel discussion are to 1) expose the unique challenges of performing in extreme environments, 2) uncover valuable similarities between seemingly different environments, and 3) present how lessons learned in one environment can be applied to others to improve human performance. Panelists address these goals with discussions on specific topics including combat aviation (Barnett), stress in extreme environments (Cuevas), barometric pressure changes and the human body (Fletcher), and military operations (Lampton).

Keywords: Extreme environments, Aviation, Stress, High-altitude; Military operations

INTRODUCTION

A major thrust of human performance research is to understand how humans adapt, endure, and succeed in settings that possess extraordinary physical and psychological stressors. Suedfeld (1987) labeled these settings "extreme and unusual environments" and described four major categories of extreme environments (EEs). Normal environments are standard for a particular group but are considered extreme because of "...physical or resource availability characteristics that militate against comfort and survival." (p. 865). Suedfeld lists situations exhibiting high social density or stimulus input that can be damaging, as well as unique situations like prison; however, he argues these settings may not fully qualify as extreme. Instrumental environments are entered voluntarily by individuals or groups for a specific purpose. In most cases, the individual is "selected, trained and equipped" to achieve a goal and typically "...share a value system that considers the goal to be worth reaching despite discomfort and danger" (p. 865). Accordingly, human missions to Mars, winter-over expeditions at the Earth's poles, or seclusion in underwater research habitats constitute instrumental EEs. A third category is the recreational environment which is entered voluntarily to achieve some personal goal or experience novel settings or events. Extreme sporting activities, such as mountaineering, cave diving, or solo dog-sledding, fall in this category. Suedfeld notes, however, the line between instrumental and recreational EEs can change with unforeseen events. A recreational hike in Rocky Mountain National Park can quickly become an instrumental EE if a mild Spring day turns into a blizzard. The last category, traumatic environments, captures extreme conditions imposed on individuals unwillingly. Suedfeld makes a distinction between "natural" traumatic EEs like natural disasters, and "man-made" events such as explosions, industrial accidents, and some medical emergency events, and combat situations.

Suedfeld (1987) further defined EEs by characterizing physical, interactive, and psychological parameters that are present in many EEs. Outlined in Table 1, Suedfeld argued normal, instrumental, recreational and traumatic EEs may possess some or all of following features.

Table 1. *Features of Extreme Environments*

Parameter	Features of Environment
Physical	• Survival impossible without advanced technology, but may serve as natural habitat for some human groups (e.g., the Artic, high mountains, deserts) • Highly hazardous • Inhabited only on exploratory or experimental bases (e.g., outer space, ocean floor) • Environments during and immediately after drastic disruption of normal attributes that involve high degree of danger and major alteration in physical characteristics (e.g., safe environments transformed by earthquake, hurricane, or battle).
Interactive	Factors related to person-environment interactions including: • Availability of information • Ease of communication within and outside environment • Mobility or physical restriction • Environment complexity • Status implications of being in environment • Degree of isolation from other members of one's group and from other groups • Whether individual is there voluntarily • Actual and expected duration • Control • Predictability • Privacy and territorial integrity • Extent to which environment pervades an individual's life
Psychological	Factors related to how an individual perceives and copes with environment, rather than environment itself, including: • How individuals perceive themselves • Degree of preparation • Training • Fitness • Personality characteristics • Affective interactions • Group and individual morale • Motivation • Cohesiveness and group structure • Leadership

Note. Descriptions from Suedfeld (1987), pp. 864-865.

This categorization serves as a framework for the present discussion on human performance in EEs. Additional conceptualizations by Manzey and Lorenz (1999), Morphew (1999), and Suedfeld and Steel (2000) offer slightly different interpretations of EEs but can be summarized into one succinct definition as *settings that possess extraordinary technological, social, and physical components that require significant human adaptation for successful interaction and performance* (Barnett & Kring, 2003).

Given this definition, a wide variety of occupations and activities can be labeled "extreme" including those faced by deep sea divers, firefighters, astronauts, and military personnel on the ground and in the air. On the surface, activities in these EEs are seemingly different, for example, when comparing the activities of a firefighter and an astronaut on a long-duration mission. However, at a deeper level, EEs share several common features that suggest findings from one domain may have relevance to efforts to understand human performance in other extreme domains.

Toward this end, panelists will endeavor to 1) expose the unique features and challenges of performing in EEs, 2) discuss valuable similarities between seemingly different EEs, and 3) present how lessons learned in one environment can be applied to others to improve human performance and safety. As summarized below, the discussion will begin with a general overview of EEs and common features within. Then, panelists will address specific examples and aspects of performing in extreme settings. First is a description of the stressors faced when challenging environments converge as in the case of aviation in a combat context. Next is a discussion of a transactional approach to investigating the effects of stress in EEs. The final two panelists address the challenges associated with performing at high-altitudes and during military operations in urban combat, respectively.

Combat Aviation – John Barnett

The human performance challenges posed by extreme environments can be exacerbated when such environments interact. Such is the case with combat aviation, which has all the challenges of commercial and private flying, but includes the additional responsibility of conducting a military mission. This section of the panel discusses how environmental variables interact with psychological factors to affect human performance in combat aviation in general, and also how the major military aviation missions, fighter, bomber, transport, rotary-wing, and special mission aircraft, have some differences in environmental elements.

In commercial aviation, safety is the principle concern, whereas in combat flying, safety and mission accomplishment have equal priority. For this reason, risk is higher with combat aviation, with a corresponding increase in performance stress and fear. In addition, considering the speeds of even large combat aircraft, events tend to happen quickly. For example, a bombing run against a defended ground target may last 90 seconds or less; whereas in fighter gun combat, the proverbial "dogfight" or "furball," the target may present itself for only a few seconds. This fast pace places considerable time pressure stress on aircrew members.

In addition to these common environmental stressors, each type of aircraft often has unique stressors due to its specific combat mission. For example, long-range aircraft, such as bombers, tankers, and transports, may add boredom and fatigue to the list. Conversely, fighter/attack aircraft often engage in high-G maneuvers not practiced by larger aircraft. The following addresses the missions and special environmental factors of different types of aircraft.

- Fighter/Attack
  - *Missions.* The typical missions of fighter/attack aircraft include Defensive Counter-Air, (air-to-air missions), Escort (protective escort for other aircraft) Close Air Support (bombing in close proximity to friendly ground troops) and Interdiction (bombing enemy ground forces).
  - *Specific environmental factors.* These include complex maneuvers which often result in high G-loading on the pilot and tend to be three-dimensional in nature. The complexity of such maneuvers increases the probability of spatial disorientation.
- Bomber
  - *Missions.* Interdiction and Strategic Attack (bombing deep inside enemy defenses). Recently heavy bombers such as B-52s and B-1s have included Close Air Support to their repertoire.
  - *Specific environmental factors.* Factors associated with long range flying include boredom and fatigue, which tends to reduce situation awareness. Dehydration is also a common problem due to extended exposure to very dry air associated with most aircraft pressurization systems. The boredom of long-range flight is generally interrupted by a brief, high-stress dash through a defended target area.
- Tanker/Transport
  - *Missions.* Long- and short-range air drop/cargo transport, and air refueling.
  - *Specific environmental factors.* These aircraft have the same long-endurance flying factors as bombers. In addition, they may begin or end their missions on airfields with minimal facilities and doubtful security.
- Rotary wing aircraft.
  - *Missions.* Typically reconnaissance/scouting, transport, ground attack, or rescue.
  - *Specific environmental factors.* Most combat helicopters traverse enemy territory at very low altitudes, which increases the complexity of the pilot's task of navigating while avoiding obstacles at relatively high speeds, thus intensifying performance stress. Helicopters also tend to have more intense vibration than fixed-wing aircraft.
- Special mission aircraft

- *Missions.* Long-range reconnaissance, command and control, airborne radar, and Special Operations missions, among others.
- *Specific environmental factors.* Many of these aircraft fly long-endurance missions, like bombers, tankers, and transports. They are often considered High-Value Air Assets (HVAA) and consequently are priority targets for enemy fighters and air defenses.

A Transactional Approach to Investigating Stressor Effects in Extreme Environments – Haydee M. Cuevas

To optimize human performance in complex operational environments, it is critical that researchers explore the underlying mechanisms by which psychological, physiological and/or environmental stressors may negatively impact the human operator (Manzey & Lorenz, 1999; Suedfeld, 2001). Toward this end, adopting a *transactional* approach to investigating stressor effects may lead to a greater understanding of the complex processes by which humans adapt psychologically and physically to the adverse conditions encountered in extreme environments (e.g., aerospace, arctic exploration, military combat).

Transactional approaches conceptualize stress as occurring in the nature of the "transaction" (i.e., interaction) between the individual and the stimulus environment, emphasizing the role of cognitive appraisal (i.e., perceived ability to cope with the situation) (Lazarus & Folkman, 1984; Stokes & Kite, 1994). Specifically, the transactional model highlights how the stress response is influenced by the degree to which one perceives (i.e., appraises) an event as threatening and/or perceives (i.e., appraises) one's ability to cope with the threat (i.e., resources available) as insufficient (Baum, Singer, & Baum, 1981; Lazarus & Folkman, 1984) Further, individual differences in operator characteristics (e.g., personality traits, coping strategies) may differentially impact one's perception and subsequent response to a potentially stressful event (e.g., Bowers, Weaver, & Morgan, 1996; Carver, Scheier, & Weintraub, 1989; Cox & Ferguson, 1991). Therefore, interventions are clearly warranted to positively influence this cognitive appraisal process and promote successful human performance under stress in extreme environments. Strategies can be targeted at either: (1) *fitting the individual to the task* through personnel selection (e.g., Hogan & Lesser, 1996; Suedfeld, 2001) and training (e.g., Driskell & Johnston, 1998); or (2) *fitting the task to the individual* via psychosocial support mechanisms (e.g., Holland, 2000; Manzey & Lorenz, 1999) and application of human factors design principles (e.g., Albery & Woolford, 1997; Wickens, 2000). Ultimately, the goal is to ensure that operators perceive a strong sense of control over their response in any challenging situation.

Barometric Changes and the Human Body – James F. Fletcher

Throughout time, various concerns have been levied on the effects of changing barometric pressure on the human body. As humans, we have subjected ourselves to various environmental extremes. Some of these environments have been utterly devastating while others have had no effect.

Early documentation, as early as 1519, when Cortez and his armies attacked Mexico, or 25 years later when Pizarro attacked Quito, Peru only to lose thousands of Spaniards, Indians and horses have been exposed to the ravages of altitude induced illness. The Jesuit Father Jose de Acosta noted after five crossings of the Andes Mountains "Not only men feel this, animals do too, and that sometimes stop and no spur can make them advance." After 1900, more vigorous investigations of human exposure to hyperbaric (compressed air) environments related to expanding caisson workers and air diving led to progressive understanding of decompression sickness.

The advent of technology has not changed the human condition. The equipment has advanced and so too has the behavioral/physical conditioning, but the physiology of the human body has remained the same. Bubbles continue to generate when exposed to the reduction of barometric pressure and inversely, gasses continue to compress into solution when the barometric pressures increase. This discussion will concern the exposure of the human condition to the extreme environments of undersea, mountaineering, high-performance flight, and space flight.

Military Operations – Donald Lampton

This section of the panel will describe the human performance challenges of military operations, with a particular focus on training to deal with the unique stressors associated with urban combat. In common with most other extreme environments, human performance in combat is a function of many factors, including personnel, equipment, organization, doctrine, and training. However, overshadowing both the cognitive and physical demands is the unique nature of combat itself, described by the military historian N. T. Dupuy (1987) as the constant danger of death from lethal weapons employed by opponents with deadly intent.

Modern urban combat in particular presents a very extreme environment that necessitates the execution of highly developed cognitive, physical, and social skills with little margin for error. The various factors that contribute to the difficulty of urban combat, and the corresponding performance challenges for small teams will be outlined. The focus will be on cognitive performance aspects such as rapid decision making under stress, command and control, and acquiring and maintaining situation awareness. Current and developing approaches to measuring and training situation awareness for the small unit team and team leader will be described. In addition, a new approach to the analysis of verbal communications will be presented.

CONCLUSION

The ultimate goal of this panel is to open the door to increased scientific collaboration. It is our hope that stimulating a dialogue on human performance in EEs will encourage other researchers and applied personnel to share experiences and empirical results and promote a common understanding of features shared by EEs.

Panelists also discuss ways to enhance human performance research and applications. We argue that operators, engineers, managers, and scientists from many distinct disciplines must work collectively to define principal theoretical and empirical issues and formulate viable solutions to performance decrements in extreme settings. Like the emergence of human factors psychology, which bridged the gap between engineering and the behavioral sciences, there is a need to facilitate communication between once solitary scientific fields and disciplines, to promote the sharing of ideas and information, and to bring together academics with practitioners in applied settings. This unified effort is an essential step in sustaining and enhancing performance in all extreme environments.

REFERENCES

Albery, W. B. & Woolford, B. (1997). Design for macrogravity and microgravity environments. In G. Salvendy (Ed.), Handbook of human factors and ergonomics (pp. 935-963). New York, NY: John Wiley & Sons.

Barnett, J., & Kring, J. P. (2003, October). Human performance in extreme environments: A preliminary taxonomy of shared factors. Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting (pp. 961-964). Santa Monica, CA: Human Factors and Ergonomics Society.

Baum, A., Singer, J. E., & Baum, C. S. (1981). Stress and the environment. Journal of Social Issues, 37, 4-35.

Bowers, C. A., Weaver, J. L., & Morgan, B. B. (1996). Moderating the performance effects of stressors. In J. E. Driskell & E. Salas (Eds.), Stress and human performance (pp. 163-192). Mahwah, NJ: LEA.

Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing coping strategies: A theoretically based approach. Journal of Personality & Social Psychology, 56, 267-283.

Cox, T. & Ferguson, E. (1991). Individual differences, stress, and coping. In C. L. Cooper & R. Payne (Eds.), Personality and stress: Individual differences in the stress process (7-30). Chichester, England: John Wiley & Sons.

Driskell, J. E. & Johnston, J. H. (1998). Stress exposure training. In J. A. Cannon-Bowers & E. Salas (Eds.), Making decisions under stress: Implications for individual and team training (pp. 191-217). Washington, DC: APA.

Dupuy, N. T. (1987). Understanding war: History and theory of combat. Falls Chruch, VA: NOVA Publications.

Hogan, J. & Lesser, M. (1996). Selection of personnel for hazardous performance. In J. E. Driskell & E. Salas (Eds.), Stress and human performance (pp. 195-222). Mahwah, NJ: LEA.

Holland, A. W. (2000). Psychology of spaceflight. Journal of Human Performance in Extreme Environments, 5 (1), 4-20.

Lazarus, R. S. & Folkman, S (1984). Stress, appraisal, and coping. New York, NY: Springer.

Manzey, D. & Lorenz, B. (1999). Human performance during spaceflight. Journal of Human Performance in Extreme Environments, *4*, 8-13, [Article reprinted from Proceedings of the 9th International Symposium on Aviation Psychology, Columbus, OH, 1997].

Morphew, M. E. (1999). The challenges of long-duration spaceflight to astronaut safety, performance, and mission success. Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting (pp. 56-60). Santa Monica, CA: Human Factors and Ergonomics Society.

Stokes, A. & Kite, K. (1994). Flight stress: Stress, fatigue, and performance in aviation. Aldershot, England: Avebury Aviation.

Suedfeld, P. (1987). Extreme and unusual environments. In D. Stokols & I. Altman (Eds.), Handbook of environmental psychology (pp. 863-885). New York; Wiley & Sons.

Suedfeld, P. (2001). Applying positive psychology in the study of extreme environments. Journal of Human Performance in Extreme Environments, 6 (1), 21-25.

Suedfeld, P., & Steel, G. D. (2000). The environmental psychology of capsule habitats. Annual Review of Psychology, 51, 227-253.

Wickens, C. D. (2000). Designing for stress. Journal of Human Performance in Extreme Environments, 5 (1), 98-106.

SUPPORTING DRIVERS' COGNITIVE MAP CONSTRUCTION WITH VISUAL GEO-CENTERED AND AUDITORY EGO-CENTERED GUIDANCE: INTERFERENCE OR IMPROVED PERFORMANCE?

Hiroshi Furukawa
University of Tsukuba, Japan

Carryl L. Baldwin and Ellen M. Carpenter
Old Dominion University

ABSTRACT

Commercially available in-vehicle routing and navigational systems (IRANS) present a generic form of route guidance information to all users. However, a growing body of literature suggests that drivers differ in their navigational strategies and abilities. The current investigation was designed to examine the impact of IRANS display modality on drivers' ability to navigate through and form cognitive maps of unfamiliar areas as a function of drivers' self-reported navigational strategy and ability. Drivers were required to navigate through unfamiliar areas along specified routes in a high-fidelity driving simulator using an ego-centered auditory route guidance system (ARGS), a geo-centered visual-map guidance system (VMGS) or both the ARGS and the VMGS. Drivers in general reported lower subjective ratings of workload when using the ARGS either by itself or in combination with the VMGS. However, drivers reporting a high degree of awareness of cardinal orientation and a tendency to use survey style navigational strategy benefited from use of the VMGS, relative to both the ARGS and the ARGS in combination with the VMGS. The current results warrant further investigation of the influence of individual differences in order to design appropriate navigational aids for supporting drivers of all types.

Keywords: Navigational aids; Area-learning task; Survey map; Driving simulator

INTRODUCTION

Invehicle routing and navigational systems (IRANS) are one of the many important types of in-vehicle technologies (IVTs) found in the modern automobile. IRANS potential advantages for the driver include ease in finding destinations, avoidance of traffic congestion and delays, shorter travel routes, fewer instances of disorientation or getting lost, shorter duration routes, greater confidence, and less stressful driving experiences (Eby & Kostyniuk, 1999). Despite these many advantages, IRANS have the potential to increase the attentional processing requirements or mental workload of the driving task. Due to the potential for IRANS to increase mental workload, the most effective system is one that assists the driver in establishing a cognitive map of the route to be taken through an unfamiliar area in the most effective way. Developing an internal cognitive map of the route to be taken decreases the information processing requirements of obtaining navigational information and ultimately decreases reliance on the system in the shortest amount of time.

Currently available systems can be categorized by key distinguishing factors including display modality, and geo-versus ego-centered display orientations. Display modality refers to whether navigational information is presented through visual, auditory or both visual and auditory channels. The second key distinguishing characteristics is whether navigational information is presented in a geo-centered orientation (north-up) or ego-centered (driver-forward view) orientation.

In addition to these key design characteristics, a growing body of literature suggests that drivers differ in their preference for and utilization of differing types of navigational information (Baldwin & Reiss, 2000; Carpenter, Baldwin, & Furukawa, in press; Lawton, 1994, 1996; Takeuchi, 1992; Thorndyke & Hayes-Roth, 1982). Constructs used to identify individual differences in drivers' navigational styles and abilities appear to remain stable across geographical location and cultural ethnicity (Carpenter et al., in press; Lawton, 2001). Important constructs include preference for a route (point by point) versus survey (global overview), use and memory for landmarks and general awareness of orientation. Current IRANS typically combine auditory "route" style navigational instructions with a visual map presenting an overview or "survey" of the area. Drivers' ability to utilize navigational information from different guidance systems may therefore depend on drivers' navigational strategy preferences as much as the modality used for presenting the information.

The aim of the current investigation was to examine the influence of individual differences in drivers' navigational style and ability on their ability to navigate through and form cognitive maps of unfamiliar areas using IRANS displays of differing types. Specifically, drivers' preferred navigational style and overall navigational abilities were assessed and their ability to develop a cognitive map after driving through an unfamiliar area using one of three styles of navigational aids was examined. It was predicted that drivers who relied on a route-style navigational strategy would benefit most (construct a more accurate cognitive map) when using an ego-centered ARGS, relative to a VMGS. Conversely, it was predicted that drivers who reported preference for survey strategy navigational information would demonstrate better cognitive map formation when using a geo-centered VMGS. Drivers' navigational performance in general was expected to follow these same trends.

METHOD

The current investigation was designed to examine the relative influence of existing navigational formats, specifically ego-centered auditory route style navigational instructions and visual maps presenting a geo-centered survey of the driving area on cognitive map formation.

Participants

Twenty female and fourteen male university students (thirty-six in total) whose ages ranged from 19 to 42 years (mean 23.7) voluntarily participated in this experiment. All participants reported that they drove the car almost everyday and had normal or corrected to normal vision and hearing.

Equipment and Materials

A high fidelity driving simulator (Capital I-Sim Driving Simulator, made by General Electric) was used to examine the efficacy of navigational aids for the navigational task as well as area-learning task. The simulator consists of three 40-inch screens, capable of presenting a 180-degree driver's front view. Participants controlled the simulated car using a steering wheel, accelerator and a brake pedal.

Routes. Two intersecting routes were constructed for each of the three urban areas. Each route had two turns and crossed each other at three intersections. A salient landmark was present on or near each of the intersections, such as a parked panel truck, a construction sign, tall trees, a fire engine, and a group of people. Participants were familiarized with the specific landmarks to be encountered in each route prior to beginning the route-learning task. The three urban areas represented different parts of the city with no overlap between the areas.

IRANS format. Three formats of navigational aids were implemented. One consisted of visual only (VMGS), a second consisted of auditory only (ARGS) and the third consisted of concurrent presentation of both VMGS and ARGS.

VMGS. The geo-centered visual-map guidance system (VMGS) format consisted of a visual map displayed on a liquid crystal display that was set up in the dashboard area on the right-hand side of the drivers' seat just below the simulated front windshield. The display location required participants to move their heads to the lower right to see the map (a typical display location for actual IRANS). The navigational map was drawn using geo-centered, north up coordination. Previous research has provided initial evidence that geo-centered maps may be more effective than ego-centered maps in facilitating cognitive map construction during navigational tasks (Azekura, 2003). The driver's location while traveling through the route was presented on the moving map display in real-time.

ARGS. The ego-centered auditory route guidance system (ARGS) format was presented via the existent audio system of the simulator. Terse auditory commands were recorded from a native English-speaking female speaking at a normal conversational level of approximately 65 dB and then digitized. Commands consisting of, "Turn left," "Turn right," or "Continue forward" were presented at each intersection to guide participants along the specified route. Auditory commands were always presented in an ego-centered (driver front view) perspective.

Procedure

Participants completed two navigation-related questionnaires (obtained from Takeuchi, 1992 and Lawton, 1994). The former assesses three types of self-evaluated perceived ability in space cognition, which are ability of using maps, memory for visual landmarks, and awareness of orientation (modified classification based on results from an independent factor-analysis using data reported in Carpenter et al., in press). The latter depicted self-reported preferred strategies for wayfinding tasks in normal life: route strategy and survey strategy (see Carpenter et al., in press for further description).

The navigational aid format was counterbalanced across areas. In each area, participants drove a simulated vehicle along the two predetermined routes using VMGS, ARGS, or both VMGS and ARGS guidance. Participants were instructed to watch for specific landmarks along the route and then were asked a series of questions designed to ascertain the accuracy and breadth of their cognitive map construction. The questions pertained to cardinal relationships between a landmark and a starting or an ending point. There were six questions for each route. For example, "The tall trees are to the _____ of the starting point," where the alternatives were North, South, East, West, NE, NW, SE, and SW. The score for exact answers was 2, and 1 was assigned to answers deviating by 45 degrees, e.g., answers of "NE" or "NW" for the correct answer "North." Participants answered the questions for each route immediately after driving through it. Following the questions pertaining to the second route in each area, they answered the same type of queries about the overall area in which they had driven. There were six questions pertaining to each overall area. Following completion of all area questions participants completed the NASA-TLX as a subjective index of mental workload for the navigational task using each type of navigational aid, not the difficulty of the questions that followed each route.

RESULTS

Grouping with Sense of Direction and Wayfinding Strategies

To examine the relationships between the efficacy of each type of navigational aid and individual differences in space cognition ability and wayfinding strategies, the participants were classified into two groups based on the results of the questionnaires. In Extreme Grouping, "Lower" is a group of participants whose total points are less than "mean – standard deviation (SD)," and "Higher" is a group of participants whose total points are greater than "mean + SD." In Coarse Grouping, the threshold for "Lower" is "mean – 1/2 SD" and "mean + 1/2 SD" for "Higher." Table 1 shows the number of the participants in each group.

Accuracy of Cognitive Map Knowledge

A repeated measures ANOVA was performed to examine participants accuracy for the questions pertaining to route and area as a function of the navigational aid used. The ANOVA test for all the participants revealed no significant differences among the types of navigational aids with respect to accuracy of overall cognitive map knowledge, nor among the types of aids for either the cognitive map construction task of local area routes or total area.

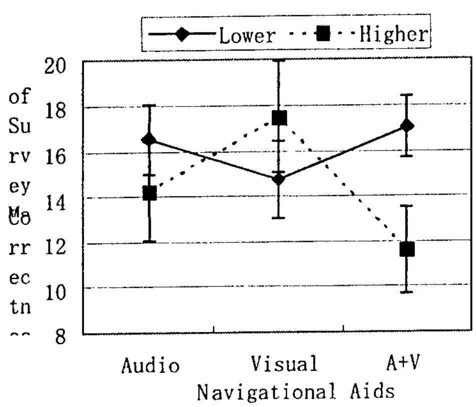


Figure 1: Individual differences as a function of navigational aid and awareness of orientation

Table 1. The number of participants in each group classified by their ability in space cognition or wayfinding strategies.

Grouping	"Extreme"		"Coarse"	
Groups	Lower	Higher	Lower	Higher
Maps	6	7	10	10
Landmarks	3	6	11	11
Orientation	6	7	16	8
Route Strategy	6	3	10	11
Survey Strategy	4	7	15	10

Individual Differences. A significant interaction was observed between the types of navigational aids and the ability in awareness of orientation with the Coarse grouping (p=0.025*). Figure 1 shows the means and standard errors in the conditions. The results of pairwise comparison (Bonferroni) shows that participants with lower ability in awareness of orientation answered more questions correctly than those with higher ability when they were using the VMGS and the ARGS (p=0.030*).

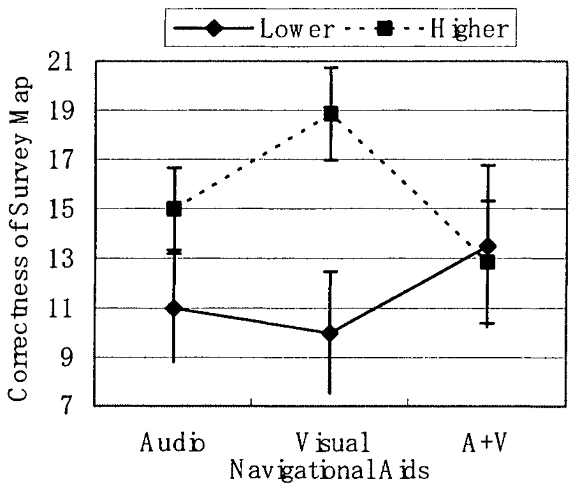


Figure 2: Individual differences as a function of navigational aid and survey strategy (Coarse Grouping)

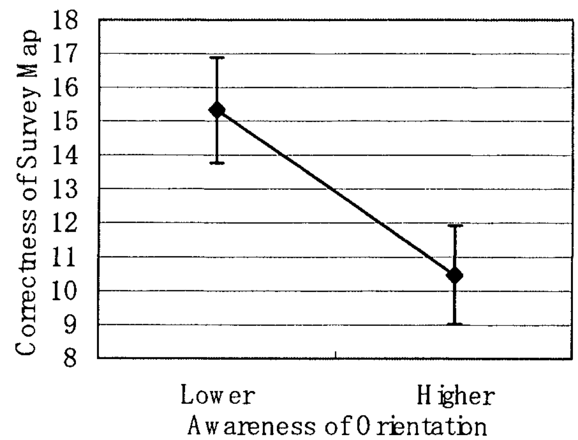


Figure 3. Individual differences as a function of awareness of orientation

In the task of construction of local area knowledge, a strong nonsignificant trend was observed for the interaction between the type of navigational aid and the preference of survey strategy regardless of type of grouping (p=0.053 at Extreme and p=0.056 at Coarse). The results of pairwise comparison shows that participants preferring a survey strategy generally tended to answer more questions correctly relative to those not preferring a survey strategy when they were using only the geo-centered map regardless of type of grouping (p=0.019* at Extreme, Figure 2, and p=0.011* at Coarse). With the Extreme grouping, the participants with lower ability in awareness of orientation answered more questions correctly than those with higher ability (p=0.044*). This finding is counterintuitive, as we would expect that people with higher ability in awareness of orientation would be better at construction of cognitive maps than people with lower ability. The means and standard errors are depicted in Figure 3.

Mental Workload

Participants' workload was significantly lower when using the ARGS or the VMGS in combination with the ARGS relative to the VMGS only (p=0.006** and p=0.000**, respectively).

DISCUSSION

The results indicate that use of the auditory ego-centered information may support drivers' navigation without harm to cognitive map development with two important exceptions. Individual differences in performance were observed as a function of navigational aid and navigational strategy as assessed by Takeuchi's and Lawton's questionnaires.

Individual Differences Related to Aids and Awareness of Orientation

With the coarse grouping, participants with lower ability in awareness of orientation answered more cognitive map assessment questions correctly relative to those with higher ability when they were presented with the VMGS and ARGS concurrently. People with high ability in awareness of orientation benefited most from the VMGS format only and suffered performance decrements when presented with the concurrent ARGS aid. This result suggests that the auditory ego-centered information may interfere with the process of construction of cognitive map knowledge by people with higher ability in awareness of orientation. There may have been a resource cost associated with trying to ignore the auditory information. If persons with lower ability in awareness of orientation were relying on the auditory aid, it would likely have been easier for them to disregard the visual display.

The results on NASA-TLX indicated that participants' workload for the navigational task was significantly lower when they were able to use the auditory guidance aid either by itself or in combination with the visual map. The auditory ego-centered information may be particularly useful for people with lower ability in awareness of orientation. The geo-centered visual map required participants to reference their driver front view to the north up map, a task that people lower in awareness of orientation have difficulty with. People with higher ability in awareness of orientation appear to be able to use the auditory aid to navigate the route, however the auditory ego-centered information appears to harm their ability to form a cognitive map

Individual Differences Related to Aids and Survey Strategy

In the task of construction of a cognitive map during the navigation task, participants preferring a survey strategy tended to perform better when they were using the geo-centered map only, relative to those not preferring a survey strategy. However, there was no significant difference between the two groups under the condition with the concurrent geo-centered map and auditory ego-centered aid. This result may indicate that, similar to persons with a high ability in awareness of orientation, the auditory ego-centered information disrupts cognitive map construction for people preferring a survey strategy. However as previously stated, the ability to use the auditory ego-centered information appears helpful in performing the navigation task and, at least for people who do not prefer a survey strategy, the auditory information does not appear to disrupt cognitive map construction.

Individual Differences Related to Awareness of Orientation

With the extreme grouping, participants with lower ability in awareness of orientation correctly answered more of the cognitive map assessment questions than those with higher ability. There are two at least two possible explanations for this result. It is possible that people with extremely high ability in awareness of orientation may actually just have lower ability in cognitive map construction. A more plausible explanation is that participants reporting a higher ability in awareness of orientation may have an over-reliance on their ability. The results of the NASA-TLX indicated that there were no significant differences between participants with the higher and lower ability on perceived mental workload in performing the navigational task. This finding, along with the lower cognitive map assessment performance, lends support to the possibility that persons scoring higher in awareness of orientation may have a tendency toward over-reliance on their ability. Further research is needed to examine this issue.

CONCLUSION

This experimental study emphasizes the importance of considering individual differences in navigational strategy and ability when designing in-vehicle routing and navigational systems. Ego-centered auditory aids may mitigate the difficulties in navigation with geo-centered maps without cognitive resource interference for some drivers. However, for other drivers the auditory aids may present an additional source of distraction that does not affect the navigation task directly, but rather interrupts the formation of a cognitive map of the area being navigated. Further research and analysis on the cognitive processes involved in the navigational task and area-learning task are necessary to identify appropriate navigational aids for supporting both tasks simultaneously.

REFERENCES

Azekura, M. (2003). Difference in drivers' space cognition from map coordinate system during navigational tasks. Unpublished manuscript, University of Tsukuba, Tsukuba, Japan.

Baldwin, C. L., & Reiss, R. (2000, October 2000). Preferred Navigational Style Assessment as a Method of Optimizing IRANS Displays. Paper presented at the Fourth Conference on Automation Technology and Human Performance and the Third Conference on Situation Awareness in Complex Systems, Savannah, Georgia.

Carpenter, E., Baldwin, C. L., & Furukawa, H. (in press, March 2004). Cross-cultural analysis of navigational strategy: further implications for IRANS design. Paper presented at the Proceedings of the Human Performance, Situation Awareness, and Automation Conference, Daytona Beach, FL.

Eby, D. W., & Kostyniuk, L. P. (1999). An on-the-road comparison of in-vehicle navigation assistance systems. Human Factors, 41(2), 295-311.

Lawton, C. A. (1994). Gender differences in way-finding strategies: Relationship to spatial ability and spatial anxiety. Sex Roles, 30(11-12), 765-779.

Lawton, C. A. (1996). Strategies for indoor wayfinding: The role of orientation. Journal of Environmental Psychology, 16(2), 137-145.

Lawton, C. A. (2001). Gender and regional differences in spatial referents used in direction giving. Sex Roles, 44(5-6), 321-337.

Takeuchi, Y. (1992). Sense of direction and its relationship with geographical orientation, personality traits and mental ability. Japanese Journal of Educational Psychology, 40, 47-53.

Thorndyke, P. W., & Hayes-Roth, B. (1982). Differences in spatial knowledge acquired from maps and navigation. Cognitive Psychology, 14(4), 560-589.

HUMAN PERFORMANCE OF A TELEROBOTIC MANIPULATION TASK AS A FUNCTION OF VIEWING DISTANCE IN MONOSCOPIC AND STEREOSCOPIC VIEWING CONDITIONS

Clayton J. Hutto, Dennis A. Vincenzi, Steve Hall and Sathya Gangadharan
Embry-Riddle Aeronautical University

ABSTRACT

Telerobotic systems help extend human capability into hazardous or remote environments, thereby helping to increase human safety. Since many telerobotic tasks require fine manipulations conducted within the near visual field, it is important to understand relative performance differences that may exist as a function of the viewing distance from the manipulation task location. The purpose of this study was to determine what differences in human performance exists for teleoperation tasks at differences distances (20 cm, 60 cm, 100 cm, 150 cm, 200 cm, 250 cm) under either monoscopic or stereoscopic viewing conditions. Human performance was determined as measured by the average time taken to perform a simple telerobotic placement task, as well as by the overall accuracy of the placement (out of ten attempts) for each participant. Results of 180 naïve telerobotic operators' performance indicated significantly better execution of accurate and timely manipulation in stereoscopic viewing conditions, as well as significantly more accurate placement at closer distances. These results support previous research, and can now be applied to distances of less than 250 cm.

Keywords: Telerobotics, Robotics, Monoscopic and Stereoscopic Viewing, Performance

INTRODUCTION

The advantages of telerobotic systems to gainfully extend human capabilities into unstructured remote environments makes the use of telerobotic systems an ideal method of increasing human safety by removing the human from potentially hazardous environments such as outer space, undersea, and remote bomb disposal (Sheridan, 1992). Many telerobotic tasks require fine manipulation of objects within the remote environment, and are therefore conducted within the near visual distance. It is therefore important for telerobotic operators to understand human performance differences of manipulation tasks within the near vision operating range of telerobotic systems.

The human visual system is primary method humans use to gain information about their environment (Chapanis, 1996). Numerous depth cues interact to provide humans with a sense of depth for their environment. Monoscopic depth cues such as interposition (occlusion), lighting effects (shading and shadows), linear and geometric perspectives, texture gradients, and size constancy of familiar objects are very helpful for relaying three-dimensional information in two-dimensional formats, such as pictures or video monitor displays. The human visual system, however, is much more adept at interpreting depth information due to stereopsis, which is the ability to extract depth information from binocular cues (Coren, Ward, & Enns, 1999). Binocular cues such as retinal disparity (the angular offset between retinal images in the left and right eyes), and oculomotor cues such as vergence movements (rotation of the eyes to a point in space), and accommodation (compression or expansion of the lens to focus at a particular distance) all combine to produce human stereopsis (Coren, Ward, & Enns, 1999).

Some cues dominate others in certain situations (Cutting & Vishton, 1995). For example, a person attempting to thread a needle primarily uses stereopsis to determine the relative locations of the end of the thread and the eye of the needle, and usually brings the objects close to the eyes to increase the intensity of binocular and oculomotor cues. However, a submarine pilot is unlikely to use stereoscopic cues to determine the distance to a far-off buoy, instead relying on multiple pictorial depth cues (Pfautz, 1996). A principle factor for the dominance of one cue over another is the distance from the observer to the objects of interest in the environment (Nagata, 1991).

Stereoscopic depth cues such as binocular disparity, accommodation, and convergence are most useful at distances less than 2 meters (Surdick, Davis, King, Corso, Shapiro, Hodges, & Elliot, 1994). Previous studies have also shown that the estimated magnitudes of perceived distances for targets beyond this range are often erroneous (Kunnapas, 1968), and that binocular disparity, accommodation, and convergence cues are severely degraded for distances outside of normal arm length (Nagata, 1991; Boff, Kaufman & Thomas, 1986).

Numerous studies have investigated the effects of stereoscopic and indirect video viewing conditions within the context of aviation (Haskell & Wickens, 1993; Ellis, McGreevy & Hitchcock, 1987), for scientific visualization (Wickens, Merwin, & Lin, 1994; Sollenberger & Milgrim, 1993), and for remote operations (Massimo

& Sheridan, 1994; Pepper & Hightower 1984; Drascic, 1991; Drascic, Milgrim, & Grodski, 1989; Lumelsky, 1991). Surprisingly few of these studies were conducted for visual conditions less than 2 meters. Massimo & Sheridan (1994), for example, measured mean tasks times but did not find significant differences for teleoperator performance in direct versus video viewing. Massimo & Sheridan go on to state that further investigation is required for telerobotic manipulation tasks at viewing distances less than eight feet in order to take advantage of stereo vision (Massimo & Sheridan, 1994). It is therefore important to understand what differences or relationships, if any, exist for performance of teleoperation tasks at different viewing distances. In this study, the authors attempt to evaluate the extent that the viewing distance between telerobotic operators and the objects affects their performance of a manipulative task. It is hypothesized that as distance increases, performance will decrease due to the decrease in available depth information afforded the telerobotic operator. The performance loss over distance will be evidenced both by decreased placement accuracy and by increased average times-to-completion for the telerobotic manipulation task. Moreover, it is hypothesized that the additional availability of depth information in stereoscopic viewing mediums will significantly improve telerobotic operator performance at any distance within the near visual field (less than 250 cm), and will be evidenced by a superior times-to-completion and placement accuracy.

METHOD

Participants

A total of 180 naïve participants (131 male and 49 female) volunteered for this study. Ages ranged from 18 to 47 (mean = 21.58, SD = 3.57). None of the participants had previous experience with telerobotics. Three participants were replaced after reporting having visual acuity worse than 20/20 or known depth perception problems; all other participants reported normal or corrected to normal vision, and no problems with depth perception.

Apparatus

A Sony monoscopic video camera with model number CCD-TR87 was used in this study to relay visual information of the task environment to the telerobotic operator. A fifteen-inch Panasonic color monitor, model number CT13R14V, was used for the two-dimensional video display of the remote environment. The telerobotic system was a Questech Robot Manipulator Arm model number TCM, which was modified to allow the remote control to reach to a distance greater than 15 feet. A plastic ring measuring 3.81 cm in total diameter (center aperture measuring 2.7 cm), and a wooden dowel rod post with a diameter of 2.25 cm and a length of 40.5 cm, which was vertically fixed within the telerobotic arm work space, were used to evaluate performance of a remote manipulation task with the telerobotic system.

Design

This study examined the differences in performance of a telerobotic-placing as a function of viewing distance from the remote environment. Performance measurements were recorded for six viewing distances within the near visual field (20 cm, 60 cm, 100 cm, 150 cm, 200 cm, 250 cm), corresponding to both direct stereoscopic and indirect monoscopic video conditions. Dependent variables included the accuracy of the placing task (measured by the number of times out of ten attempts that a participant successfully dropped a ring completely to the bottom of a dowel post), and the time to completion for the task, which was measured in seconds for each attempted drop from the first motion of the robotic arm to the release of the ring. The study was conducted as a fully between-subjects experimental design.

Procedure

Each participant was shown the experimental apparatus with the telerobotic manipulator arm holding the ring, and was then instructed in the use of the manipulator arm. Upon the completion of the instructions for the telerobotic manipulator arm, the participants were asked to drop a plastic ring measuring 3.81 cm in total diameter (center aperture measuring 2.7 cm) over a wooden dowel rod post with a diameter of 2.25 cm and a length of 40.5 cm, which was vertically fixed within the telerobotic arm work space. The manipulator arm was reset to the same start position for each trial with the plastic ring being held in the arm's gripper. Participants were not allowed to view the work area while the manipulator arm was being reset to the start position by the researcher.

Accuracy was judged by the amount of times that the plastic ring fell to the bottom of the dowel rod out of ten drops (successful drops counted as "hits"). Rings that did not fall completely to the bottom of the dowel (i.e., rings that were hung up on the top of the dowel) or rings that missed the dowel were considered errors and were not counted as hits. Time was measured in seconds beginning from the first movement of the telerobotic manipulator arm and ending with the release of the ring.

In the direct stereoscopic view condition, a chinrest was used to ensure that all participants were at eye level with the top of the dowel, and to ensure each participant viewed the apparatus from the same distance. For the indirect view, a monoscopic video camera was leveled with the top of the dowel at the appropriate distance, and adjusted to approximate the same field of view and visual angel as the direct viewing condition.

RESULTS

This study examines human performance as a function of distance for teleoperators who performed a simple robotic placing task in the near visual field (20 cm to 250 cm) while viewing the environment either directly (with stereovision), or indirectly (with monoscopic video and 2D display). *Table 1* summarizes performance data as measured by the average time to completion for each of the experimental groups. *Figure 1* is a graphical representation of the group means presented in *Table 1*.

Table 1: Group Means and Standard Deviations for Average Time to Completion

Viewing Medium	Viewing Distance	Mean Time (s)	Std. Deviation	N
Direct View	20 cm	28.95	8.99	15
	60 cm	31.01	7.27	15
	100 cm	31.34	9.59	15
	150 cm	33.98	9.37	15
	200 cm	34.48	11.49	15
	250 cm	42.09	11.10	15
	Total	33.64	10.36	90
Indirect View	20 cm	54.97	20.66	15
	60 cm	55.69	15.62	15
	100 cm	54.45	10.27	15
	150 cm	55.93	13.95	15
	200 cm	54.84	27.62	15
	250 cm	56.07	26.25	15
	Total	55.33	19.53	90

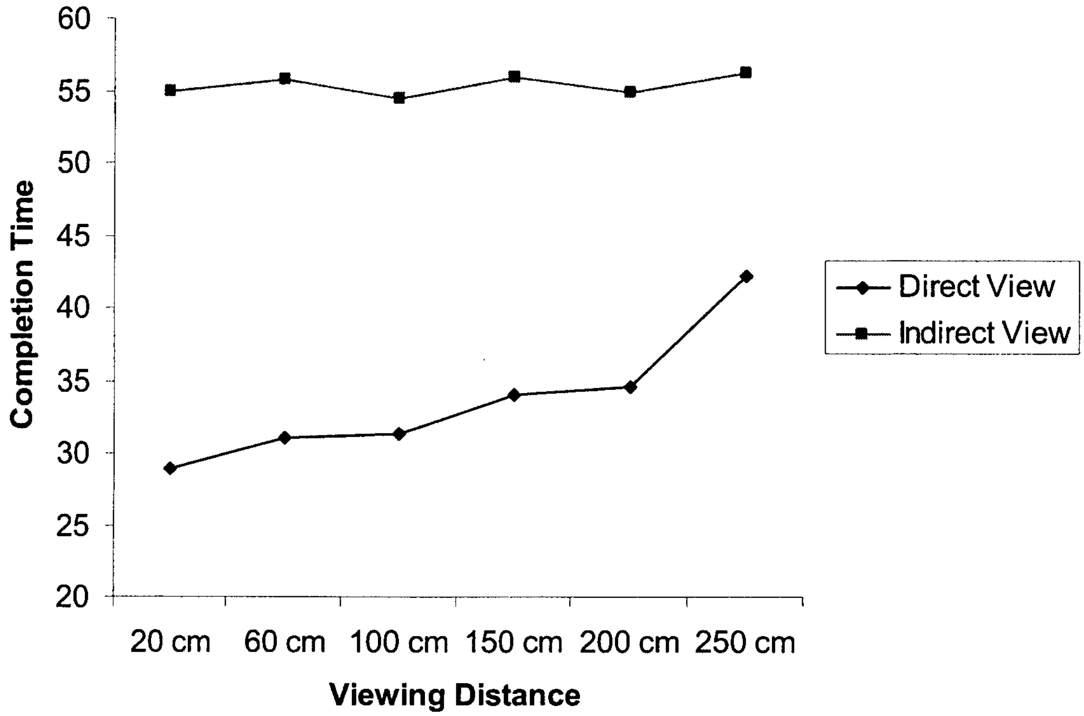


Figure 1: Line graphs represent group means for average time to completion in each experimental group

Additionally, the group means as measured by the drop accuracy (out of ten trials) for each participant are reported in Table 2 for each of the experimental groups, and Figure 2 is a graphical representation of the information in Table 2.

Table 2: Group Means and Standard Deviations for Drop Accuracy out of Ten Trials

Viewing Medium	Viewing Distance	Mean Accuracy (out of ten)	Std. Deviation	N
	20 cm	8.73	1.03	15
	60 cm	7.73	1.79	15
	100 cm	6.33	2.44	15
Direct View	150 cm	6.73	2.19	15
	200 cm	5.53	2.13	15
	250 cm	4.87	2.33	15
	Total	6.66	2.37	90
	20 cm	3.27	2.55	15
	60 cm	3.07	1.67	15
	100 cm	2.73	1.28	15
Indirect View	150 cm	2.67	1.99	15
	200 cm	2.07	1.49	15
	250 cm	1.67	1.80	15
	Total	2.58	1.87	90

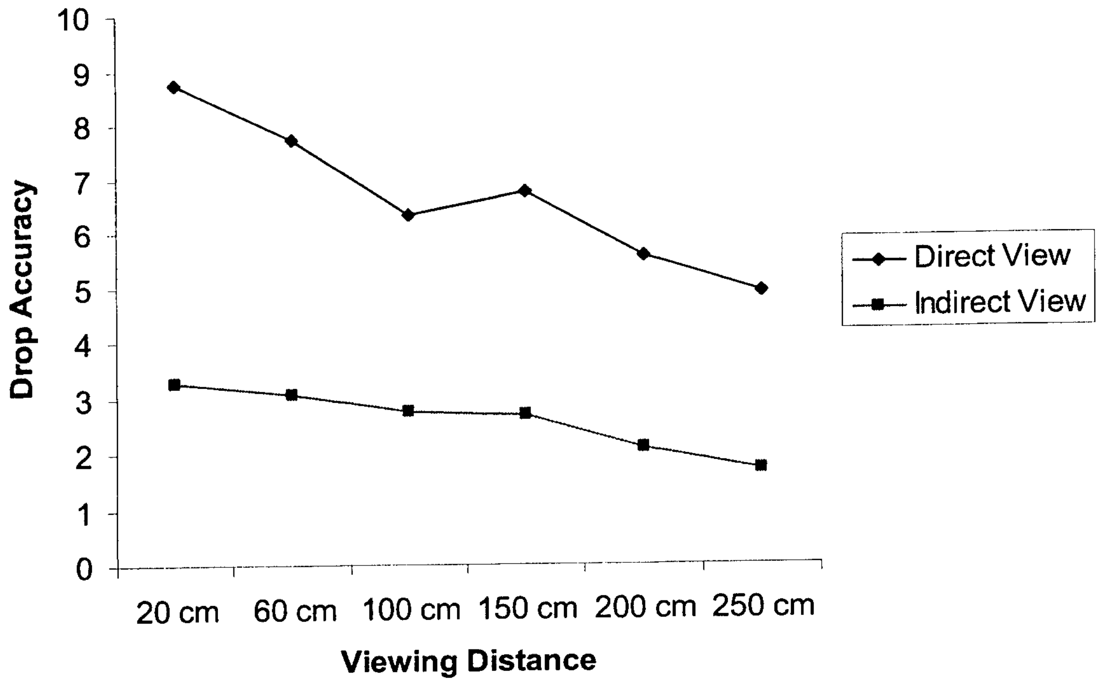


Figure 2: Line graphs represent group means for drop accuracy in each experimental group

The data were analyzed using a univariate analysis of variance for each dependent measure, and indicates a significant difference of the effects of view (direct/stereoscopic versus indirect/monoscopic) on human performance as measured both by average time to completion $F(1, 168) = 84.89, p < .01$, and for drop accuracy (number of hits out of ten attempted drops) $F(1, 168) = 198.28, p < .01$. Human performance as measured by the drop accuracy also indicate a significant difference for the effects of distance $F(5, 168) = 8.00, p < .01$, but not for time to completion $F(5, 168) = 0.76, p = .58$. *Table 3* presents additional source information regarding the results.

DISCUSSION

Based on the results of this study, we can report a significant difference in human performance of a telerobotic manipulation across viewing medium conditions based on both the average completion times and drop accuracy for 180 naïve telerobotic operators. Additionally, a significant difference in human performance as measured by drop accuracy alone was observed as the viewing distance increased. That is to say that those operators completing the manipulation task under the direct viewing condition performed much better than those under the indirect viewing condition, and a general decline in performance occurred as viewing distance increased.

These results support the research hypothesis that human performance will decrease as a function of distance. Specifically, telerobotic operators performed significantly better at closer viewing distances, although only the performance measured by placement accuracy was significant, and much better in the direct (stereoscopic) viewing condition due to the additional binocular depth cue advantages afforded to them. These findings are consistent with previous research that reports advantages of stereoscopic viewing over monoscopic viewing (Barfield & Rosenberg, 1995; McLean, Prescott, & Podhorodeski, 1994; Yeh & Silverstein, 1992), and can now be applied to viewing distances less than 250 cm.

Table 3: Source Data for Univariate Analysis of Variance

Average Time to Completion

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
View (Direct / Indirect)	21161.85	1	21161.85	84.89	.000	.336	84.89	1.000
Distance	947.85	5	189.57	0.76	.580	.022	3.80	0.269
VIEW * DISTANCE	683.07	5	136.61	0.55	.740	.016	2.74	0.199

Drop Accuracy

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
View (Direct / Indirect)	748.272	1	748.27	198.28	.000	.541	198.28	1.000
Distance	150.917	5	30.18	8.00	.000	.192	39.99	1.000
VIEW * DISTANCE	27.361	5	5.47	1.45	.209	.041	7.25	0.502

a: Computed using alpha = .05

The magnitude of the differences between the means for each dependent measure suggests that telerobotic operators rely more heavily on the stereoscopic cues that are available in binocular vision *in general* than on the *intensity* of those cues as they relate to viewing distances less than 250 cm. This also suggests that viewing medium (a determinate of available depth information) should be considered a more relevant factor than viewing distance for operators when performing telerobotic tasks at ranges within the near visual field. The results of this study demonstrate how the increase in the awareness of depth information as a function of distance in monoscopic and stereoscopic viewing conditions translates directly into better human performance of a remote telerobotic manipulation task for distances less than 250 cm.

REFERENCES

Barfield, W., and Rosenberg, C. (1995). Judgments of Azimuth and Elevation as a Function of Monoscopic and Binocular Depth Cues Using a Perspective Display. Human Factors, 37*(1)*, 173-181.

Boff, K., Kaufman, L., & Thomas, J. (1986). *H*andbook of Perception and Human Performance, Volume I, Sensory Process and Perception. John Wiley and Sons, New York.

Chapanis, A. (1996). Human Factors in Systems Engineering. New York: John Wiley & Sons.

Coren, S., Ward, L.M., & Enns, J.T. (1999). Sensation and Perception (5th ed., pp. 253-267). Fort Worth: Harcourt College Publishers.

Cutting, J. & Vishton, P. (1995). Perceiving Layout and Knowing Distance: The Integration, Relative Potency and Contextual use of Different Information about Depth. In W. Epstein & S. Rogers (Eds.), Perception of Space and Motion (pp. 69-118). New York: Academic Press.

Drascic, D.(1991). Skill acquisition and task performance in teleoperation using monoscopic and stereoscopic video remote viewing. Proceedings of the Human Factors Society 35th Annual Meeting – *1991*. 1367-1371.

Drascic, D., Milgrim, P., & Grodski, J. (1989). Learning effects in telemanipulation with monoscopic versus stereoscopic remote viewing. IEEE International Conference on Systems, Man and Cybernetics, 3, 1244-1249.

Ellis, S.R., McGreevy, M.W., & Hitchcock, R.J. (1987). Perspective traffic display format and airline pilot traffic avoidance. Human Factors, 29, 371-382.

Haskell, I.D., & Wickens, C.D. (1993). A theoretical and empirical comparison of two- and three-dimensional displays for aviation. International Journal of Aviation Psychology, 3, 87-109.

Kunnapas, T. (1968). Distance perception as a function of available visual cues. Journal of Experimental Psychology, *77(4)*, 523-529.

Lumelsky, V. (1991). On human performance in telerobotics. IEEE Transactions on Systems, Man and Cybernetics, 21, 971-981.

Massimino, M.J., & Sheridan, T.B. (1994). Teleoperator performance with varying force and visual feedback. Human Factors, 36*(1)*, 145-157.

McLean, G. F., Prescott, B., and Podhorodeski, R. P. (1994). Teleoperated System Performance Evaluation. IEEE Transactions on Systems, Man, and Cybernetics, 24*(5)*, 796 - 804.

Nagata, S. (1991). How to Reinforce Perception of Depth in Single Two-Dimensional Pictures. In S.R. Ellis, M. Kaiser, & J. Grunwald (Eds.) Pictorial communication in virtual and real environments (pp. 182-195). London: Taylor & Francis.

Pepper, R.L., & Hightower, J.D. (1984). Research Issues in Teleoperator Systems. Proceedings of the Human Factors Society 28th Annual Meeting – 1984, 803-807.

Pfautz, J. (1996). Distortion of Depth Perception in a Virtual Environment Application. Master's Dissertation, Massachusetts Institute of Technology.

Sheridan, T.B. (1992). Telerobotics, automation, and human supervisory control. Cambridge, MA: The MIT Press.

Surdick, R.T., Davis, E., King, R., Corso, G., Shapiro, A., Hodges, L., & Elliot, K. (1994). Relevant Cues for the Visual Perception of Depth: Is Where You See It Where It Is? Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting, 1305-1309.

Wickens, C.D., Merwin, D., & Lin, E. (1994). Implications of graphics enhancements for the visualization of scientific data: Dimensional integrality, stereopsis, motion, and mesh. Human Factors, 36, 44-61.

Sollenberger, R.J., & Milgrim, P. (1993). Effects of stereoscopic and rotational displays in a three-dimensional path-tracing task. Human Factors, 35, 483-501.

Yeh, Y. Y., and Silverstein, L. D. (1992). Spatial Judgments with Monoscopic and Stereoscopic Presentation of Perspective Displays. Human Factors, 34, 583-600.

ENHANCING TEAM TRAINING AND PERFORMANCE WITH THE DEVELOPMENT OF HUMAN CENTERED PERFORMANCE ASSESSMENT TOOLS

John E. Deaton, Benjamin Bell and Jennifer Fowlkes
CHI Systems, Inc.

Clint Bowers and Florian Jentsch
University of Central Florida

Meredith A. Bell
Florida Institute of Technology

Abstract
Assessing team performance is a difficult task for instructors because of factors such as high workload and the difficulty observing relevant aspects of performance. Automated performance measurement systems may provide assistance to instructors and improve the quality of performance assessment. In this effort, research was performed to compare the ability of flight instructors to give diagnostic ratings of team performance with or without computer assistance. Computer assistance was provided by the EPIC (Enhancing Performance with Improved Coordination) tool, which was designed to assist instructors in evaluating team performance during simulation-based training. Flight instructors watched a video showing a search and rescue mission and were asked to rate the aircrew depicted in the video on the three elements of mutual support (backup, flight discipline, and communication). The performance of two groups was compared: an experimental group that received EPIC assistance and a control group that did not. It was found that the instructor group operating with EPIC had more differentiated and accurate ratings of team performance compared to the control group. These results suggest that approaches such as EPIC can serve to compliment and augment instructor capabilities, providing more powerful and diagnostic team performance assessment systems.

Keywords: Team Performance Assessment; Mutual Support; Aircrew Training; Automation

INTRODUCTION

Training teams of military and civilian operators such as aircrews is critical to effective performance in operational environments increasingly characterized as dynamic, integrated, and information-rich. In order to assure that team training provides the necessary levels of readiness for crews operating in such complex environments, advances in *team performance assessment* are needed, both to enhance the value of time spent in training, and to provide overall measures for evaluating the training approaches being practiced. Unfortunately, instructors confront high workload situations when simultaneously evaluating individual and team performance, and can rapidly become saturated while monitoring multiple aspects of mission performance. Moreover, much of the work that team members perform may be difficult for humans to observe, and thus assessments of teamwork may be based on just a portion of relevant performance. These problems are well documented in the literature (e.g., Baker & Salas, 1992), and thus, support tools are needed that can assist instructors and improve the quality of team performance assessments. In this paper, we report on a prototype system, EPIC (Enhancing Performance with Improved Coordination), designed to assist instructors through automated performance assessment. Below we describe our concept for EPIC and then describe research performed to begin to validate the effects of EPIC features on instructor assessments of team performance.

EPIC

EPIC is intended to help flight instructors gain a more complete picture of team performance during simulation-based training than they would have otherwise. The intent is for EPIC to: 1) keep track of scenario events, 2) alert instructors of impending events using triggers such as location, time, or patterns among entities, 3) collect and monitor information that instructors would typically not be able to observe, and 4) allow instructors to focus on those performance dimensions that are best evaluated by human observers. EPIC will be tuned to detect and recognize events in training scenarios that serve as measurement opportunities. Once events are detected, EPIC will initiate follow-on actions such as alerting the instructor to perform measurements, initiating automated measurements, or inserting additional events that provide potentially more diagnostic measurement opportunities. It is anticipated that the improved team performance diagnosis enabled by EPIC can serve to enhance the development of team-specific skills, reveal the strengths and limitations of current team training approaches, and enhance crew performance and readiness. The intent of the research described in this paper is to begin to assess the impact of EPIC features on instructor ratings of team performance.

Mutual Support

For the application of EPIC described herein, the naval tactical aviation domain was targeted and "mutual support" was selected as the construct for measurement. Mutual support is defined as: "That support which units render each other against an enemy, because of their assigned tasks, their position relative to each other and to the enemy, and their inherent capabilities." (Department of Defense Joint Publication 1-02). Mutual support has tactical importance to aircrews and is highly related to success and survivability of the Navy's basic air fighting unit, the two aircraft "section." Through a review of both Navy and civilian aviation documents, three dimensions of mutual support were identified: flight discipline, communication, and backup. In addition, process behaviors were identified that enable mutual support and that provide indications of success in each of the three dimensions. The dimensions and process behaviors are shown in Figure 1.

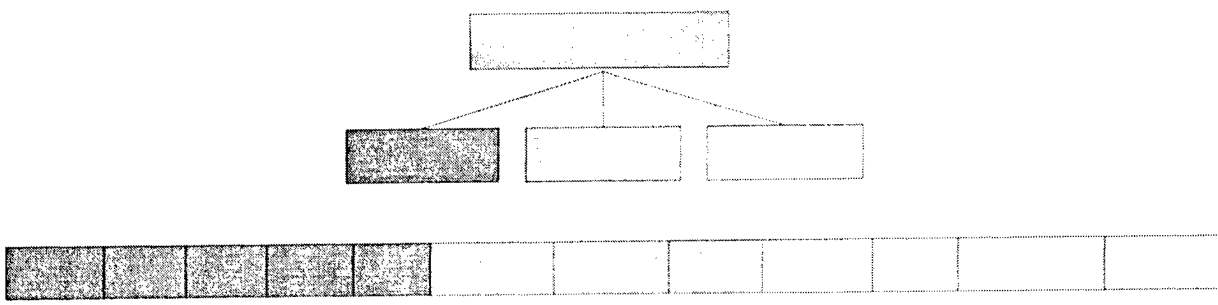


Figure 1. Mutual Support Construct (ROE: Rules of Engagement, SOP: Standard Operating Procedures).

METHOD

Design

The approach used was to compare the ability of instructors to give differentiated and accurate ratings of performance with or without assistance from EPIC. To facilitate consistency across participants, an experimental design was used in which flight instructors watched a video-taped performance of two pilots flying a joint mission in two distributed, PC-based flight simulators. The pilots depicted in the video followed a script that had them perform at three distinct performance levels across the three dimensions of mutual support: (a) below standard on flight discipline, (b) standard on communications, and (c) above standard on backup. Due to the difficulties of observation inherent in the video-tape, which mimicked the difficulties found in the real world described above, it was anticipated that instructors without EPIC would have difficulty forming a differentiated (and thus more valid) set of ratings across the dimensions than would those who had EPIC available. In particular, it was expected that EPIC could provide instructors with specific information about the pilots' performance in the area of flight discipline that the instructors would normally have difficulty observing. The mission scenario consisted of a two-ship flight to airdrop rescue goods to shipwrecked survivors of two yachts. Within the scenario, opportunities were imbedded to demonstrate good and poor performance in the three dimensions of mutual support, such as a failure to maintain horizontal and vertical separation between aircraft (flight discipline), the use of standard air traffic control-related calls (communications), and mutual warning among the aircraft in the flight of conflicting traffic (backup).

Participants

Twenty subjects participated in the research and were randomly assigned to either the treatment group (receiving simulated EPIC alerts and cues) or control group. Participants were flight instructors at the Florida Institute of Technology flight program.

Materials

The video observed by the participants was filmed in three segments. The first segment provided a brief introduction to the study, the study tasks, and the constructs of mutual support. The second segment depicted the preflight brief. For this segment, a video style similar to an interview was chosen, with two camera views alternately showing the two pilots as they discussed the upcoming flight. For the third section of the video, the actual flight, up to four camera views were combined into a simultaneous

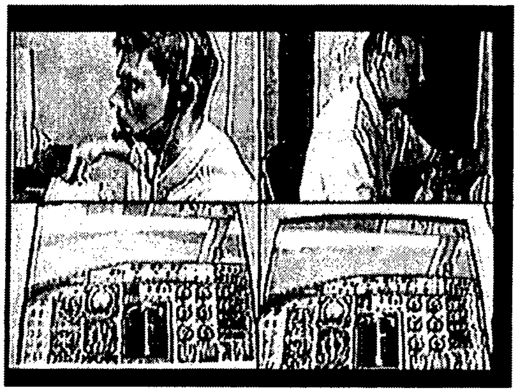


Figure 2. Frame from Mission Videotape.

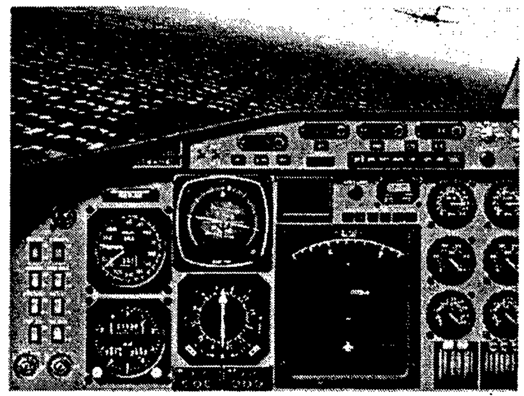


Figure 3. En-route in Formation.

display, using a quad view feature (Figure 2). The quad view showed: (a) pilot aircraft 1, (b) pilot aircraft 2, (c) simulation screen/cockpit view for aircraft 1, and (d) simulation screen/cockpit view for aircraft 2. The quad view alternated with a single, larger screen capture of the instrument panel and out-the-window view for aircraft 2 (Figure 3) which allowed participants a better look at the instruments and flight parameters.

Other study materials included instructions and response materials (i.e., grade sheets) for the study participants. The grade sheets instructed the respondents to give global ratings of performance in each of the three behavioral dimensions of mutual support. They also included examples of critical behaviors in the three dimensions, which were aligned with the behaviors presented in the first segment of the video.

Procedure

Participants viewed the videotaped simulation of the mutual support mission. Participants in the EPIC (treatment) group were additionally provided with a display that gave periodic reports simulating the functionality envisioned for EPIC. Specifically, the display was updated in synchronization with events depicted in the video, and it provided the EPIC participants with information pertaining to performance related to flight discipline. This information was based on data that readily could be collected from flight simulators (e.g., heading, altitude, airspeed). Participants were required to rate the performance of the flight crew portrayed on the videotape regarding their performance on the three mutual support dimensions using the following scale: 1 = Unsatisfactory; 2 = Below Standard; 3 = Standard; 4 = Above Standard; or 5 = Excellent.

RESULTS

A 2 (between) x 3 (within) mixed-model analysis of variance (ANOVA) was calculated in which the between subject variable was group (EPIC vs. non-EPIC condition), and the within subject variable was mutual support dimension (flight discipline, communication, and backup). The results showed a non-significant main effect for EPIC vs. non-EPIC condition, $F(1,18) = .70$, $p = .42$. The mean for the EPIC condition was 2.80 ($s = 0.20$), while the mean for the non-EPIC condition was 3.03 ($s = .20$). The main effect for mutual support dimension was significant, $F(2,36) = 26.65$, $p < .01$. The means for the three performance dimensions were 2.30 ($s = 0.16$), 2.75 ($s = 0.13$), and 3.70 ($s = 0.24$) for Flight Discipline, Communication, and Backup, respectively.

The interaction of EPIC/non-EPIC condition and mutual support dimension was also significant, $F(2,36) = 3.17$, $p = .05$. Figure 4 shows the interaction effect. The graph shows virtually no difference between Flight Discipline and Communication for the non-EPIC condition, and an increased performance rating for the Backup dimension. The means for the non-EPIC condition for Flight Discipline, Communication, and Backup were 2.70 ($s = 0.22$), 2.70 ($s = 0.18$), and 3.70 ($s = 0.33$), respectively. Instructors in the EPIC condition, however, showed differences across all three dimensions. The means for the EPIC condition for Flight Discipline, Communication, and Backup were 1.90 ($s = 0.22$), 2.80 ($s = 0.18$), and 3.70 ($s = 0.33$), respectively. That is, instructors in the non-EPIC condition assigned similar ratings to flight discipline and communication, when in fact the crew's demonstrated performance on flight discipline was much lower. Conversely, instructors who used EPIC showed clear differentiations in their ratings across the three dimensions. Thus, a comparison of Figure 4 with the performance that a priori had been modeled in the video shows that the instructors in the EPIC condition gave ratings which were entirely congruent with the performance shown in the video. Instructors in the non-EPIC condition, however, deviated from the accurate ratings in at least one of the three areas, namely in flight discipline.

DISCUSSION

As expected, the EPIC prototype display modeled in the study was effective in helping instructors assign more differentiated and more accurate performance ratings to the crew. In the video viewed by all participants, the crew was scripted to perform at "Below Standard" for flight discipline, at "Standard" for communication, and at "Above Standard" for backup. Only the instructors who had access to EPIC

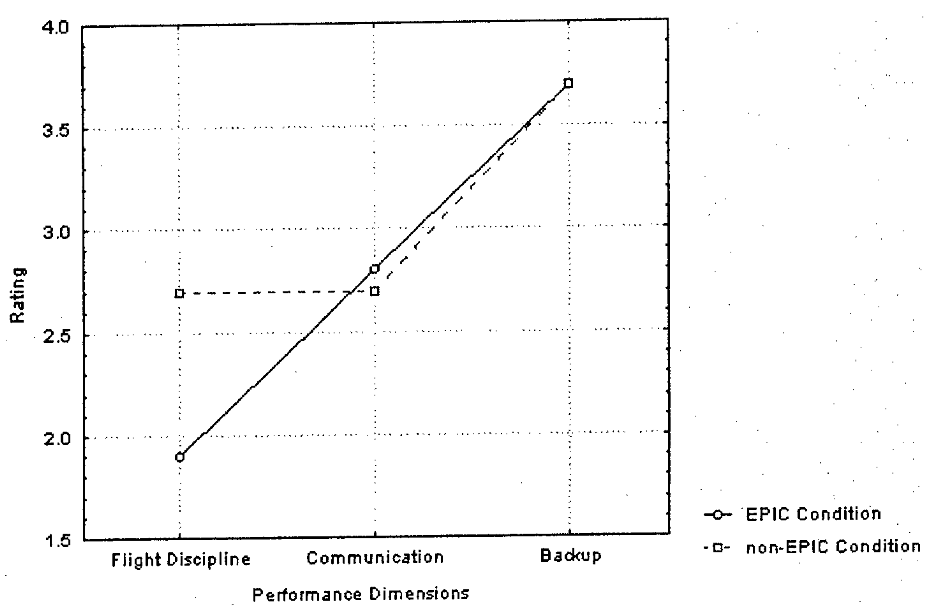


Figure 4. Performance ratings as a function of EPIC/non-EPIC conditions and performance dimensions.

correctly assigned "Below Standard" ratings to the crews in the area of flight discipline. Instructors without EPIC, conversely, assigned an average rating of "Standard" to the crews in this area, thereby showing a halo effect with communication. Interestingly, providing the EPIC data to instructors on the flight discipline dimension did not lead to a "reverse halo" effect. That is, in our study, instructors with EPIC continued to assign the correct ratings of "Standard" and "Above Standard" to communication and backup, respectively. Thus, unlike in a previous Navy policy-capturing study (K. Jentsch, 13 September 2003, personal communication) with combat-information-center teams, the provision of automated performance data in one area did not lead to less accurate ratings in other performance areas.

CONCLUSION

EPIC is intended to enhance team training and improve combat readiness by providing automated capabilities for team assessment in the context of scenario-based training. EPIC can reduce instructor workload and provide access to a broader sample of relevant performance, thereby improving measurement. The present study found that the instructor group operating with EPIC had more differentiated and accurate ratings of team performance compared to the control group. These results suggest that approaches such as EPIC can serve to compliment and augment instructor capabilities, providing more powerful and diagnostic team performance assessment systems.

ACKNOWLEDGEMENTS

This work was supported through a Phase I Navy SBIR, contract N61339-03-C-0033.

REFERENCES

Baker, D.P., & Salas, E. (1992). Principles for measuring teamwork skills. Human Factors, 34, 469-475.

DISPLAYS

QUANTIFICATION OF VISUAL CLUTTER USING A COMPUTATIONAL MODEL OF HUMAN PERCEPTION: AN APPLICATION FOR HEAD-UP DISPLAYS

Michael Zuschlag
Volpe National Transportation Systems Center

ABSTRACT

A means of quantifying the cluttering effects of symbols is needed to evaluate the impact of displaying an increasing volume of information on aviation displays such as head-up displays. Human visual perception has been successfully modeled by algorithms that process an image through a bank of visual filters for a range of spatial frequencies and orientations. The model proposed here derives a vector of "feature density" values from these filtered images where each value represents the degree to which the image contains a particular spatial frequency and orientation. Differences in these feature densities between a target and a context is used to calculate the degree the target is salient relative to the context.

Keywords: Head-up displays; Clutter; Image analysis; Visual perception models

INTRODUCTION

Advanced technology is bringing an increasing volume of information to the flight deck that must be displayed in the relatively limited area of the flight deck displays. Display symbols must be designed such that the symbols are salient, and thus easy to read, but not so dominant that they create clutter by visually interfering with other significant objects. The compromises a designer must make between salience and clutter can be seen in head-up displays (HUDs), which are quite sensitive to the cluttering effects of symbology (e.g., Ververs and Wickens, 1998). While qualitative design guidelines emphasize minimizing HUD clutter (Newman, 1995), new technologies such as enhanced vision systems imply HUDs will required to display even more information. Designers and other display evaluators would be greatly aided if a means of quantifying the level of clutter were available so that the salience of symbols can be more optimally balanced.

A Model to Calculate Visual Salience

Salience as Average Color Difference

As a first approximation, assume a monochrome display, as is the case with current aviation HUDs. The degree a monochrome target, o (i.e., a HUD symbol), is salient with respect to a context, i, is related to the color contrast between the color of the target and the color of the context, where color includes both luminance and chromatic components. The perceptual difference in any two colors can be represented by their Euclidean distance in L*u*v* space (Wyszeski and Stiles, 1982). It is assumed that perceptual salience has an inverse exponential relationship to perceptual color difference. Thus, the salience of target o in the context of i should be related to average salience of the color differences of each point of i:

$$S_{oi}(0) = \frac{1}{A_i}\left[\iint 1 - \exp\left(-\beta\, p_{\Delta,xy} \right) dx\, dy \right],$$

where A_i is the area of the context for the target, $p_{\Delta,xy}$ is the L*u*v* distance between the target color and the color of a point at coordinates x and y in the context, and β is a constant to be empirically determined.

This inverse exponential relationship implies that after a certain level of color contrast, additional contrast has little effect on human performance, which is consistent with experimental research on HUDs (Weintraub and Ensing, 1992).

An application of this formula is illustrated in Figure 1, where a HUD symbol, a Bray-style flight path marker (FPM) (Weintraub and Ensing, 1992), is compared to uniform backgrounds of 0%, 75%, and 87% gray. In this example, the Red-Green-Blue (RGB) color values of the background images were assumed to be of the sRGB color space (International Electrotechnical Commission, 1999) in order to convert them to L*u*v* difference distances. The resulting $S_{oi}(0)$'s are shown, where higher value represents greater salience. The parameter β was rather arbitrarily set to 0.05. In practice, this value would be determined by fitting the model to human performance.

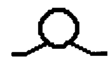
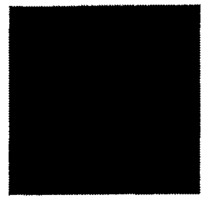
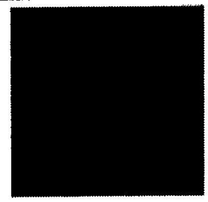
	(a)	(b)	(c)
Target and Context Note: For printing purposes, a black rather than bright green FPM is used.			
Context gray scale	0%	75%	87%
$S_{oi}(0)$	0.993	0.729	0.372

Figure 1. Calculated salience that compares average background color to target color.

The calculated salience agrees with intuition; the value of $S_{oi}(0)$ decreases as the contrast between the target and its context decreases. However, $S_{oi}(0)$ itself does not take into account the cluttering effects of any visual features that may reside within the context. In a HUD, these features may represent variations in the background texture (e.g., features of cloud or terrain), objects within the out the window (OTW) scene (e.g., traffic and runways), other nearby HUD symbols, and possibly overlaying textures from an enhanced vision or synthetic vision system. Consider Figure 2. The value of $S_{oi}(0)$ for (a) is about the same as (b). However, one would probably expect the target in (b) to be more difficult to see. Thus, in addition to $S_{oi}(0)$, one needs to account for the degree the context has features similar in shape and color to the target.

	(a)	(b)
Target and context		
$S_{oi}(0)$	0.729	0.752

Figure 2. Failure of average color difference in accounting for cluttering effects of texture.

Salience as Differences in Features

In artificial intelligence research, certain successful models of computational visual feature detection are based on results from low-level human and primate visual perception studies (Doll, McWhorter, Wasilewski, and Schmieder, 1998; Wilson, 1991). These models analyze an image for a range of spatial frequencies and orientations (Bergen and Landy, 1989). The greater a target differs from its context in the amplitudes of the spatial frequencies across the orientations, the more salient the target (Itti, Koch, and Niebur, 1998).

Specifically, let **I** be a two-dimensional array representing the perceptual salience of each pixel in an image compared to the target's color (i.e., $1 - \exp(-\beta\, p_{\Delta,xy})$), where the image may be the context or the target itself. Given a monochrome and transparent HUD, the array element values for any HUD symbol, including the target, are all 0 except for the background, which is set to 1, so the array represents the HUD symbol against a high contrast background.

The features of an image with respect to the target's color are then quantified as illustrated in Figure 3.

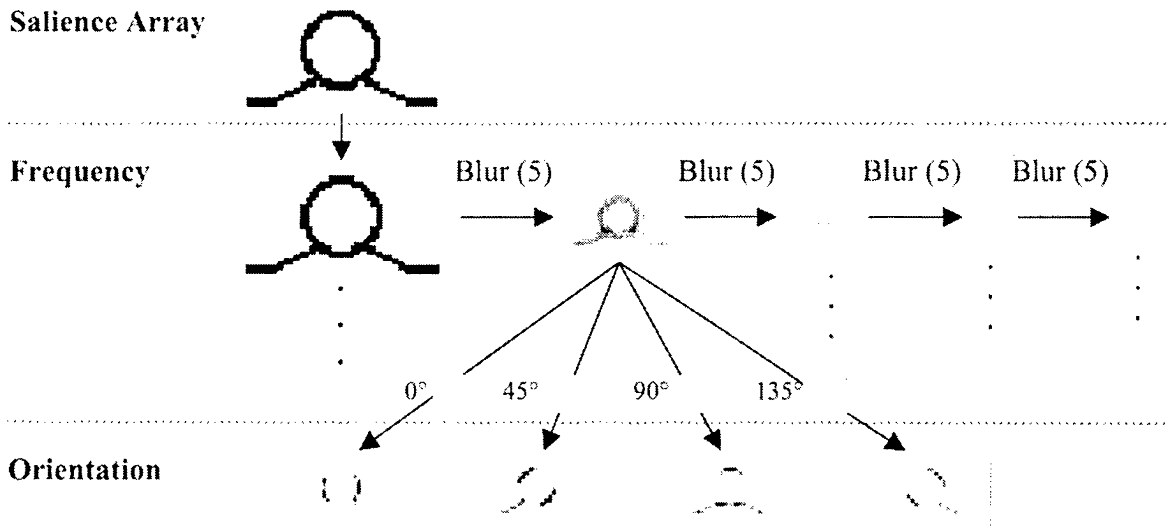


Figure 3. Algorithm for feature detection.

First, a range of frequencies for the spatial filtering is accomplished by building a "pyramid" of images of successively lower frequency v, where each successive image $\mathbf{I}_{v/2}$ is half the width and height of its predecessor $\mathbf{I}_v$. This done by first blurring a copy of the predecessor image as follows:

$$\mathbf{I}_{blurred} = \mathbf{I}_v * b * b^{\mathrm{T}}, \quad b = \{\ 0.05, 0.25, 0.40, 0.25, 0.05\ \}.$$

Then shrinking it to half its dimensions by summing the value of each set of four adjacent pixels:

$$p_{v/2,\,x/2,\,y/2} = (p_{blurred,xy} + p_{blurred,x+1,y} + p_{blurred,xy+1} + p_{blurred,x+1,y+1}),$$

where p_{xy} is a pixel at position (x,y) in $\mathbf{I}$.

This is done four times resulting in five octaves of frequency filtering, spanning the detectors for spatial frequencies found in the visual cortex (Wilson and Gelb, 1984).

Then, four spatially filtered arrays are generated for each frequency by convolving the array first by a five-element Gaussian vector then an orthogonal three-element approximately Gaborian vector. This is done for vectors angled at 0, 45, 90, and 135 degrees, which again roughly corresponds to detectors in the cortex. An absolute value is taken of the resulting element values. For example, the 0 degree filtering of an image corresponding to spatial frequency v is:

$$\mathbf{I}_{v,0} = |\ \mathbf{I}_v * b * g^{\mathrm{T}}\ |, \quad g = \{\ -0.5, 1.0, -0.5\ \}.$$

While the 90 degree filtering is:

$$\mathbf{I}_{v,90} = |\ \mathbf{I}_v * b^{\mathrm{T}} * g\ |.$$

Thus, for each input image $\mathbf{I}$, the image analysis yields 20 output arrays, $\mathbf{I}_{v\theta}$ (5 frequencies · 4 orientations). In a sense, high values of the elements of $\mathbf{I}_{v\theta}$ represent an edge at orientation θ where image color changes with respect to the target color. For the same L*u*v* distances, changes towards the target color are weighted more than changes away owing to the transforming the L*u*v* distances by $1 - \exp(-\beta\, p_{\Delta,xy})$. Uniform images have no edges, so all elements of such a $\mathbf{I}_{v\theta}$ are 0.

Let the feature density of an image $f_{i,v\theta}$, represent the degree that image i has features per unit area of spatial frequency v and orientation θ that are similar in color to the target color. This is calculated by summing all array elements, $p_{v\theta,xy}$, of $\mathbf{I}_{v\theta}$ and dividing by the area of the image, A_i:

$$f_{i,v\theta} = \frac{1}{A_i}\sum_x \sum_y p_{v\theta,xy}.$$

In this manner, the 20 feature density values are calculated for both the target and its context. Let $S_{oi}(v,\theta)$ be the salience of target o in context image i with respect to features of v and θ. For a target that overlays and combines with the context, much as a HUD symbol would combine with the OTW view or enhanced vision system imagery, target salience is considered to be proportional to the degree the target adds features to the context:

$$S_{oi}(v,\theta) = w_{v\theta} \frac{f_{o,v\theta}}{(f_{o,v\theta} + f_{i,v\theta})},$$

where $w_{v\theta}$ is an empirically derived weight representing the significance of the corresponding feature in perceptions of salience. Note that when o is compared to a featureless uniform context, i, each $S_{oi}(v,\theta)$ simply equals $w_{v\theta}$.

For a target that is presented proximal to the context, such as a HUD symbol with respect to other HUD symbols, target salience is considered to be proportional to the degree the target has different features from the context, weighted by a function of the target and context spatial separation $d(i,o)$:

$$S_{oi}(v,\theta) = d(i,o)\, w_{v\theta} \frac{| f_{o,v\theta} - f_{i,v\theta} |}{(f_{o,v\theta} + f_{i,v\theta})}.$$

Overall, the salience of a target o with respect to the context i is the combined effects of $S_{oi}(0)$ and all $S_{oi}(v,\theta)$. That is, the salience of the features must be weighted by the background salience, compensating, in a sense, for the salience of target's background pixels being set to 1.0. Thus, total salience, S_{oi}, is:

$$S_{oi} = S_{oi}(0) \sum_{v} \sum_{\theta} S_{oi}(v,\theta).$$

Model Performance

As a demonstration of this model, consider Figure 4. With the FPM symbol acting as the target and three backgrounds of varying clutter each acting as contexts, $S_{oi}(0)$ and $\Sigma\Sigma S_{oi}(v,\theta)$ are calculated with respect to the target's color. Relatively arbitrary parameters are used: all $w_{v\theta} = 0.05$, $\beta = 0.05$.

As can be seen in Figure 4 moving from (a) to (c), $S_{oi}(0)$ decreases as more cluttering features are added to the context, as the average color becomes darker and thus more like the color of the target (i.e., black). Note also how the $\Sigma\Sigma S_{oi}(v,\theta)$ sharply decreases with additional features, with the total calculated salience of Figure 4(c) being 0.274. Contrast that now to Figure 2(a), for which $S_{oi}(0) =$ 0.729, and $\Sigma\Sigma S_{oi}(v,\theta) = 1$ (all 20 $S_{oi}(v,\theta) = w_{v,\theta} = 0.05$), resulting in a substantially higher calculated overall salience of 0.729. Indeed, using these arbitrary parameters, Figure 4(c) is rated as less salient than even Figure 1(c) (total salience = 0.372), which more or less corresponds to intuition.

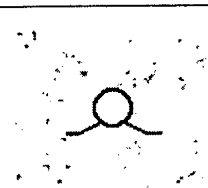
	(a)	(b)	(c)
Target and context			
$S_{oi}(0)$	0.988	0.920	0.752
$\Sigma\Sigma S_{oi}(v,\theta)$	0.926	0.591	0.364
Total Salience	0.916	0.544	0.274

Figure 4. Salience calculated from differences in average color and features.

As another illustration, consider Figure 5. Here, average grayscale and $S_{oi}(0)$ for the contexts are relatively constant and only the features of each context are varied. The context for Figure 5(a), dominated by high vertical frequencies, has few features in common with the FPM, so the model rates the FPM to be more salient there. In contrast, the FPM has strong diagonal features of high to low frequencies, and thus the salience is rated lower in Figure 5(b). This is fairly consistent with intuition; a better correspondence to human experience can be expected with more systematic parameter fitting.

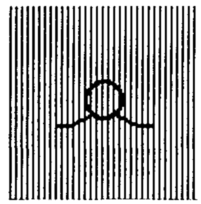
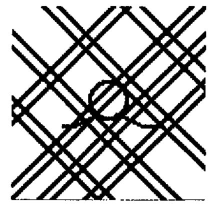
	(a)	(b)
Target and context		
Average grayscale	67%	67%
$S_{oi}(0)$	0.665	0.694
$\Sigma\, \Sigma\, S_{oi}(\nu,\theta)$	0.568	0.343
Total Salience	0.378	0.238

Figure 5. Effects of different features on calculated salience, holding average color difference constant.

CONCLUSION

The model presented here performs in accordance with one's intuition in accounting for the degree clutter interferes with symbol salience, suggesting that this approach is promising. The final verdict will depend on experimental validation in which the model's predictions will be compared to human performance. Before it can be used to evaluate actual HUDs, however, a number of details need to be addressed. Firstly, the spatial separation function $d(i,o)$ must be specified. Secondly, most objects viewed in and/or through a HUD vary in shape and color, thus ultimately a sample of representative of images for each object is necessary. Thirdly, for the sake of fast processing, it is preferred if one can evaluate the feature densities, $f_{i,\nu\theta}$, of each *component* of the context then somehow calculate their joint effect on each target; this calculation may be effectively approximated by a simple sum of the individual effects. Fourthly, in actual HUDs, the true color of HUD symbology is affected by the color of the background. Furthermore, HUDs are designed to vary in brightness to maintain a constant contrast ratio. The significance of these characteristics needs to be addressed. These characteristics may simplify the implementation of the model: a constant contrast ratio implies the luminance contribution to $S_{oi}(0)$ is constant so that one only needs to estimate the chromatic differences between the OTW view and the HUD symbols.

If successful, this approach can ultimately be generalized to other aviation displays such as navigation displays. Application to more traditional aviation displays may be on the one hand simpler, as most aviation displays have a uniform background (typically black). On the other hand, most aviation displays are not monochrome, implying a need to evaluate the salience of each object with respect to multiple target colors.

ACKNOWLEDGEMENTS

This study was supported by FAA Office of the Chief Scientific and Technical Advisor for Human Factors, AAR-100, working in cooperation with the FAA Transport Airplane Directorate, AMN-111. The author would like to express appreciation to the following people who made this document possible: Dale Dunford of the FAA, Chuck Oman, Miwa Hayashi, Pawan Sinha, and James DiCarlo of the MIT, and Dan Hannon of the Volpe Center.

REFERENCES

Bergen, J. R. and Landy, M. S. (1989). Computational modeling of visual texture segregation. In M. S. Landy, and J. A. Movshon (Eds.), Computational models of visual processing (pp253-271). Cambridge, MA: MIT Press.

Doll, T. J., McWhorter, S. W., Wasilewski, A. A., and Schmieder, D. E. (1998). Robust, sensor-independent target detection and recognition based on computational models of human vision. Optical Engineering, 37, 2006-2021.

International Electrotechnical Commission (1999). Multimedia systems and equipment - Colour measurement and management - Part 2-1: Colour management - Default RGB colour space - sRGB (IEC 61966-2-1). Geneva: International Electrotechnical Commission.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, pp1254-1259.

Newman, R. L. (1995). Head-up displays: Designing the way ahead. Hants, England: Ashgate Publishing.

Ververs, P. M. and Wickens, C. D. (1998). Head-up displays: Effects of clutter, display intensity, and display location on pilot performance. The International Journal of Aviation Psychology, 8, 377-403.

Weintraub, D. J. and Ensing, M. (1992). Human factors issues in head-up display design: The book of HUD. CSERIAC State-of-the-art report SOAR 92-2.

Wilson, H. R. (1991). Psychophysical models of spatial vision and hyperacuity. In D. M. Regan (Ed.) Spatial Vision (pp64-86). New York: MacMillan Publishers.

Wilson, H. R. and Gelb, D. J. (1984). Modified line-element theory for spatial-frequency and width discrimination. Journal of the Optical Society of America, 1, 124-131.

Wyszecki, G., and Stiles, W. S. (1982). Color science: Concepts and methods, Quantitative data and formulae, (2nd ed.). New York: John Wiley & Sons.

EFFECTS OF HEAD-UP DISPLAY AIRSPEED INDICATOR AND ALTIMETER FORMATS ON PILOT SCANNING AND ATTENTION SWITCHING

Miwa Hayashi and Charles M. Oman
Massachusetts Institute of Technology, Cambridge, MA, USA

Michael Zuschlag
John A. Volpe National Transportation Systems Center, Cambridge, MA, USA

ABSTRACT

The effects of the rotating pointers and gradation marks of head-up display (HUD) airspeed indicator (ASI) and altimeter symbology formats were examined. The effects of the gradation marks were of special interest, as being able to remove them would help reduce display clutter. The three formats examined included: rotating pointers with gradation marks, rotating pointers without gradation marks, and digits only. The pilots' eye-movement data collected during flight simulations indicated significant changes in both ASI and altimeter fixation durations between the rotating-pointer formats and digits only, but no difference between the rotating-pointer formats themselves. However, the differences between them were found in the vertical speed indicator fixations and the flight task strategies estimated by Hidden Markov Model analysis. Results provided first empirical support for the potential value of the gradation marks.

Keywords: Aircraft display; display clutter; eye movements; Hidden Markov Model (HMM).

INTRODUCTION

A head-up display (HUD) is a transparent display that provides flight information in the pilot's primary field of view, superimposed on the outside scene. The present study examined the effects of the rotating pointers and gradation marks of HUD airspeed indicator (ASI) and altimeter symbology formats. The rotating pointers are known to provide a certain degree of motion cue in peripheral vision and, by formulating expectancy of the displayed values, to facilitate quicker instrument reading (Senders, Webb, & Baker, 1955). The value of the gradation marks, however, has not been well understood. If their contribution is small, eliminating the gradation marks may become a valid option to reduce display clutter and the potential occlusion of the outside scene.

Prior to developing a military standard for HUD symbology, the US Air Force (USAF) had conducted a flight simulator study to investigate various HUD ASI/altimeter formats (Ercoline & Gillingham, 1990; Weinstein, Gillingham, & Ercoline, 1994). They found no difference between the rotating pointer formats with and without gradation marks in terms of the RMS airspeed or altitude error or subjective ratings, although both rotating-pointer formats did better than the other formats they tested, including digits-only and vertical-tape formats. The resulting military standard (MILSTD, 1996) requires the rotating pointer format to be used for ASI/altimeter symbology. The military standard also requires the gradation marks to be present, as they are still believed to provide additional advantages, despite the negative findings of the USAF study.

The present study reinvestigated the value of the rotating pointers and the gradation marks by examining pilots' scan and attention patterns, in addition to their performance and preferences. The study was conducted as part of the effort to develop civil HUD design guidelines. Three ASI/altimeter formats similar to the ones used in the USAF study were compared (Figure 1): rotating pointers with gradation marks and digits readout (PGD), rotating pointers with digits readout but no gradation marks (PD), and digits only (D).

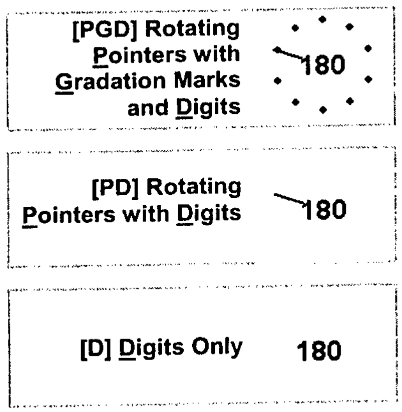


Figure 1. Three ASI/altimeter formats

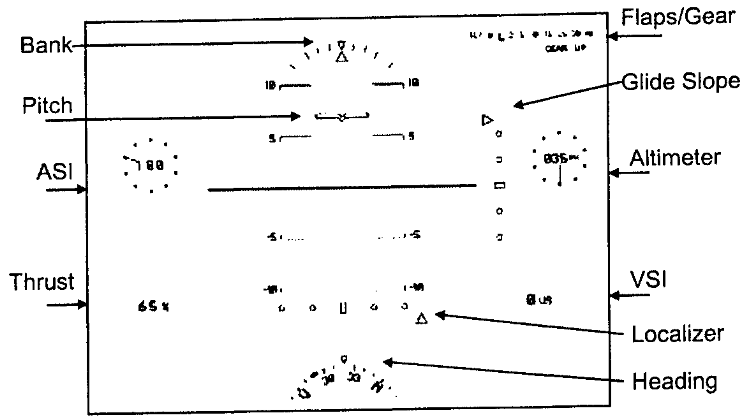


Figure 2. HUD symbology (with PGD)

METHOD

Pilot Participants

Six airline transport pilots, including 3 captains and 3 first officers, participated in the study. The pilots' total flight time as of the date of the experiment ranged from 4000 to 17500 hours. One of the captains had previous experience flying approaches with HUD-equipped aircraft.

Flight Simulation

A fixed-base flight simulator configured with Boeing 737-400 flight dynamics was used. The HUD symbology (Figure 2) was projected on a screen approximately 180 inches from the pilot's eyes. The symbology was depicted in bright green on a black background. The projection area subtended a visual angle of 21° horizontally and 16° vertically.

In the ILS simulation scenario, the aircraft was initially positioned at either side of the localizer course at an intercept angle of about 25°. Each approach had five segments: (i) straight and level at 3500 ft, 180 knots; (ii) constant-airspeed descent at 180 knots to 2000 ft; (iii) straight and level at 2000 ft, gear down and flaps lowered to approach configuration, slow to 150 knots; (iv) level turn to intercept the localizer at 2000 ft, 150 knots; and (v) final descent along the glide path to 1000 ft at 150 knots. Data collection ended when the aircraft passed 1000 ft, but the flight continued until reaching the decision height (370 ft), and then the pilot initiated a go-around. The flight segment lengths were (i) 2.3, (ii) 4.5, (iii & iv combined) 5.5, and (v) 3.2 nautical miles, respectively. Each approach took approximately 7 minutes to complete.

Data Collection

The pilots' eye-movement data were collected with a head-mounted eye camera (RK-726PCI/RK-620PC, ISCAN, Inc., Burlington, MA) and a magnetic head tracker (InsideTRAK, Polhemus, Colchester, VT) at the rate of 60 Hz. Flight variables were recorded at 1 Hz. In addition, the pilots' verbal reports of their current intentions or attitude indicator readings (i.e., "pitch" or "bank") were recorded on videotape.

Each pilot flew 9 data-collection approaches, 3 approaches for each format in balanced order. Before the data collection approaches, each pilot received a briefing and made several practice approaches. After all the approaches were completed, pilots were asked to provide their subjective preference between each pair of symbology formats by marking them on a continuous preference scale (Figure 3).

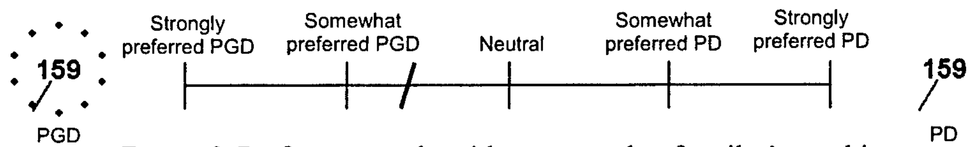


Figure 3. Preference scale with an example of a pilot's marking for the preference between "PGD vs. PD."

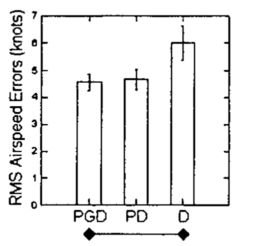


Figure 4. Grand means and standard errors of RMS Airspeed Error. Diamonds connected by a line indicate a significant difference between the two formats ($p < 0.05$) computed by pairwise comparison.

RESULTS

Root Mean Square Airspeed and Altitude Errors

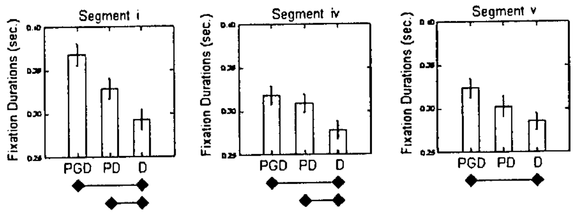
Root Mean Square (RMS) airspeed error was computed for segments (i), (ii), (iv), and (v) from the assigned airspeeds, 180, 180, 150, and 150 knots, respectively. A generalized linear model (GLM) repeated measures analysis (SYSTAT 10, SPSS, Inc.) was applied, with the main effect variables being Segment, Format, and Trial Block (block 1 included the first three approaches, block 2 the second three, and block 3 the last three). The results showed that the airspeed error was significantly reduced when PGD was used compared to when D was used ($df = 2$, $F = 4.167$, $p = 0.048$). Figure 4 plots the grand means of all pilots for each format. The result is consistent with that of the USAF study, although the difference between PD and D did not reach statistical significance in this study.

RMS altitude deviation was also computed for segments (i), (iii), and (iv) from the assigned altitudes, 3500, 2000, and 2000 ft, respectively. The same GLM repeated measures analysis was performed. Unlike in the USAF study, no significant format effect was found in this study.

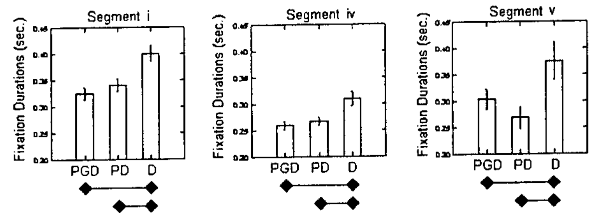
Fixation durations on each HUD symbology were computed from the eye-movement data. Due to positively skewed distributions, the values of durations were transformed by taking natural logarithms. Since each format had a different number of fixations (i.e., "unbalanced" data), mixed regression repeated measures analysis (SYSTAT 10, SPSS, Inc.) was applied instead of a GLM. The main effect variables were Format and Trial Block. Analyses were performed for each flight segment. Figure 5 shows the grand means of fixation durations on the ASI, altimeter, and vertical speed indicator (VSI) and pairwise comparison results in selected flight segments: (i) straight and level, (iv) level turn to intercept the localizer, and (v) final descent. As seen in Figure 5, the fixation durations on the ASI and altimeter showed opposite trends; the durations on the ASI tended to be longer when PGD or PD was used than when D was used, while those on the altimeter tended to be shorter when PGD or PD was used than when D was used. No difference between the two rotating-pointer formats (PDG and PD) was found in the ASI and altimeter fixations. However, a difference between them appeared in the fixations on VSI; the durations on the VSI tended to be longer when PD was used than when PGD or D was used.

Symbology Fixation Durations and Look Rates

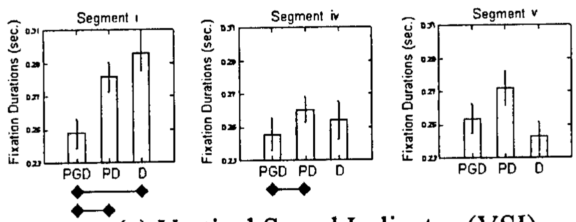
The GLM repeated measures analysis with the main effect variables being Segment, Format, and Trial Block also revealed significantly higher VSI look rates (i.e., the frequency of visits per second) when PD or D was used than when PGD was used ($df = 2$, $F = 5.867$, $p = 0.021$).



(a) Airspeed Indicator (ASI)



(b) Altimeter



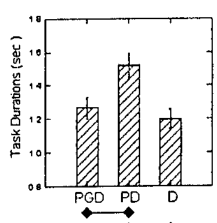
(c) Vertical Speed Indicator (VSI)

Figure 5. Grand means and standard errors of fixation durations (the values before taking logarithms) on (a) ASI, (b) altimeter, and (c) VSI in segments (i), (iv), and (v). Diamonds connected by a line indicate a significant difference between the two formats ($p < 0.05$) computed by pairwise comparison.

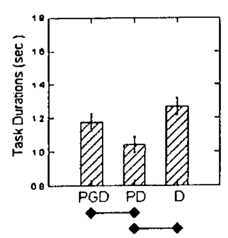
Flight Task Durations (HMM Analysis)

During instrument flight, pilots usually have several "sets" of instruments to crosscheck together—vertical-tracking instruments (pitch, altimeter, VSI, and glide slope), horizontal- tracking instruments (bank, heading, and localizer), and airspeed-tracking instruments (pitch, ASI, and thrust). A Hidden Markov Model (HMM) based analysis tool has been proposed by Hayashi, Oman, & Zuschlag (2003) to compute the pattern in pilots' scanning, or the sequence of pilots' attention switching among these tracking tasks, from pilots' eye-movement data.

HMM analysis was applied to the eye-movement data from flight segment (iv). The pilots' verbal reports in segment (iv) were used to train the HMM. Analysis showed that the durations on the vertical-tracking task were significantly longer and those on the airspeed-tracking task were significantly shorter when PD was used than when PGD or D (Figure 6).



(a) Vertical-Tracking Task



(b) Airspeed-Tracking Task

Figure 6. Grand means and standard errors of (a) vertical-tracking task durations and (b) airspeed- tracking task durations. Diamonds connected by a line indicate a significant difference between the two formats ($p < 0.05$) computed by pairwise comparison.

Pilots' Preference Rankings

The positions of the pilots' markings on the preference scales (Figure 3) were converted to preference scores by measuring the distance from the opposite side of the scale. The scores of the same format were added within each pilot and ranks were assigned (3 for the most preferred, 2 for the second most, and 1 for the least preferred). The rank sum of all pilots indicated that the most preferred format was PGD (rank sum = 16), the second most preferred was PD (12), and the least preferred was D (8) (Friedman test statistic = 5.33, $df = 2$, $p = 0.070$).

DISCUSSION

When PGD or PD was used, the fixation durations on the ASI increased and the RMS airspeed error decreased, compared to when D was used. The increased fixation durations indicate that reading the rotating-pointer movements took extra fixation time, but the pilots could effectively utilize the information to reduce airspeed error.

The fixation durations on the altimeter, on the other hand, decreased when PGD or PD was used, compared to when D was used. In some segments, the fixation duration means of PGD and PD were even shorter than that of the ASI in D format. However, one should be aware that the altitude tends to move faster than the airspeed, and, thus, even a relatively short fixation may have been sufficient to observe the rotating pointer motions. In addition, most flight segments in this experiment were defined by the altitude rather than the airspeed, such as "straight and level." Thus, the pilot may have perceived the altitude as more important than the airspeed. Therefore, it was possible that the pilots took more fixation time to read the altitude and its movement than the airspeed when D was presented.

An interesting difference between PGD and PD appeared in the VSI fixations, and this may help in understanding the value of the gradation marks. When PD was used, the VSI look rates increased and the fixation durations also increased. This may imply that the pilots did not utilize much altitude rate information with PD despite the presence of the rotating pointers. HMM analysis also showed that the pilots spent more time on the vertical-tracking task when PD was used, possibly as the result of the increased fixation demands on VSI, and that extra time was taken away from the airspeed-tracking task. Although the RMS airspeed and altimeter error levels stayed about the same between PGD and PD, this strategy change may have caused the pilots' slight preference for PGD over PD.

CONCLUSION

The rotating pointers (PGD and PD) resulted in smaller airspeed error and higher pilot preference ratings. These results were consistent with the USAF study. Unlike their study, this study did not find any significant format effect in the altitude error.

In addition, the eye-movement data analysis provided further insights into the effects of the rotating pointers and gradation marks. For both the ASI and the altimeter, significant changes in the fixation durations were observed between the rotating-pointer formats (PGD and PD) and digits-only format (D). The results, combined with the RMS airspeed and altitude error findings, confirmed that the rotating pointer formats provide superior scanning efficiency over the digits-only format. The differences between PGD and PD were found in the VSI fixation patterns and the vertical- and airspeed-tracking task durations estimated by the HMM analysis. The increased attentional demand for the vertical-tracking task when PD was used may explain why the pilots slightly preferred PGD over PD. The results provide empirical support for the common belief in the potential advantages of gradation marks.

ACKNOWLEDGEMENTS

The authors would like to express thanks to Dale Dunford of the FAA ANM-111 and Tom McCloy of AAR-100 for supporting this work. Great appreciation is also expressed for the pilots who volunteered their time to participate in this study. Thanks also to Andrew Kendra of the Volpe Center and Rikki Razdan of ISCAN, Inc. for technical assistance, to Tao Yue of MIT for OpenGL programming, and to Carl Quesnel for editorial assistance. The research presented in this paper was supported and funded by the FAA through the Volpe Center under DTRS57-01-D-30043.

REFERENCES

Ercoline, W. R., & Gillingham, K. K. (1990). Effects of variations in head-up display airspeed and altitude representations on basic flight performance. Proceedings of the Human Factors Society 34th Annual Meeting, Oct. 8-12, Orlando, FL, 1547-1551.

Hayashi, M., Oman, C. M., & Zuschlag, M. (2003). Hidden markov models as a tool to measure pilot attention switching during simulated ILS approaches. Proceedings of the 12th International Symposium on Aviation Psychology, Apr. 14-17, Dayton, OH, 502-507.

MILSTD (1996). MIL-STD 1787B Aircraft military symbology: United States Air Force.

Senders, J. W., Webb, I. B., & Baker, C. A. (1955). The peripheral viewing of dials. Journal of Applied Psychology, 39, 433-436.

Weinstein, L. F., Gillingham, K. K., & Ercoline, W. R. (1994). United States Air Force head-up display control and performance symbology evaluations. Aviation, Space, and Environmental Medicine, 65(5), A20-A30.

SYNTHETIC VISION SYSTEM DISPLAY GUIDANCE, INTEGRATION, AND VISIBILITY EFFECTS ON FLIGHTPATH TRACKING, SITUATION AWARENESS, AND MENTAL WORKLOAD

Amy L. Alexander, Christopher D. Wickens, & Thomas J. Hardy
University of Illinois, Aviation Human Factors Division

ABSTRACT

The current research was designed to examine the effects of display guidance (tunnel/datalink commands), integration (instruments overlaid/separate), and outside world visibility (simulated VMC/IMC) on flightpath tracking, situation (traffic) awareness, and mental workload within the context of a synthetic vision system. Fourteen pilots flew a series of eight curved approaches over rugged terrain in a high-fidelity simulator. The results revealed that flightpath tracking and traffic detection performance was superior while flying with the tunnel compared to implementing datalink commands. Mental workload was also rated lower in the tunnel than in the datalink condition. While an overlaid instrument panel slightly benefited vertical flightpath tracking compared to the separated condition, this was offset by a six-second cost to traffic detection. Outside world visibility did not mediate these findings. The practical applications of this work include recommending guidance and display layout configurations in the design of synthetic vision systems, as well as informing the human factors community of the effects of these configurations on hazard awareness and mental workload.

Keywords: synthetic vision, displays, guidance, flightpath tracking, traffic awareness

INTRODUCTION

Synthetic Vision Systems (SVS) are being developed for the display of information needed by the pilot in order to safely and efficiently navigate under challenging-terrain or low-visibility conditions (Alexander, Wickens, & Hardy, submitted; Prinzel, Comstock, Glaab, Kramer, & Jarvis, 2004; Schnell, Kwon, Merchant, & Etherington, in press; Scott, 2001; Stark, Comstock, Prinzel, Burdette, & Scerbo, 2001; Williams, 2002). Such systems are specifically designed to increase **situation awareness** of terrain and possibly other hazards (i.e., traffic, weather) with the primary objective of reducing controlled flight into terrain (CFIT; Wiener, 1977) accidents. We particularly focus on maintaining situation awareness of traffic hazards in the current study.

The present research examined the effects of display guidance and integration, as well as outside-world visibility (simulated VMC vs. IMC), within an SVS context. Display guidance was either automated by providing a tunnel-in-the-sky, or was non-automated by providing datalink commands for the pilots to implement. The degree of display integration was manipulated by either superimposing the instruments on the Primary Flight Display (PFD; overlaid) or placing them in a separate panel next to the PFD (separated). At the heart of such manipulations is the tradeoff which often exists between clutter and scanning. On the one hand, superimposing information within a single display panel allows for minimal scanning demands in accessing that information, but may cause excessive clutter which could slow the retrieval of specific information needs (i.e., traffic; Yeh, Merlo, Wickens, & Brandenburg, 2003). On the other hand, separating different information databases across panels relieves clutter but necessitates increased scanning when information across panels must be integrated.

In comparing display guidance options, the presence of an automated tunnel would be expected to support superior flightpath tracking given (a) the availability of flightpath preview, and (b) the less-complex nature of simply following a tunnel versus cognitively transforming altitude, heading, and vertical descent rate datalink commands into the appropriate physical actions. In fact, previous research has indeed supported the use of a tunnel for routine flightpath maintenance (Alexander et al., submitted; Beringer, 1999; Fadden, Ververs, & Wickens, 2001; Schnell et al., in press). It remains unclear, however, how using this ego-referenced PFD to host traffic depiction will affect performance.

In terms of the clutter/scan tradeoff previously discussed in light of integrated displays, we might expect the overlaid display (i.e., instruments superimposed on the PFD) to better support traffic detection due to the overall decreased scanning demands. On the other hand, the separated display (i.e., instruments displayed in a separate panel than the PFD) might better facilitate traffic detection by reducing the amount of clutter which might obscure airborne hazards. One study (Stark, 2003) has systematically examined integrated (overlaid) displays within an SVS context and found benefits to flightpath maintenance borne by the integrated display condition. This experiment, however, did not look at the issue of traffic detection.

METHODS

Participants

Fourteen certified flight instructors (4 female, 10 male, mean age = 24 years) flew a sequence of eight flight scenarios following curved paths over rugged terrain to an airport in a high-fidelity flight simulator. The mean total flight experience was 715 hours with a mean of 111.5 instrument flight hours. All pilots were paid $8/hour for approximately 4 hours of participation.

Displays

The experiment consisted of 8 conditions broken down according to the type of guidance offered, where the instrument symbology was located, and the weather status of the outside world. Figure 1 represents the 2x2x2 design schematically. The four display suites grouped together on the left represent those conditions in which a tunnel was provided for flightpath guidance (in the upper left panel of each suite, overlaying the synthetic terrain, represented by the mountains); the suites on the right illustrate tunnel-absent conditions in which flightpath commands were issues as text within the datalink display panel. These datalink commands (i.e., headings, altitudes, vertical descent rates) were designed to mimic the information content which drove the properties of the now-absent tunnel. Necessary information to accurately follow these commands was depicted by a heading indicator and vertical situation display in the instrument panel (symbolically represented by the 2 round gauges) and the NAV display. The top row represents those cases where the instrument symbology was overlaid, or superimposed, on the PFD, while the instrument symbology was presented in a separate display panel next to the PFD in the bottom row. Within each foursome display suite, the pair on the left were encountered during IMC while the pair on the right were encountered during VMC.

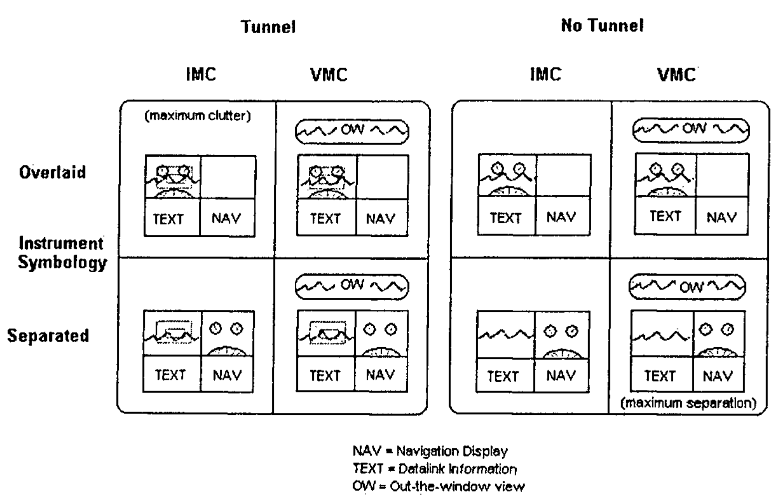


Figure 1. Schematic representations of the experimental displays.

Tasks and Experimental Design

While following the paths through use of a tunnel or by datalink commands, pilots were also required to detect periodic airborne hazards once they appeared in the computer-generated imagery of the SVS sky. All traffic was also presented in the NAV display which acted as a CDTI with broad coverage. All pilots flew 8 scenarios, one with each display suite shown in Figure 1, in a counterbalanced order.

RESULTS

Flight Performance

Analysis of the log-transformed error data revealed that both vertical and lateral flightpath deviations were greater when flying by datalink commands (vertical $\underline{M}$ = 30.0m; lateral $\underline{M}$ = 76.7m) than with the automated tunnel (vertical $\underline{M}$ = 5.49m, $\underline{F}(1, 13) = 324$, $\underline{p} < .001$; lateral $\underline{M}$ = 7.89m, $\underline{F}(1, 13) = 965$, $\underline{p} < .001$), as shown in Figure 2. The automated tunnel provides constant integrated feedback to the pilots regarding their position relative to the flightpath, and also makes available important preview information, allowing for the anticipation of upcoming turns. The only effect of overlay on flightpath tracking was seen in the vertical dimension in which an overlaid instrument panel produced a small 5 meter benefit ($\underline{F}(1, 13) = 11.3$, $\underline{p} < .01$) compared to the separated condition.

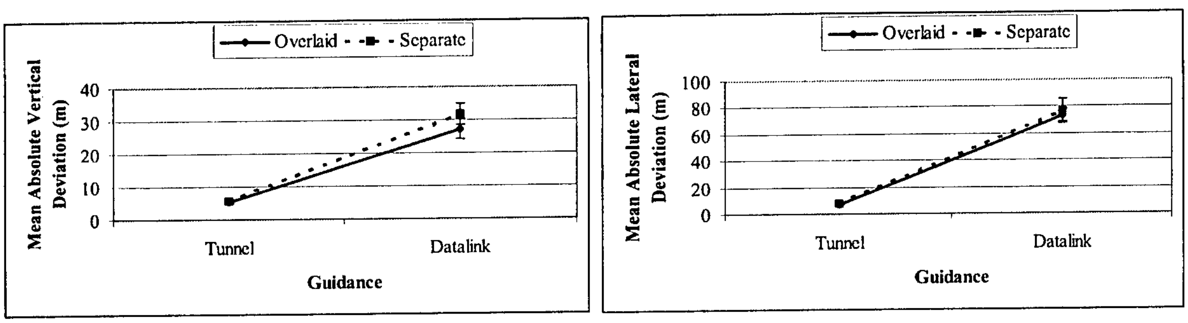


Figure 2. Flightpath deviation results collapsed across outside world visibility. Left: vertical deviations, Right: lateral deviations. Type of guidance (tunnel vs. datalink) and integration (instruments overlaid vs. separate) are represented within each graph.

Situation (Traffic) Awareness

Traffic detection times shown in Figure 3 illustrate that detection times were as much as five seconds slower when flying by datalink commands ($\underline{M}$ = 18.1s) compared with the tunnel ($\underline{M}$ = 13.3, $\underline{F}(1, 13) = 15.9$, $\underline{p} < .01$). We infer this to be a result of the increased cognitive demands imposed by the datalink condition, perhaps causing An attentional tunneling effect in which the pilot's resources are depleted by flightpath maintenance and therefore deter speedy traffic detection. In terms of instrument symbology location, traffic awareness was best supported by the separated SVS display ($\underline{M}$ = 12.6s, $\underline{F}(1, 13) = 35.0$, $\underline{p} < .001$). Traffic detection times were as much as six seconds slower with the overlaid SVS display ($\underline{M}$ = 19.0s), an effect presumably due to the effects of clutter.

Mental Workload

NASA-TLX ratings revealed that subjective mental workload was higher when flying by the datalink commands than with the automated tunnel ($F(1, 13) = 43.0$, $p < .001$). Given the nature of these flightpaths (i.e., curved, step-down approaches), the datalink commands imposed greater workload both physically and subjectively in that the pilots were required to constantly scan between the commands and their instruments to determine whether they were on the path or not.

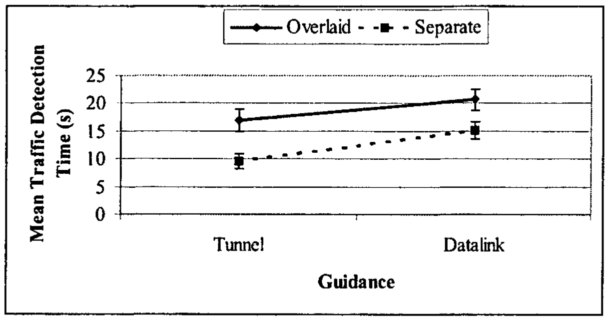


Figure 3. Traffic detection times collapsed across outside world visibility.

DISCUSSION

The goal of the current study was to examine the effects of display guidance, integration, and outside world visibility on flightpath tracking, situation (traffic) awareness, and mental workload. In accordance with previous research, the presence of a tunnel clearly benefited flightpath tracking. Importantly, the added display elements inherent to the tunnel did not disrupt traffic detection within the same panel. In fact, pilots were on average 5 seconds faster at detecting traffic when flying with the tunnel than when implementing datalink commands. Both of these performance advantages of flying the tunnel may be attributed to two factors: (1) the fact that preview was available with the tunnel such that pilots could anticipate turns, and (2) the lower cognitive demands given that pilots did not have to transform commands into actions with the tunnel as required by the datalink condition. Furthermore, in the datalink condition, pilots were required to integrate the separately-presented lateral deviations (from a flightpath on the NAV display) and vertical deviations (represented within the vertical situation display). The tunnel, on the other hand, was a single, integrated object display for which performance benefits have previously been found (Haskell & Wickens, 1993). Mental workload ratings further support this idea that flying with the datalink commands was more challenging than flying with the tunnel.

In contrast, there did appear to be a slight tradeoff in terms of integration effects on flightpath tracking and traffic detection. While the overlaid condition supported superior vertical tracking (although this was only a 5 meter benefit), traffic detection was slowed by as much as 6 seconds when compared to the separated conditions. Given the size of these effects, one might conclude that the costs of the overlaid condition to traffic detection outweigh the minor benefit to vertical tracking.

CONCLUSION

In conclusion, the results of this study continue to support the implementation of a tunnel to support flightpath maintenance. Furthermore, we have now shown that the presence of a tunnel also supports faster detection of traffic hosted on the SVS display panel when compared to a more cognitively-demanding guidance option such as issuing datalink commands. Given that traffic detection was faster in the separated as compared to the overlaid condition, especially with the interest of maintaining situation awareness of hazards in mind, it would be recommended that the instrument panel remain separate from the PFD. Other aspects of this experiment, including visual scanning, traffic change detection, and response to off –normal events, may be found in Wickens, Alexander, Hardy, Horrey, and Thomas (in preparation).

ACKNOWLEDGMENTS

This research was supported by grant NASA-NAG-1-03014, for which Dr. Bettina Beard was the scientific monitor. The authors wish to acknowledge the invaluable support of Ron Carbonari, Jonathan Sivier, Roger Marsh, and Sharon Yeakel for developing the simulation used in this experiment. Any opinions, findings, and conclusions or recommendations in this publication are those of the authors and do not necessarily reflect the views of NASA.

REFERENCES

Alexander, A.L., Wickens, C.D., & Hardy, T.J. (submitted). Synthetic Vision Systems: The Effects of Guidance Symbology, Display Size, and Field of View. Human Factors.

Beringer, D. B. (1999). Flight command-guidance indicators or pathway displays: My way or the highway? Proceedings of the 10th International Symposium on Aviation Psychology (pp. 57-62). Columbus, OH: The Ohio State University.

Fadden, S., Ververs, P. M., & Wickens, C. D. (2001). Pathway HUDS: Are they viable? Human Factors, *43*(2), 173-193.

Haskell, I.D. & Wickens, C.D. (1993). Two- and three-dimensional displays for aviation: A theoretical and empirical comparison. International Journal of Aviation Psychology, *3*(2), 87-109.

Prinzel, L.J., Comstock, J.R., Glaab, L.J., Kramer, L.J., & Arthur, J.J. (2004). The efficacy of head-down and head-up synthetic vision display concepts for retro- and forward-fit of commercial aircraft. International Journal of Aviation Psychology, *14* (1), 53-77.

Schnell, T., Kwon, Y., Merchant, S., & Etherington, T. (in press). Improved flight technical performance in flight decks equipped with synthetic vision information system displays. International Journal of Aviation Psychology.

Stark, J.M. (2003). Investigating display integration in candidate synthetic vision system displays. Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting (pp. 134-138). Santa Monica, CA: HFES.

Stark, J.M., Comstock, J.R., Prinzel, L.J., Burdette, D.W., & Scerbo, M.W. (2001). A preliminary examination of situation awareness and pilot performance in a synthetic vision environment. Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society (pp. 40-43). Santa Monica, CA: HFES.

Wickens, C.D., Alexander, A.L., Hardy, T.J., Horrey, W.J., and Thomas, L.C. (in preparation). Traffic and flight guidance depiction on a synthetic vision system display: The effects of clutter on performance and visual attention allocation.

Wiener, E.L. (1977). Controlled flight into terrain accidents: System-induced errors. Human Factors, *19*, 171.

Williams, K. (2002). Impact of aviation highway-in-the-sky displays on pilot situation awareness. Human Factors, *44*, 18-27.

Yeh, M., Merlo, J.L., Wickens, C.D., Brandenburg, D.L. (2003). Head up versus head down: The costs of imprecision, unreliability, and visual clutter on cue effectiveness for display signaling. Human Factors, *45*(3), 390-407.

Relative Height in 3D Aircraft Displays

Patrik Lif
Swedish Defence Research Agency, Sweden

Torbjörn Alm
Linköping Institute of Technology, Sweden

ABSTRACT

Our work is motivated by three considerations. First, relative height is a key factor in aviation safety, not least from the pilots' perspective. Second, in a series of experiments on 3D air-traffic displays we have shown the benefits of 3D presentation. In the experiment here we evaluate the utility of relative height in a 3D display. Third, our task is an example of a focused attention task. While there are findings suggesting that 3D displays are not suitable at all for focused attention tasks, our hypothesis was that this suitability may be dependent on the cues in the display design. To address these three points, three different cue alternatives were investigated in scenarios with own-ship and targets on a 3D pictorial air traffic display. We found that the addition of more elaborated cues (depth cues and height cues) significantly improved pilots' assessments of relative height. This means that more nuanced design guidelines could be used.

Keywords: 3D displays; relative height; monocular depth cues; monocular height cues; spatial relations

INTRODUCTION

Earlier research investigating the applicability of 3D aircraft or traffic control displays has been carried out in static scenarios (Andersson & Alm, 2003; Mazur & Reising, 1990; McGreevy & Ellis, 1986). One obvious reason for this was that the computer (PC) capacity at the time did not allow dynamics to any large extent. This restriction was later eliminated, which made it possible for any research lab to go into dynamics. Since aviation is very dynamic it is quite natural to mirror this fact in experimental settings. It is also important to emphasize that the time aspect is embedded in dynamic experiments, while not in static settings. Time is important in investigations including Situation Awareness (SA) since SA is built up over time (Endsley, 1995). Analogous, this is also true for investigations where elements of SA (like relative height) are in focus. With this background it is our opinion that experimental studies carried out in static scenarios need to be replicated in dynamic settings in order verify the conclusions made. In our own research we have found differences of such magnitude between the two approaches that we have reconsidered our own results from static scenarios (Alm, Andersson, & Öberg, 2003). These differences mainly refer to the subjects' difficulty in understanding the situation as a whole if they only had a static view of the situation, while in the dynamic experiments the results were much improved indicating better understanding.

In our research the main concept for measuring the understanding of display content (3D pictorial) was to let subjects assess various spatial relations, that is, elements of SA (Endsley, 1995). In this experiment differences in height between own-ship and other objects (aircraft) were investigated.

In our series of experiments we have used three measures as dependent variables:

- 3D bearing between own-ship and target symbols
- Estimation of future point of collision
- Relative height between own-ship and other aircraft.

The first two are examples of integrated measures (and tasks), while the last measure only focuses on one dimension, height relations. Subsequently, the dependent measure changed from an absolute, metric to a relative, non-metric estimate. The reasoning behind this change was that there is evidence for that 3D displays do not support "focused attention tasks" (Haskell & Wickens, 1993) such as assessing "at what distance" or "at what height". In our case the corresponding question was "which difference is the closest or most distant". The statement from Haskell & Wickens was based on a comparative study of 2D and 3D displays. Since the 3D display design was not manipulated we found it interesting to do so in order to have a more nuanced opinion on the feasibility of 3D displays in focused attention tasks. The change from absolute measures to relative was motivated by the difficulties in absolute distance estimations in 3D displays, but this also corresponds with real flight situations, where height differences are more important to the pilot than the absolute altitudes, and where even relative heights may be hard to assess using the 2D displays of today.

METHOD

Design

A within subject design with three experimental conditions was carried out. The conditions were Sphere-alone, Sphere with drop-line, and Sphere with cone.

Subjects

28 subjects participated voluntary, receiving a small token fee in return, 14 male and 14 female. The subjects were undergraduate students at the University of Linköping with no pilot education. All subjects had normal color vision and normal or corrected to normal vision acuity.

Apparatus

The experiment was carried out in the generic vehicle simulator at the VR laboratory at Linköping Institute of Technology. The perspective display was shown head-down on a fifteen-inch color LCD computer display with a resolution of 1024 x 768 pixels, and at a distance of about 70 cm from the subject. A keyboard was added in front of the subjects for answer input. The environment projection system was closed down since the experimental focus was to evaluate a tactical head-down display. The simulator allows for manual or automatic piloting. In this case with non-pilots as subjects the own-ship flight was automatic.

Stimuli

The ground in the perspective display was a lit and shaded gray topographic landscape. A black "north-in" oriented grid was added on top of the terrain to enhance the linear perspective. Each grid square covered approximately 250 m^2 of terrain. Sky and sea were presented in different blue colors.

The FOV of the perspective display was 80° horizontally and 60° vertically. The viewpoint was 618 meters above the own aircraft location at a distance of 2000 meters. The aspect angle was 18° looking down which placed the own aircraft symbol in the centre of the display. Figure 1 illustrates the position of the viewpoint relative to the own-aircraft.

The color of the own-ship symbol was white and the target symbols were orange. At a certain moment in the scenario three target symbols were highlighted by changing color to yellow, green and red. Three design alternatives were investigated. In all of them spheres were used as symbol shapes, which is consistent with one conclusion from earlier studies (Andersson & Alm, 2003). The sphere shape does not change depending on viewing angles or headings and thereby was best identified among a set of other simple symbol shapes. Another motivation is that size could be used as a depth cue using the same nominal size for all objects. What were varied in the design alternatives were additional cues for height and depth estimation.

Three display alternatives were evaluated. One contained sphere symbols with no additional attributes (Figure 2). In the second design alternative spheres with drop-lines were used. This design included a horizontal tic mark indicting the own-ship altitude applied to the target symbol drop-line (Figure 3). The third alternative had sphere symbols with a transparent reference surface in grey through the own-ship level and blue transparent cones between this plane and the target symbols. The cones were oriented with the tops towards the target spheres and the bases towards the transparent reference surface. The cones had equal base diameters, which consequently meant varying cone angles depending on relative altitude to the plane (Figure 4). Compared with the display format used in the experiment, the three figures below are cropped under the horizon, which was visible through all scenarios. The figures are also cropped below the own-ship symbol.

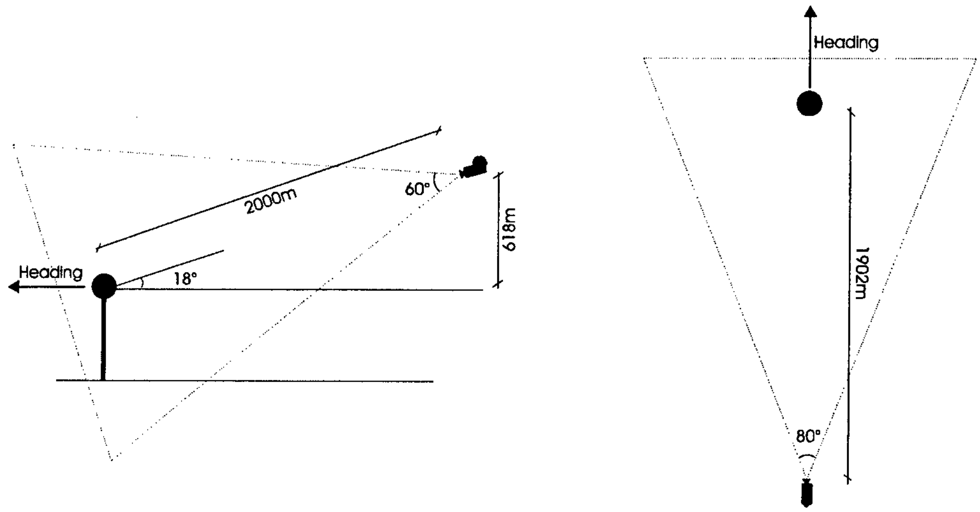


Figure 1. Horizontal and vertical views of viewpoint and own-ship relations.

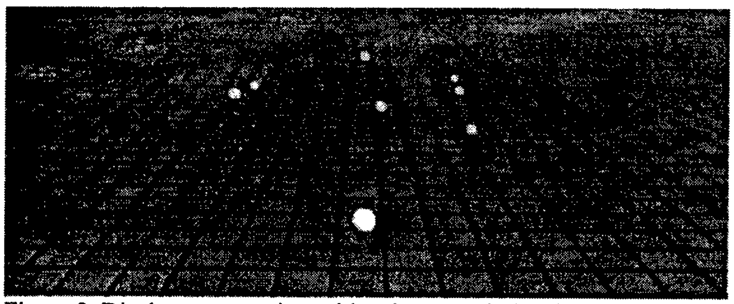


Figure 2. Display presentation with sphere symbols with no additional information (experimental setting #1).

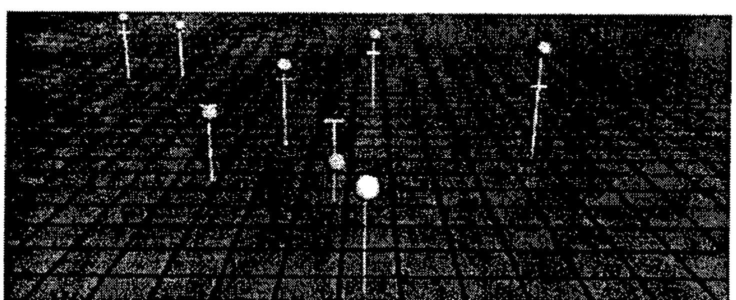


Figure 3. Display presentation with sphere symbols and drop-lines (experimental setting #2).

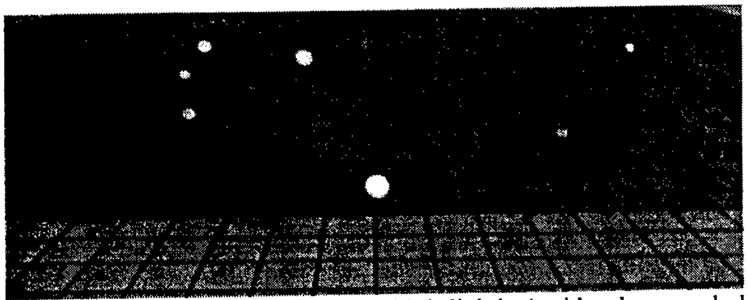


Figure 4. Display presentation (after high-lighting) with sphere symbols, transparent referencesurface, and cones (experimental setting #3).

The scenarios used in this experiment consisted of 7 target symbols and the own-ship symbol. The own-ship had constant course (north), altitude (3000 m), and speed. The targets had constant courses, constant but different altitudes and the same constant speeds (1.6 times own-ship speed). The target courses were chosen to keep the symbols within the display frame through each whole scenario. 16 different scenarios were developed and applied to each of the three design alternatives, all together 48 scenarios for each subject.

In order to have balanced target appearances and a variation in heights, the following scenario "rules" were inserted: The virtual space was divided into 13 levels, one at the own-ship altitude and 6 levels above and 6 levels below. The targets appeared at levels above and below in a balanced way and within these two sectors the appearances at the specific levels were randomly distributed.

The sequence of each scenario followed the same scheme: 1. Own-ship and target symbols are present, 2. After 8 seconds a question appeared on the top of the display, either "most distant" or "closest" which means either the target at the closest or the nearest height with reference to the own-ship altitude, 3. After two more seconds three target symbols started to twinkle in 10 Hz for 0.5 seconds and then changed to yellow, green, and red, respectively. The not highlighted target symbols remained orange. 4. Answering by pressing one of the three color coded buttons as quickly as possible after answer decision. Maximum response time was set to 5.5 seconds. 5. After response or maximum response time, next scenario was started. All 48 scenarios were carried out in one sequence for each of the three experimental conditions.

Procedure

The subjects were given verbal instructions about the purpose of the experiment, the three design alternatives, the subjects' roles and activities in the experiment before the experiment was started. The subjects were also shown how the response tool was operated before doing the training round under supervision. The subjects were told to prioritize correctness over fast response.

The experiment took between 90 and 120 minutes for each subject. This included instructions, one training session, three experiment sessions and five minutes of rest between the sessions.

RESULTS

Comparisons were made between the tree conditions with number of 'correct answers' and 'time to answer' as depending variables.

One-way Anova showed a significant effect of number of correct answers, $F(2, 54) = 235.7$, $p < .0001$. Tukey's HSD Post Hoc test showed that the Sphere-alone condition was more difficult than both Sphere with drop-line ($p<.005$) and than Sphere with Cone ($p< .005$), as shown in Figure 5. There was no significant difference between Sphere with drop-line and Sphere with cone ($p>.05$).

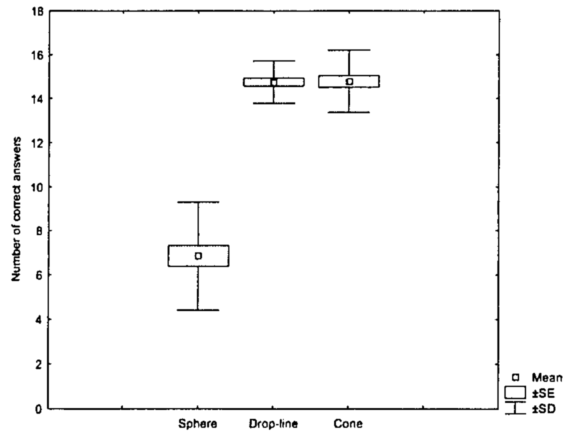


Figure 5. Number of correct answers for the three experimental conditions, Sphere, Sphere with cone, and Sphere with drop-line.

These results clearly shows that the subjects could solve the task almost perfect with additional cues (maximum number of correct answers are 16), while with no additional cue the number of correct answers were close to chance.

One-way Anova showed a significant effect of number of time to answer, $F(2, 54) = 54.9$, $p<.0001$. Tukey's HSD Post Hoc test showed that the Sphere-alone condition took longer time so answer than both Sphere with drop-line ($p<.005$) and than Sphere with Cone ($p<.005$), as shown in Figure 6. There was no significant difference between Sphere with drop-line and Sphere with cone ($p<.05$).

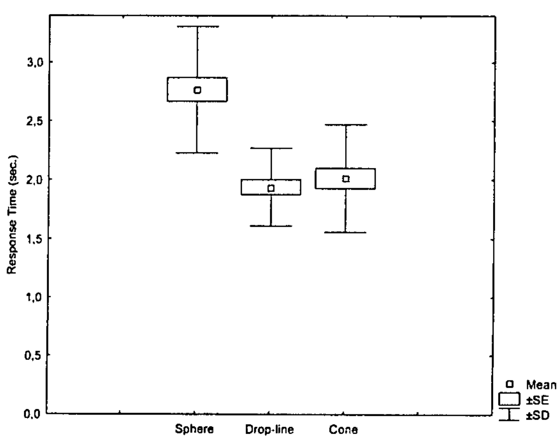


Figure 6. Time to answer for the three experimental conditions, Sphere, Sphere with cone, and Sphere with drop-line.

The results of response time clearly strengthen the picture of the usability of using additional cues in this focused attention task. From a practical point of view, one second is of significant importance in the aviation area. However, the most important part in this analysis is correctness, which has a strong coupling to aviation safety.

DISCUSSION

These results point on a need for additional cues like drop-lines with horizontal tic marks or cones in order to support relative height estimations. It is interesting that there are no differences in assessment results between the two additional cue alternatives despite the obvious difference in design concepts. The logical choice should be to recommend the simplest solution (drop-line) to minimize cluttering.

The necessity to insert additional cues like drop-lines contradict the results from our own studies with integrated measures (3D bearings, collision points, where no additional cues were needed) but are very much in line with other research (Ware, 2000). Additional cues seem to be necessary in focused attention tasks. These results prompt to following recommendation:

If you have an integrated task using a 3D display and must change to a focused attention task it will be enough to add cues to the 3D display. There is no need to change display format.

This should carry benefits from an operator perspective since the change between 2D and 3D formats is mentally demanding. These demands could be referred to perceptual problems in the mapping procedure between objects presented in different formats (displays). The heuristic of "visual momentum across displays" describes techniques to overcome such problems (Woods, 1994), but obviously the most effective solution must be not to change format at all.

A concluding subject for discussion is that metric measures in one single dimension (a focused attention task) of the three dimensional space could be problematic because of distortion. This distortion differs with the other two dimensions (Lind, Bingham, & Forsell, 2003; Todd, Tittle, & Norman, 1995), in this case the x- and y-dimensions. The distortion problem emanates from the use of different scales along the axes and also with the chosen field of view. Even if there is no manipulation of scales, the field of view problem exists as long as more than one dimension is included in the 3D format (Smallman, Manes, & Cowen, 2003). The distortion should be important also with non-metric measures, as in this experiment. However, this was not further analyzed but could be of interest in future research activities.

ACKNOWLEDGEMENTS

This research was funded by the Swedish Aviation Research Program (NFFP) and the Swedish Defence Material Administration. Special thanks to our research colleagues Kip Smith, Sidney Dekker, and Lars Eriksson for valuable advice.

REFERENCES

Alm, T., Andersson, P., & Öberg, M. (2003). Evaluation of Monocular Depth Cues in 3D Aircraft Displays. In Proceedings of International Symposium on Aviation Psychology 2003.

Andersson, P., & Alm, T. (2003). Perception Aspects on Perspective Aircraft Displays. Displays, 24, 1-13.

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors, 37(1), 32-64.

Haskell, I. D., & Wickens, C. D. (1993). Two- and Three-Dimensional Displays for Aviation: A Theoretical and Empirical Comparison. The International Journal of Aviation Psychology, 3(2), 87-109.

Lind, M., Bingham, G. P., & Forsell, C. (2003). Metric 3D Structure in Visualizations.

Mazur, K. M., & Reising, J. M. (1990). The relative effictiveness of Three Visual Depth Cues in a Dynamic Air Situation Display. In Proceedings of Human Factors and Ergonomics Society 41st annual meeting.

McGreevy, M. W., & Ellis, S. R. (1986). The Effect of Perspective Geometry on Judged Direction in Spatial Information Instruments. Human Factors, 28*(4)*, 439-456.

Smallman, H. S., Manes, D. I., & Cowen, M. B. (2003). Measuring and modelling the missinterpretation of 3-D perspective views. Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting 2003, 1615-1619.

Todd, J. T., Tittle, J. S., & Norman, J. F. (1995). Distortions of three-demensional space in the perceptual analysis of motion and stereo. Perception, 24, 75-86.

Ware, C. (2000). Space Perception and the Display of Data in Space, Information Visualization Perception for design (pp. 273-310). San Diego: Academic Press.

Woods, D. D. (1994). Visual Momentum: A concept to improve the cognitive coupling of person and computer. International Journal of Man-Machine Studies, 21, 229-244.

APPROACH FOR DESIGNING AN INTEGRATED AIR TRAFFIC CONTROL WORKSTATION AROUND THE NEEDS OF CONTROLLERS

Stephanie Kreseen-Imbembo, Steve Fadden, and Paul Mafera
Booz Allen Hamilton

ABSTRACT

Over the next several years, new tools and capabilities will be introduced to the en route environment to enable sector controllers to deal with projected increases in air traffic. En route sector controllers will use new procedures, tools, and systems to provide separation assurance and maintain traffic flow. Decision support tools and automated systems will assist controllers in performing tasks such as spacing, communications, and maintaining awareness of traffic constraints. The introduction of such capabilities to the en route environment will likely affect the manner in which sector controllers manage information to control traffic. To ensure that controllers will receive suitable information to support their tasks and maintain situation awareness, it is necessary to explore how new system components may be integrated into the existing environment prior to deployment. By understanding how en route sector controllers perform their tasks and use information in the present, new systems can be developed to support controller information and task requirements in the future. This paper will present one approach for gathering information about the current tasks of en route controllers and assessing how the introduction of new systems and procedures will affect such tasks in the future.

Keywords: Air Traffic Control; User Interface; Display Integration

INTRODUCTION

Despite a temporary decline in air traffic volume since the events of September 11, 2001, air traffic has been on the rise, and the FAA anticipates traffic growth to continue beyond pre-September 11 levels over the next several years (Aerospace Forecasts, FAA, 2003; NAS Operational Evolution Plan, FAA, 2002). As traffic continues to rise in the National Airspace System (NAS), new systems and equipment will be introduced to increase the safety and efficiency of NAS service providers as they provide services to greater numbers of aircraft. Many of these new systems are targeted at the Air Route Traffic Control Center (ARTCC) environment, where en route air traffic controllers provide separation assurance and efficient flow management to aircraft crossing through en route sectors of airspace. Although new systems, tools, and equipment have the potential to enable controllers to perform tasks with greater effectiveness and lower workload, they may also have an opposite effect if their introduction is poorly planned and not integrated with controller task and information requirements.

This paper presents a process for gathering information about the current tasks performed by en route controllers, validating the steps and information required for controllers to perform their tasks, and assessing how new systems and tools can support these tasks. In addition, we examine the potential costs and benefits associated with integrating multiple systems and components to reduce the total number of displays, interfaces, input devices, and hardware components associated with en route sector controller workstations. The need for such an approach stems not only from the introduction of new systems and displays to the en route environment, but also the current use of multiple displays and input devices in the en route environment. The approach described here enables engineers to take into account the current baseline of en route operations, systems, tasks and information needs, in an effort to develop systems that better support en route controllers through intelligently integrating new and existing functionalities into the future workstation user interface.

METHOD AND DISCUSSION

A six-step process was used to understand how en route controllers currently perform their tasks, and how modifications to the controller workstation interface can be made to continue supporting tasks as new systems are introduced. The steps of this process included the following:

1. Conducting a literature review of human factors (human machine interface and task performance) issues associated with en route controller performance, including the identification of existing en route sector controller task flow models
2. Observing operational en route sector controllers to validate task flow models and elicit additional human machine interface issues
3. Creating task flow diagrams to capture the steps involved in the performance of controller tasks
4. Associating task flow steps with the information needs and information sources required to complete the task
5. Conducting a feasibility assessment to map controller information needs supported by information sources in the current system to information sources proposed in new systems
6. Developing interface prototypes to support controller tasks and serve as a basis for eliciting feedback from en route controllers and other FAA stakeholders.

The initial literature review was performed to identify human factors concerns associated with en route tasks and the current tools and equipment used by controllers. The review process also facilitated the understanding and identification of frequent and critical controller tasks, including the results of cognitive task analyses (Redding, Cannon, Lierman, Ryder, Seamster, Purcell, 1992) and tasks identified in operational concepts (ATS Concept of Operations, FAA, 1997). Overall, a total of ten primary tasks were identified for further validation, including monitoring airspace, monitoring traffic for separation, resolving traffic conflicts, transferring flights to the next sector, and receiving approaching flights.

This task analysis work was used as a basis for the formation of baseline task analyses for en route controllers in the current operational environment. Because some equipment and decision support tools have been modified or added to the en route environment since the task analyses were performance, we conducted observations of en route controllers to validate and modify the baseline task analyses to reflect the current operational environment. Air traffic controllers and other center personnel were observed at two different air route traffic control centers (ARTCC) to gain a better understanding of how tasks are performed in the current environment. The knowledge gained from the observations was used to validate the steps within each task and identify information needs required by controllers to complete the task steps. Researchers also used the observation sessions as an opportunity to identify interface issues that may be addressed through future design and integration activities, and to discuss potential design solutions with controllers during their breaks.

The data collected through the literature review and task observations were used to identify task goals and create final task flow diagrams for en route sector controllers. ARTCC observations provided researchers with a working knowledge of how sector controllers perform their tasks in the context of the current NAS environment, with contemporary tools and information sources. In general, the task steps performed by controllers in the current operational environment were found to be similar to the steps performed in previous environments, although some aspects of the tasks (especially the sources from which controllers received their information) had changed. These changes primarily reflected differences associated with the removal of Flight Progress Strips (FPS), the addition of the User Request Evaluation Tool (URET), and the addition of traffic restriction and flow information through the Enhanced Status Information System (ESIS) Status Indicator Area (SIA) and the ESIS Traffic Situation Display (TSD).

The final task flow diagrams that were developed are similar to the example shown in Figure 1 below. (Note that the diagram in Figure 1 is provided as a notional example only, and is not a direct result of the current study.) In each task flow, individual task steps were assigned (through the use of swim lanes) to the controller position commonly responsible for performing the task: R-side, D-side or Tracker. This assignment of steps to different controller positions facilitated the identification of each controller's primary tasks, responsibilities, and information needs, in the event that developing position-specific user interfaces was necessary.

The process of performing field observations at the ARTCCs provided insight into the information needs and information sources of en route controllers. Controller information needs were identified by associating each task step with the information sources consulted to complete the task step (see Figure 1). Example information sources include flight path and track information on the Main Display Monitor (MDM) radar display, flight plan information through the FPS or URET display, pilot communications through VSCS, status and traffic flow information through Enhanced Status Information System (ESIS), and knowledge, rules, and procedures memorized by controllers. Associating information sources with controller positions and individual task steps facilitated the design process by highlighting information elements that may be collocated or integrated in the future.

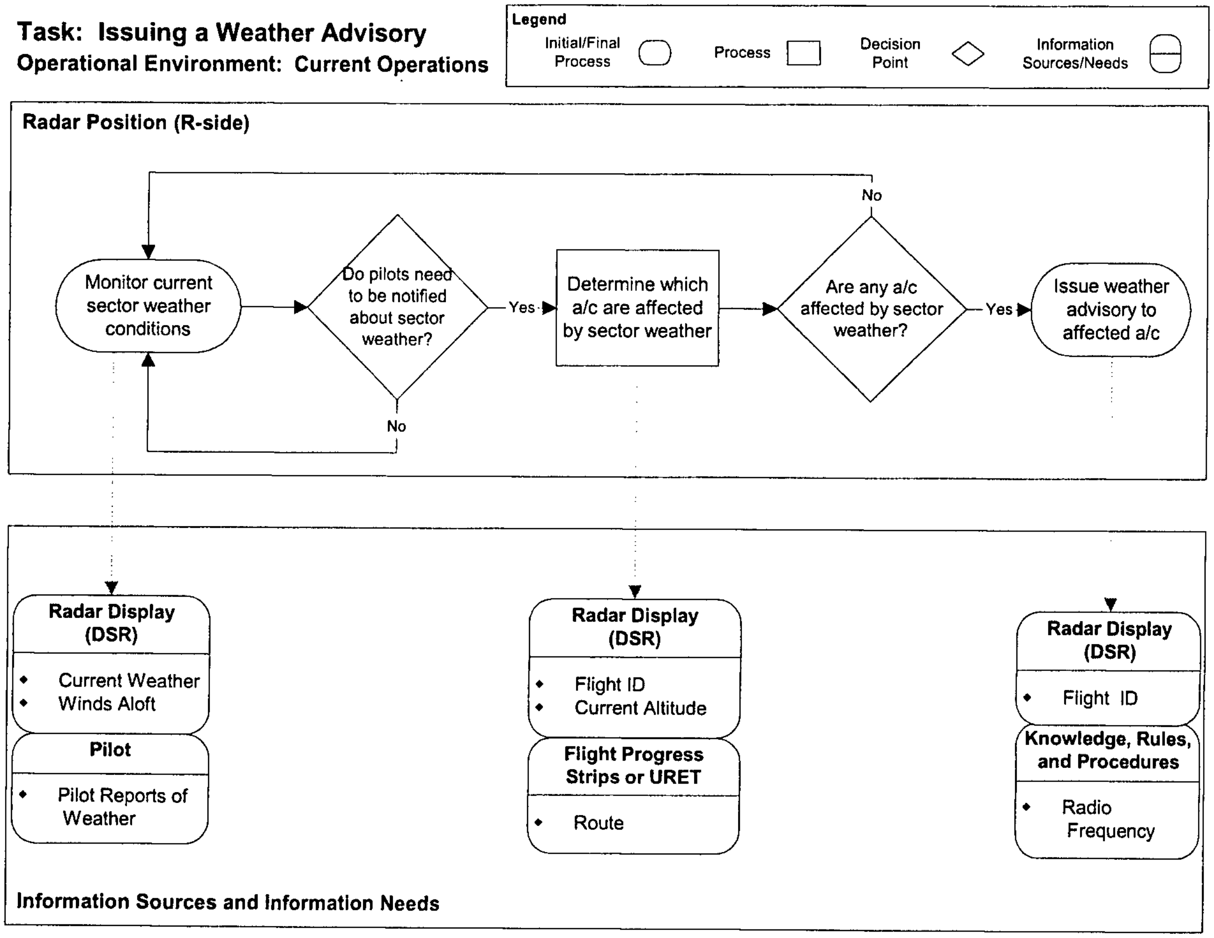


Figure 1. Notional task flow diagram depicting the en route sector controller task of issuing a weather advisory.

A feasibility assessment was conducted to determine if controllers would be able to access the information necessary to complete each task step in the future en route environment. Multiple information sources were consulted to identify the systems proposed for the future NAS environment, including the Operational Evolution Plan Version 5.0 (OEP), the NAS Architecture Version 4.0, and the Blueprint for NAS Modernization. The information sources and interfaces in the current environment were compared against those proposed for the future environment, and feasibility was assessed (as best as possible) against several factors, including the distance between the current and future information sources, the similarity in information format, the compatibility between information sources and cognitive task requirements, and the necessity for information to be integrated on a common display.

The results of the feasibility assessment were used to drive the identification of display elements and candidates for prototyping. Prototyping activities included developing displays and input devices to support controller task performance, emphasizing display consolidation and interface element integration when feasible. These prototypes were presented to controllers and other operational personnel for feedback regarding their usability, suitability, and acceptability, and were redesigned in response to the feedback.

CONCLUSION

Failure to consider user needs early in the design process can render tools and displays (such as those used in the complex and dynamic air traffic control environment) unusable for multiple reasons. Some users may not be able to effectively transition to new systems, while others may require that the system be better suited to their needs (e.g. Standard Terminal Automation Replacement System (STARS), (Observations on the Federal Aviation Administration's Standard Terminal Automation Replacement System, Office of Inspector General, 1997). The task assessment and design process presented in this paper emphasizes the importance of considering user information needs and task goals early in the process. Developing an understanding of controller tasks, creating task flow diagrams, determining information needs, conducting a feasibility analysis, creating prototype designs, and gaining stakeholder feedback all provide the designer with a better understanding of how designs can be developed to support the goals and needs of the users of the systems. Because the FAA is planning to introduce many new systems to the en route environment over the next several years (National Airspace System Architecture Version 4.0, FAA, 1999; National Airspace System Concept of Operations and Vision for the Future of Aviation, RTCA, 2002), the adoption of the type of approach presented here is vital to the successful development of integrated air traffic control systems. By taking the tasks and information needs of the user population into account early in the design and development process, researchers and developers can design systems that will have a better chance of satisfying the needs of the users while also meeting the goals of the organization and its stakeholders.

REFERENCES

Federal Aviation Administration. (2003). FAA Aerospace Forecast Fiscal Years 2003-2014.

Federal Aviation Administration. (1997). ATS Concept of Operations for the National Airspace System in 2005.

Federal Aviation Administration. (2002). Blueprint for NAS Modernization 2002 Update.

Federal Aviation Administration. (1999). National Airspace System Architecture Version 4.0.

Federal Aviation Administration. (2002). National Airspace System Operational Evolution Plan Version 5.0.

Office of Inspector General. (1997). Observations on the Federal Aviation Administration's Standard Terminal Automation Replacement System (STARS) Report Number AV-1998-012.

Redding, R.E., Cannon, J.R., Lierman, B.L., Ryder, J.M., Seamster, T.L., Purcell, J.A. (1992). Cognitive Task Analysis of En Route Air Traffic Control: Model Extension and Validation. McLean, VA, Human Technology Inc.

Redding, R.E., Cannon, J.R., Lierman, B.L., Ryder, J.M., Seamster, T.L., Purcell, J.A. (1992). Cognitive Task Analysis of Prioritization in Air Traffic Control. McLean, VA, Human Technology Inc.

RTCA, Incorporated. (2002). National Airspace System Concept of Operations and Vision for the Future of Aviation (RTCA CONOPS).

AVIATION SAFETY BENEFITS OF NASA SYNTHETIC VISION: LOW VISIBILITY LOSS-OF-CONTROL, RUNWAY INCURSION DETECTION, AND CFIT EXPERIMENTS

Lawrence J. Prinzel III, Monica F. Hughes, Lynda J. Kramer, Jarvis J. Arthur
NASA Langley Research Center, Hampton, VA, USA

ABSTRACT

A national aviation safety goal was established to reduce the accident rate by 80% by 2007. Reducing low visibility as a causal factor in general aviation and commercial accidents may help meet that goal. The paper describes research conducted at the NASA Langley Research Center on the efficacy of synthetic vision to mitigate spatial disorientation, runway incursions, and controlled-flight-into-terrain.

Keywords: Synthetic Vision; Spatial Disorientation; CFIT; Runway Incursion; Inadvertent IMC

INTRODUCTION

Flying is safe. The worldwide commercial aviation major accident rate is low and has remained nearly constant over the past two decades. However, the demand for air travel is expected to increase over the coming two decades, more than doubling by 2017. Without an improvement in the accident rate, such an increase in traffic volume would lead to a projected 50 or more major accidents a year worldwide - a nearly weekly occurrence. Given the very tragic, and damaging effects of a single major accident, this situation would deliver an unacceptable blow to the aviation system. As a consequence, the anticipated growth of the commercial air-travel market may not reach its full potential.

Aviation Safety Program

To ensure the public trust, a national goal was established to reduce the aviation fatal accident rate by 80% by 2007. NASA stepped up to this challenge by forming the Aviation Safety Program (AvSP), which is part of the NASA Aerospace Technology Enterprise (NASA, 2001). The AvSP program has a number of research projects developing technologies to help meet the national safety goal. Among aviation safety enhancement strategies, NASA is working toward the reduction of low-visibility as a causal factor of aircraft accidents.

Synthetic Vision Systems Project

Limited visibility is the single most critical factor affecting both the safety and capacity of worldwide aviation operations. In commercial aviation alone, over 30-percent of all fatal accidents worldwide are categorized as Controlled Flight Into Terrain (CFIT), where a mechanically sound and normal functioning airplane is inadvertently flown into the ground, water, or an obstacle, principally due to the lack of outside visual reference and situational awareness (Wiener, 1977). Other types of accidents involving restricted visibility combined with compromised situational awareness include spatial disorientation and runway incursions.

The AvSP Synthetic Vision Systems (SVS) project is developing technologies with practical applications that will eliminate low visibility conditions as a causal factor to civil aircraft accidents, as well as replicate the operational benefits of flight operations in unlimited ceiling and visibility conditions, regardless of the outside weather or lighting condition. The technologies will emphasize the cost-effective use of synthetic/enhanced-vision displays; worldwide navigation, terrain, obstruction, and airport databases; and Global Positioning System (GPS)-derived navigation to eliminate "visibility-induced" (lack of visibility) errors for all aircraft categories. A major thrust of the SVS project is to develop and demonstrate affordable, certifiable display configurations that provide intuitive out-the-window terrain & obstacle information, including advanced pathway and guidance information for precision navigation, obstacle/obstruction avoidance, and runway incursion detection. SVS display concepts employ computer-generated terrain imagery, on-board databases, and precise position and navigational accuracy to create a three dimensional perspective presentation of the outside world, with necessary and sufficient information and realism, to enable operations equivalent to those of a bright, clear, sunny day regardless of the outside weather condition. The safety outcome of SVS is a display that should help reduce or even prevent CFIT, which is the single greatest contributing factor to fatal worldwide airline and general aviation accidents (Boeing, 1998). Other safety benefits include reduced runway incursions and loss-of-control accidents (Prinzel et al., 2000; 2001; 2002; 2003; Prinzel et al., in press; Williams et al., 2001).

Prevention of Spatial Disorientation

General aviation (GA) accounts for 85 percent of the total number of civil aircraft in the United States. Of the 1,820 accidents in 2002, 1,714 were general aviation with 342 fatal accidents. Although the number of accidents has decreased slightly, the accident rate remains unacceptable at 6.56 accidents per 100,000 flight hours. The majority of fatal GA accidents (67.8%) were the result of pilot-related causes. The overwhelming majority of these accidents took place during instrument meteorological conditions (IMC), which produced almost three times the rate of fatal accidents than flights under visual meteorological conditions (VMC; AOPA, 2001). To help reduce the GA accident rate, NASA is developing GA synthetic vision technologies that could help to mitigate or even prevent spatial disorientation accidents through an intuitive display for VMC-type flight in IMC.

Several experiments have been conducted to evaluate the efficacy of synthetic vision for enhancing aviation safety for GA aircraft. One of these studies focused on whether SVS could help reduce or prevent low visibility, loss-of-control accidents for low-hour visual flight rules (VFR) pilots. The objective of the experiment was to establish the benefits of a synthetic vision for inadvertent IMC (iIMC) situations wherein the VFR pilot accidentally enters clouds and loses the visual horizon. A significant number of accidents happen each year because pilots lose spatial awareness and experience loss-of-control during these iIMC events. VFR flight into IMC is a major hazard in general aviation (O'Hare & Owen, 2000), and 75-80% of accidents classified as inadvertent IMC were fatal compared to 18% of all other GA accident categories (Goh & Weigmann, 2001). Clearly, prevention of spatial disorientation accidents would significantly improve the safety of Part 91 operations. Because many of these accidents are due to a loss of visual cues, researchers at the NASA Langley Research Center (LaRC) evaluated whether synthetic vision displays could mitigate these types of accidents.

Low Visibility, Loss-Of-Control Experiment

The experiment evaluated three displays while 18 low-hour (< 400 hours) pilots executed four maneuvers during iIMC scenarios. The three displays were (a) baseline Cessna-172 instruments, (b) Electronic Attitude Indicator (EAI), and (c) SVS display (Figure 2). The baseline display represented what is currently available on GA aircraft. The EAI display was designed to be more representative of "glass cockpits" and included advanced flight symbology, such as a velocity vector. The third concept was the SVS display that was similar to the EAI display except the blue-sky/brown-ground background was replaced by synthetic terrain. The four scenarios were: straight-and-level flight while maintaining airspeed, altitude, and heading in IMC; 180° turn with a 20° bank upon entering IMC while maintaining altitude and airspeed; descend 1,000 ft. upon entering IMC while maintaining heading and airspeed; and climb 1,000 ft. upon entering IMC while maintaining heading and airspeed.

Several pilots failed to maintain pilot technical standards (PTS) with either the baseline or EAI displays. One pilot experienced a significant loss of situation awareness using the baseline display and became totally disoriented during the 180° maneuver. In comparison, pilot performance was found to be significantly better with the SVS display during each of the four maneuvers (Glabb & Takallu, 2002; Takallu, Wong, & Uenking, 2002). Future research will validate these results in a motion-based GA simulator to simulate the physiological mismatches experienced during spatial disorientation.

Controlled Flight Into Terrain

Aviation has been witness to rapid advancement in technologies that have significantly improved aviation safety. The development of attitude indicators, flight management systems, radio navigation aids, and instrument landing systems (ILS) have extended aircraft operations into weather conditions with reduced forward visibility. However, as Brooks (1997) has noted, "...while standard instrumentation has served us well, enabling aviation as we see it today, literally thousands of dead souls, victims of aviation catastrophe, offer mute and poignant testimony to its imperfections. The simple, elegant dream of soaring aloft *visually, intuitively* – bird-like – remain elusive" (Italics added, p. 17).

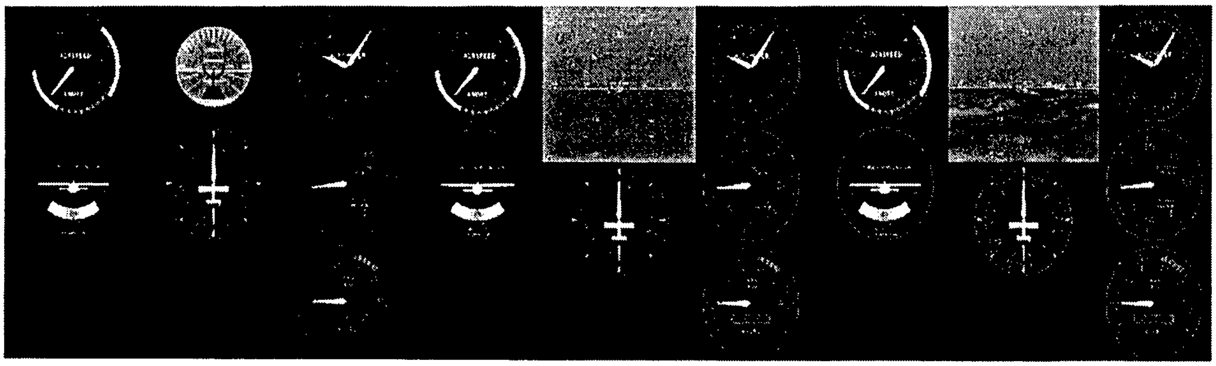


Figure 1. Three NASA Synthetic Vision Displays Used in Low Visibility, Loss-Of-Control Experiment

Pilots must cope within an alphanumeric "filter of symbology" to achieve spatial awareness, which has repeatedly met with deadly consequences. The significant number of CFIT accidents testifies to the danger of losing situation awareness with these "coded" displays (Theunnissen, 1997). Approximately 40% of all aircraft accidents are CFIT and account for 50% of all aircraft fatalities (Mathews, 1997). Because CFITs account for a significant proportion of aircraft fatalities, prevention of these accidents would significantly reduce the accident rate for both commercial and GA aircraft. Often, these accidents are caused because of limited visibility which synthetic vision may help to mitigate.

SVS displays provide a natural presentation of the outside world with information that is intuitive and easy to process. Essentially, it provides a "picture" of the outside world, rather than disparate pieces of alphanumeric information, and best supports humans' natural acquisition and encoding of the world. As the old Chinese proverb goes, "One picture is worth a thousand words". But, in aviation terms, it may be more appropriate to say, "One picture is worth a thousand alphanumerics" (Brooks, 1997) and "...a thousand lives" (Prinzel et al., 2003).

NASA research has successfully evaluated the safety and operational benefits of synthetic vision, but only during nominal, restricted visibility operations (e.g., Glaab & Takallu, 2002; Prinzel et al., 2002; Prinzel et al., in press). Although the research has consistently shown the advantage of synthetic vision compared to traditional instruments for complex approaches to terrain- (EGE, ROA, AVL) or operational-challenged airports (DFW), the true safety value of SVS would be to reduce or eliminate off-nominal situations that present significant safety risks, such as prevention of CFIT. Therefore, two experiments were conducted to evaluate the efficacy of synthetic vision for CFIT prevention.

General Aviation CFIT Experiment

The first experiment focused on general aviation and introduced an inadvertent IMC scenario with an altimeter error. The inadvertent IMC anomaly scenario was designed to show that an otherwise unavoidable CFIT situation could be prevented with synthetic vision technology. Therefore, a baseline display was not evaluated because even highly experienced pilots were unable to avoid a CFIT during preliminary testing. The displays that were tested were based on three different SVS texturing methods: Constant Color (CC), Elevation-Based Generic (EBG), and Photo-Realistic (PR). CC replicates an industry concept that the FAA has certified under the SafeFlight 21 Capstone-II program. The EBG concept uses shades of green with darker shades representing higher terrain. Finally, the PR concept was derived from 4-meter satellite imagery data. The display concepts were combined with 1, 3, or 30 arc-sec digital elevation models (DEM). A 500 x 500 ft grid fishnet was also evaluated.

Pilots flew 34 experimental runs prior to the CFIT scenario (35 total). The CFIT scenario resembled 11 of the previous 34 trials that began straight-and-level at 6500 ft MSL (4000 ft AGL) with instructions to make a left-bank turn and descend after two minutes to 5000 ft MSL (1000 ft AGL) over rising terrain. The scenario began in VMC with visibility deteriorating to IMC within one-minute elapsed time. The CFIT scenario started at 5000 ft MSL, but the altimeter showed 6500 ft MSL. Therefore, the instruction to reduce altitude by 1500 ft in effect descended the aircraft to -500 ft below the mountain peaks directly in front of the aircraft.

Only 15% (2/14) of the VFR pilots and none (0/13) of the professional pilots experienced a CFIT while using the SVS displays. One of these 14 VFR pilots had significant difficulty flying the aircraft throughout the entire experimental session and analysis showed performance to be well outside practical pilot standards; therefore, the data for this pilot should be considered an outlier. The other pilot, however, did experience a CFIT event and, during the semi-structured interview, reported awareness that something was wrong but felt captured by the incorrect MX-20 reading and failed to crosscheck. Despite this CFIT, the results provide strong evidence that synthetic vision can significantly enhance terrain awareness under low-visibility conditions that otherwise would result in an unavoidable CFIT accident.

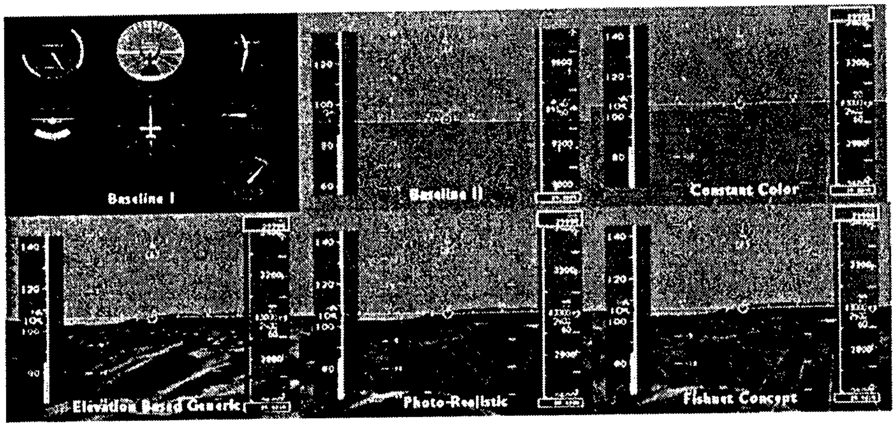


Figure 2. NASA GA Synthetic Vision Displays Used in CFIT Experiment

Commercial Aviation CFIT Experiment

The second CFIT experiment focused on commercial air transport pilots and introduced a lateral path error in flight management system guidance that brought the aircraft into close proximity with terrain during a go-around procedure. Pilots were asked to fly a circling approach to Eagle-Vail, CO (EGE) runway 07 under CAT IIIa and execute a go-around 200 ft AGL and intercept the 059 radial from SNOW VOR (SXW). The aircraft model was a Boeing 757, and both the approach and departure speed target was 140 knots. All scenarios were flown with moderate turbulence. At 200 ft AGL, a go around was executed and the climb gradient performance was degraded. The pilot raised the landing gear and the flaps were set to go-around configuration. The evaluation pilot was instructed to use speed-on-pitch to maintain 140 knots and follow the departure path that provided escape guidance through a "notch" between two mountain peaks. The run ended at the 12.0 DME point from SXW. For the CFIT scenario (run 22 of 22), the flight guidance was altered on the departure path. A Terrain Awareness Warning System (TAWS) and Vertical Situation Display (VSD), however, were available on the navigation display for both baseline and SVS. The display concepts were: (a) baseline EFIS 757 display, (b) size A (5.25" x 5.25") display with SVS, (c) size X display size (8"x10") with SVS, and (d) HUD enhanced with SVS. The order of display presentation was randomized across evaluation pilots. Twelve of the 16 evaluation pilots flew the CFIT scenario with a SVS enhanced PFD or HUD and 4 pilots flew with the Baseline display.

One significant result was that all four Baseline pilots (100%) had a CFIT event, but none (0%) of the twelve SVS pilots did. On average, pilots with a SVS display noticed the potential CFIT 53.6 seconds before impact with the terrain. Three of the 4 pilots impacted the terrain while one passed within 58 feet of a mountain peak (topped trees on mountain). Even though the baseline concept had a Radio Magnetic Indicator (RMI), TAWS and VSD enhanced ND, none of the Baseline pilots were aware until after the CFIT event had occurred. Pilots rated the baseline concept to be "moderately high" on the modified Cooper-Harper workload scale and to be "very low" for situation awareness (SART) during the departure task. SA-SWORD paired comparison rankings confirmed that SVS displays significantly enhanced situation awareness for CFIT detection.

Runway Incursion Detection

Runway incursions are a serious aviation concern. The number of reported incursions rose from 186 in 1993 to 383 in 2001, an increase of 106 percent. In 1990, the National Transportation Safety Board (NTSB) has listed runway incursions as a "top 10" of most wanted transportation safety improvements. The FAA has begun several initiatives to reduce the number of runway incursions, including an alerting system for ATC, which is relayed via voice communication to the cockpit. However, no system is currently available onboard aircraft to provide the flight crew runway incursion alerts. NASA developed a Runway Incursion Prevention System (RIPS) to help provide this information to flight crews.

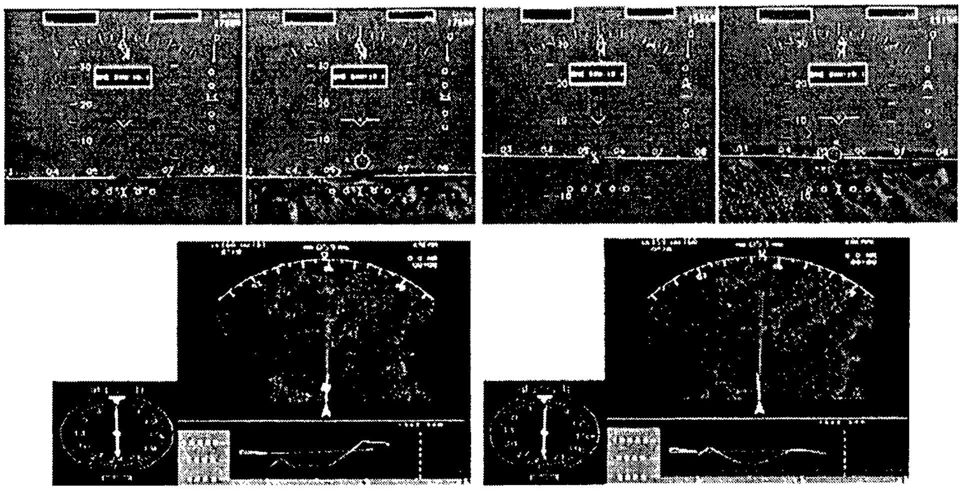


Figure 3. Commercial CFIT Displays During Nominal (Left) and CFIT (Right) Scenarios

Attention Capture Experiment

Head-up displays (HUDs) provide primary flight, navigation, and guidance information to the pilot in a forward field-of-view on a head-up transparent screen. Because HUDs reduce time head down, they enhance pilot performance and situation awareness through simultaneous scanning of both instrument data and the out-the-window scene (e.g., Wickens & Long, 1995). However, research has also documented the phenomenon of "attention capture" and problems detecting unexpected events, such as another aircraft on the active runway during landing. Because synthetic vision HUDs may present compelling near-domain information, there are concerns about whether the pilot can transition to the far domain when the synthetic terrain is removed.

Research was conducted using a rare-event scenario in which a B-737 taxied beyond the hold line and presented a runway incursion situation. The experiment was part of research to evaluate pathways displays presented on a SVS HUD while pilots flew complex, curved approaches in simulated CAT IIIa conditions. Nine 757 Captains with HUD experience participated in the experiment. Fourteen approaches using the Reno Sparks 16R Visual Arrival were made in a B-757 fixed-based simulator. In addition, a runway incursion scenario was flown in which the pilot was forced to make a go-around to avoid a 737 on the active runway. Pilots were not given the option to "de-clutter" the synthetic terrain and instead it was automatically removed just before decision height making the scenario a "worse case" for runway incursion detection.

Only one (1/9) of the commercial pilots failed to notice the transport aircraft on the active runway. During the post-experimental interview, he acknowledged that he saw the aircraft but it was too late to initiate the go-around and decided to land. The pilot felt that the situation did not pose any danger since he could land the aircraft further down the runway well beyond the incursion aircraft. Therefore, these results support that a synthetic vision HUD does not significantly decrease unexpected event detection. However, to further safeguard against incursions, the AvSP has incorporated RIPS technology to be used as part of the NASA synthetic vision system.

Runway Incursion Prevention System

RIPS integrates airborne and ground-based technologies to provide: (1) enhanced surface situation awareness to avoid blunders and (2) runway conflict alerts in order to prevent runway incidents and improve operational capability. The system monitors for potential incursions using incursion detection algorithms that provide both aural and graphical alerts. The alerts can be presented on a HUD, PFD, or electronic moving map (EMM). RIPS also enhances situation awareness by providing graphical guidance during rollout, turn-off, and taxi. The EMM displays a graphical perspective airport layout, current ownship position, traffic, and ATC instructions. Together, RIPS has been demonstrated to significantly increase situation awareness and eliminate the occurrence of runway incursions during both simulation (e.g., Jones, 2002) and flight tests (e.g., Jones, Quach, & Young, 2001).

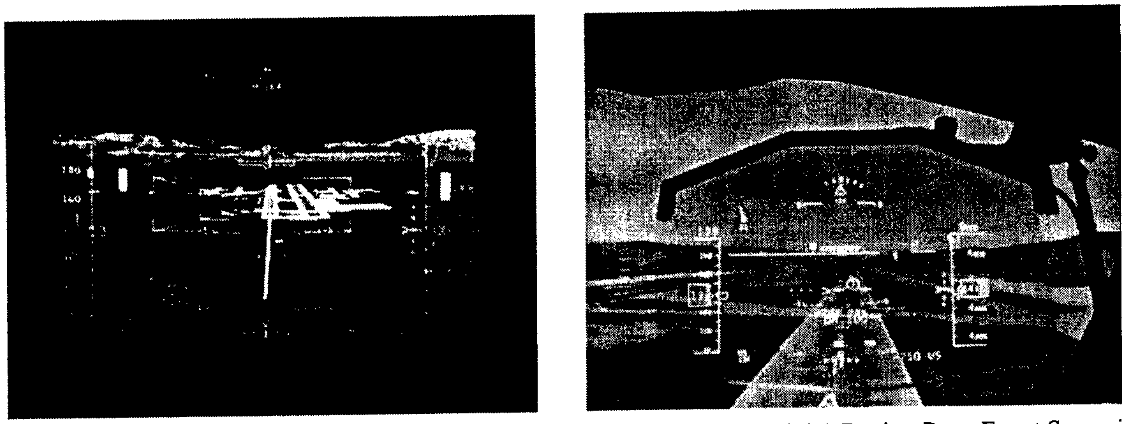


Figure 4. Head-Up Displays On Approach (Left) and At Decision Height (Right) During Rare-Event Scenario

Figure 5. Runway Conflict Alert Presented On HUD (Left) and EMM (Right)

CONCLUSIONS

The paper describes the aviation safety benefits of the NASA Synthetic Vision System, and presents a sample of research that has been conducted to demonstrate the efficacy of SVS to meeting national aviation safety goals. Synthetic vision is composed of several technologies that include SVS navigation displays; RIPS; integrity monitoring; enhanced vision sensors; taxi and surface maps; and advanced communication, navigation, and surveillance. Together, these technologies represent a comprehensive solution to problems of restricted visibility. Future research is planned for GulfStream-V and 757 flight tests that will evaluate these technologies as part of an integrated system. Research is also ongoing for simulation research, including synthetic navigation displays, 4D tunnels, helmet-mounted displays, and synthetic/enhanced sensor blending.

REFERENCES

Glabb, L.J., & Takalu, M.A. (2002). Preliminary effect of synthetic vision systems displays to reduce low-visibility loss of control and controlled flight into terrain accidents. SAE Technical Paper 2002-01-1550. Warrendale, PA: Society of Automotive Engineers.

Jones, D.R., Quach, S.D., & Young, S.D. (October 14, 2001). Runway Incursion Prevention System – Demonstration and testing at Dallas/Fort Worth International Airport. 20th Annual Digital Avionics Systems Conference (DASC). Daytona Beach, FL.

Prinzel, L.J., Hughes, M.F., Arthur, J.J., Kramer, L.J., Glaab, L.J., Bailey, R.E., Parrish, R.V., & Uenking, M.D. (2003). Synthetic Vision CFIT Experiments for GA and Commercial Aircraft: "A Picture Is Worth A Thousand Lives". Proceedings of the Human Factors & Ergonomics Society, 47, 164-168.

Prinzel, L.J., Kramer, L.J., Comstock, J.R., Bailey, R.E., Hughes, M.F., & Parrish, R.V. (2002). NASA synthetic vision EGE flight test. Proceedings of the Annual Human Factors and Ergonomics Meeting, 46, 135-139.

Prinzel, L.J., Comstock, J.R., Glaab, L.J., Kramer, L.J., Arthur, J.J., & Barry, J.S. (in press). The efficacy of head-down and head-up synthetic vision display concepts for retro- and forward-fit of commercial aircraft. International Journal of Aviation Psychology.

A COMPARISON OF AUDITORY AND VISUAL IN-VEHICLE INFORMATION DISPLAYS

Deborah Bruce
National Transportation Safety Board

Deborah A. Boehm-Davis and Karen Mahach
George Mason University

ABSTRACT

Two experiments examined the effectiveness of presenting in-vehicle auditory information to drivers as an alternative to visual information display. In the first experiment, 24 participants (12 younger and 12 older) were presented with roadway symbolic or text-based signs using either graphic images or natural voice audio recordings and asked to identify whether the displayed message matched a projected sign image. Performance accuracy and recognition times for the projected signs were equally fast and accurate using visual and auditory displays. A second experiment with 24 additional participants looked at the addition of music or talk radio background noise to the test environment. Noise did not differentially affect participant's reaction time or accuracy of performance to information. However, there was some indication that background noise increases overall reaction time. The results of both experiments indicate that auditory information displays can effectively inform drivers and should be considered where feasible.

Keywords: auditory displays, visual displays, intelligent transportation systems.

INTRODUCTION

Highway drivers are increasingly being provided with in-vehicle computer displays and roadside communication devices to acquire information concerning routes and traffic status. The design of these information systems has focused almost exclusively on providing the driver with information through visual displays. This proliferation of in-vehicle computer displays is proceeding despite the fact that reference to driver's overload from visual information has existed in the research literature for several decades (Sivak, 1996; Dewar, 1988; Zwahlen & DeBald, 1986; Treat, et al., 1977; Senders, Kristofferson, Levinson, Dietrich, & Ward, 1967). "There is now much evidence that drivers are quite often operating beyond their visual or perceptual capabilities in a number of key driving situations, including overtaking, joining or crossing high speed roads, and in a number of nighttime situations," Hills (1980). Given the heavy demand that driving places on the visual channel, it is prudent to consider alternative display modalities (Mollenhauer, Lee, Hulse, & Dingus, 1994).

Parks and Burnett (1993) concluded that drivers could spend more time looking at traffic movement and lane position when additional information is presented using auditory rather than additional visual displays. Unlike the perceptual channel for visual processing, auditory perception is not overloaded during the driving task. It is, in fact, rarely used. In a task analysis of driving behavior, McKnight and Adams (1970) found that only one percent of critical driving tasks, such as driver's identification of emergency vehicle sirens or awareness of unusual engine sounds, were hearing related.

As more visual displays are added to automobiles, not only does the magnitude of visual information processing increase, but the requirement to shift attention between different visual displays also increases. In addition to the serial requirements of visual processing time, there is added mental work required to switch attention between different visual displays. Baldwin and Schieber (1995) concluded that visual attention switching has a large decremental effect on driving performance, especially for older drivers.

A primary concern for visual displays is whether drivers can find and use information while actively driving the vehicle. The workload associated with attending to in-vehicle displays depends on the complexity of the message, interaction requirements necessary to manipulate the system, and time pressures associated with processing the information. For example, driving activities that are associated with travel planning have an elastic window of time that may or may not affect driving performance.

A different situation exists for in-vehicle systems designed to augment real-time operation and control of the vehicle. These systems, which provide the driver with information concerning traffic signs, direction of next turn, and collision avoidance, are more time critical because they focus on vehicle control activities that have a finite time frame of performance (Schofer, et al., 1997). Time requirements for processing traffic information while executing vehicle control often require drivers to perform multiple, concurrent activities, thereby increasing attentional demand.

We have long known that simultaneous tasks are more difficult to perform if they share the same sensory modality (Norman and Bobrow, 1975; Wickens, 1983). Wickens described the cognitive resource advantages of using two different perceptual channels to perform simultaneous tasks; a situation referred to as bimodal time-sharing. Using both auditory and visual displays rather than additional visual displays would allow for dual task processing through different perceptual channels and result in less task interference.

A limited amount of research on auditory displays has been applied to the driving domain. Though limited in scope, the results of research on auditory displays have generally been positive. Auditory route guidance information has been associated with more efficient driving, as measured by time and distance (Streeter, et al., 1985); auditory route guidance devices result in fewer navigational errors (Walker, Alicandri, Sedney, & Roberts, 1990); and drivers react faster and with fewer errors using auditory information systems instead of visual systems (Srinavasan & Jovanis, 1997). Yet, application of these results to design has been minimal; the ITS Human Factors Design Guide cites only Labiale (1990) and Mollenhauer, et al. (1994) to substantiate design guidelines for auditory display.

The technical aspects of in-vehicle auditory displays are not at issue, but the human factor aspects of the driver's interface with these information systems will be critical to successful implementation of ITS. A common limitation of previous research was that auditory and visual displays were not compared using the types of messages that would be suitable for auditory displays. The majority of ITS research has been associated with navigation systems and driver's experience with road maps does not easily translate to auditory information. Moreover, the environmental effects of noise and interference with auditory displays need to be further explored.

METHOD

Two experiments were conducted to address the previous shortcomings by comparing driver performance using auditory and visual display of short, well-known road sign messages. Further, the study manipulated the information content of the display (symbolic or text messages) to examine whether additional cognitive effort is required to associate graphic images with name-identified messages. In the second experiment, noise common to the driving environment was incorporated into the research to examine the effects of interference.

Experiment 1

Experiment 1 employed a mixed-factorial, repeated measure ANCOVA design. There were two between-subjects factors: age (categorized as 35 and younger or 65 and older) and sex. There were three within-subjects factors: display channel, message type, and match type. Match type was categorized based on three types of sign pairs: first presentation of matching sign pairs, a second presentation of matching sign pairs, or presentation of non-matching sign pairs. There were six trials for each of the 12 experimental conditions, totaling 72 trials. For half of those presentations, the projected sign image was preceded by an auditory pre-cue display; for the other half, the projected sign image was preceded by a visual pre-cue display. During the original preparation of the computer scenario, match/no match comparisons between the pre-cue and target signs were randomly assigned to one-third of the trials. The order of slide presentation, and therefore the sequence of graphic and text message formats, was also randomized. The experiment was conducted in the SIGNSIM laboratory of the Federal Highway Administration. A type 1 precision sound level meter was used to control sound volume. Static visual acuity and hearing ability were used as covariates in the data analyses. Participants were also presented with written questions following completion of the reaction time trials. The questionnaire asked about their preferences and concerns for receiving in-vehicle information.

Experiment 2

In the second experiment, noise distractions were added to the test environment. This experiment asks if auditory and visual displays lead to equal performance under conditions that include background noise. Furthermore, it asks if the type of background noise, music or voice, differentially affects driver performance using information presented by auditory or visual information displays.

The methods and procedures for Experiment 2 were the same as for Experiment 1, except that the environmental test conditions in Experiment 2 were altered to include a noise distraction that consisted of one of two different recordings - recorded music without lyrics or recorded talk radio. Calibrated recordings that presented sounds of similar volume and tone were developed for the purpose of the experiment. The recordings were played continuously in the SIGNSIM laboratory environment while participants responded to the experimental trials.

RESULTS

The results of experiment 1 showed no significant difference for auditory versus visual displays, $\underline{F}$ (1, 18) = .13, $\underline{p}$ = .72. The mean time for responses to auditory displays was 9.60 sec (standard deviation = 2.08 sec) and the mean time for responses to visual displays was 9.35 sec (standard deviation = 2.09 sec). There was a significant difference in reaction time based on message type, $\underline{F}$ (1, 18) = 14.44, $\underline{p}$ = .001. Recognition of text messages (11.13 sec) took appreciably longer than recognition of symbol messages (7.82 sec). Accuracy did not differ as a function of auditory or visual display.

Recognition time scores for male participants ($\underline{M}$ = 9.59 sec) were not different from those for females ($\underline{M}$ = 9.36 sec), $\underline{F}$ (1, 18) = .10, $\underline{p}$ = .75. But there was a significant interaction between sex and message type, $\underline{F}$ (1,18) =10.14, $\underline{p}$ = .005. Plots of the estimated marginal means indicated that females were less affected by message type differences than were males. It took females slightly longer on average to recognize symbol signs (8.02 sec for females compared to 7.62 sec for males) but they recognized text signs faster (10.69 sec for females compared to 11.56 sec for males).

The results of Experiment 2 confirm the effectiveness of auditory display of in-vehicle sign information shown in Experiment 1. The addition of noise to the test environment did not affect auditory displays differently than visual displays. However, in terms of an exploratory analysis, the combined data for the two experiments indicated that there was a significant effect of background distracters on in-vehicle information displays. The overall reaction times for Experiment 2 were approximately one second longer than for Experiment 1 (9.47 sec compared to 10.43 sec). An analysis of reaction time results from all 48 participants showed that the addition of a distracter noise to the test environment did have a significant effect, $\underline{F}$ (2,43) = 4.31, $\underline{p}$ = .02. This was evident despite the fact that questionnaire responses overwhelmingly indicated that participants did not think that the addition of noise presented a performance problem. The addition of noise in the second experiment did not negatively affect response accuracy.

Drivers were fairly divided on whether they thought they could view visual displays without being distracted from driving tasks, with only slightly more than half (56 %) saying that they could view a computer screen in their dashboard without affecting their driving. However, there was a definite difference if this question was considered by age group. Two thirds of older participants indicated that they did not think they could glance at a computer screen in their dashboard without affecting their driving; while more than three quarters of younger participants thought they could view a computer screen display without it affecting driving.

DISCUSSION

The value of these two experiments is that they confirm the feasibility of auditory displays for use in the driving environment. Driving places a heavy demand on the need for visual information, so it is prudent to consider whether alternative display modalities are suitable for in-vehicle information systems. Yet human factors research on Intelligent Transportation Systems has focused primarily on the use of visual displays to transmit information to drivers. The Human Factors Guidelines for Advanced Traffic Information Systems (ATIS) acknowledges that very little research has been performed to evaluate the different methods of displaying sign information with an in-vehicle system (FHWA, 1998).

Although older drivers appeared to exhibit slower reaction times than younger drivers in both experiments, the effect was only reliable in Experiment 2. Perhaps more important, in both experiments, the performance of older drivers did not vary based on whether participants received information through an auditory or a visual display. This indicates that auditory displays do not impose a differentially negative effect on older drivers.

The second experiment specifically addressed the issue of noise distracters common to the driving environment. Users of visual information systems can filter out information by redirecting their line of sight; however, it is not as easy to selectively attend to one audio message while excluding others. While the overall response time under conditions of background noise were longer, the addition of noise did not affect driver's performance using the auditory display any differently than it did when using the visual display.

Since the vast majority of driving research is conducted in a controlled simulator laboratory setting, the experimental results concerning noise distraction would suggest that high-fidelity driving simulators should include a wide range of perceptual distracters. It would then be possible to determine the extent to which in-vehicle displays are affected by auditory and visual distractions that occur on a regular basis in the driving environment. More research is needed to determine the interactive effects that environmental conditions pose for in-vehicle information displays.

The essence of ITS is to provide useful information to drivers, consequently one of the primary issues with these new information systems is not technical feasibility, but rather usability. Norman (1988) has clearly stated that effective interfaces begin with an analysis of what the person is trying to do, rather than as a metaphor for what the screen should display. This distinction between merely providing information or helping with the activity becomes clearer as we examine the past development of computer technology.

The obvious functionality of information devices is disappearing. Translating this trend into ITS system design means drivers shouldn't interact with the information technology device in their car, the technology should invisibly assist them with driving tasks. Evolution of the computer interface is now leading to, as Laurel (1993) calls it, "direct engagement."

Fully integrated, natural language information systems may not be part of the near-term ITS systems deployed in automobiles, but they should be considered during the system design process. These systems can include criterion-based or inquiry-based designs that avoid the nuisance display aspects associated with audio-alerts on cars of the past, as exemplified by "your door is ajar" announcements. Intelligent auditory display systems are technologically feasible but for reasons of cost and infrastructure requirements they will evolve over the coming decade. Unfortunately, a great deal of research development work currently underway to design visual interfaces seems to have overlooked the performance advantages of auditory displays.

REFERENCES

Baldwin, C. L., & Schieber, F. (1995). Dual task assessment of age differences in mental workload with implications for driving. Proceedings of the 39th Annual Meeting of Human Factors and Ergonomics Society, pp. 167-172.

Dewar, R. E. (1988). In-vehicle information and driver overload. International Journal of Vehicle Design, 9, pp. 557-564.

Federal Highway Administration. (1998). Human Factors Design Guidelines for Advanced Traveler Information Systems (ATIS) and Commercial Vehicle Operations (CVO). (FHWA-RD-98-057). Department of Transportation, Washington, DC.

Hills, B. L. (1980). Vision, visibility, and perception in driving. Perception. 9, pp. 183.

Labiale, G. (1990). In-car information: comparison of auditory and visual presentation. Proceedings of the 34th Annual Meeting of Human Factors and Ergonomic Society. pp. 341-346.

Laurel, B. (1993). Computers as theatre. New York: Addison-Wesley Publishing,. pp. xviii.

McKnight, A. J., & Adams, B. B. (1970). Driver education task analysis, Volume 1: Task description report, DOT HS 800 367, Department of Transportation.

Mollenhauer, M., Lee, J., Cho, K., Hulse, M., & Dingus, T. (1994). The effects of sensory modality and information priority on in-vehicle signing and information systems. Proceedings of the 38th Annual Meeting of Human Factors and Ergonomics Society. pp. 1072-1077.

Norman, D. (1988). The psychology of everyday things. New York: Basic Books.

Norman, D., & Bobrow, D. (1975). On data-limited and resource limited processing. Journal of Cognitive Psychology, 7, pp. 44-60.

Parkes, A. M., & Burnett, G. E. (1993). An evaluation of medium range advance information in route guidance displays for use in vehicles. IEEE Proceedings of Vehicle Navigation and Information Systems Conference. Institute of Electrical and Electronic Engineers.

Schofer, J. L., Koppelman, F. S., & Charlton, W. A. (1997). Perspectives on driver preferences for dynamic route guidance systems, paper 97-0146, 76th Annual Meeting, Transportation Research Board, Washington, D.C.

Senders, J. W., Kristofferson, A. B., Levinson, W. H., Dietrich, C. W., & Ward, J. L. (1967). The attentional demand of automobile driving, Highway Research Record, 195, pp. 15-33.

Sivak, M. (1996). The information that drivers use: is it indeed 90% visual?, Perception, 25. pp. 1081-1089.

Srinivansan, R., & Jovanis, P. P. (1997). Effects of selected in-vehicle route guidance systems on driver reaction times, Human Factors, 39(2), pp. 200-215.

Streeter, L. A., Vitello, D., & Wonsiewicz, S. A. (1985). How to tell people where to go: comparing navigation aids. International Journal of Man-Machine Studies. 22, pp. 549-562.

Treat, J. R., Tumbas, N. S., McDonald, D. T., Shinar, D., Hume, R. D., Mayer, R. D., Stansifer, R. L., & Castallen, N. J., (1977). Tri-level study of the cause of traffic accidents. Report No. DOT-HS-034-3-535-77, Department of Transportation.

Walker, J., Alicandri, E., Sedney, C., & Roberts, K. (1990). In-vehicle navigation devices: effects on the safety of driving performance. Tech Report FHWA-RD-90-053. Department of Transportation.

Wickens, C. D. (1983). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), Varieties of attention. Orlando, FL: Academic Press.

Zwahlen, H. T., & DeBald, D. P. (1986). Safety aspects of sophisticated in-vehicle information displays and controls. Proceedings of the 30th Annual Meeting of Human Factors and Ergonomics Society Proceedings pp. 604-609.

NONMOTION FLIGHT SIMULATION DEVELOPMENT OF SLOW FLIGHT AND STALL TASKS FOR USE IN AB INITIO FLIGHT TRAINING

Richard P. Anderson, Nickolas D. Macchiarella
Embry-Riddle Aeronautical University

ABSTRACT

The ever-increasing computational power used to drive ground-based flight simulations and flight training devices (FTDs) is enabling higher levels of fidelity at lower costs while accurately modeling specific aspects of flight. With an appropriate level of fidelity, nonmotion fight simulators can serve as a means for training ab initio pilots for slow flight and stall tasks. Due to the visual nature of these flight tasks, and the absence of full proprioceptive and vestibular cues during nonmotion simulated flight, modifications to data derived from flight testing and used in aircraft modeling can accentuate other sensory modalities to deliver an effective simulated flight training environment.

Keywords: Flight Training Device; ab initio flight training; flight modeling; flight simulation; virtual reality; psychophysical stimulation

INTRODUCTION

Training Need, Certification and Fidelity

Flight Training Devices (FTD) and full motion flight simulators have been used for years for advance flight training in the military and airlines. However, until recently ab initio flight educators did not have a compelling need to adopt these devices. Advances in simulation technologies and decreasing simulation costs have created circumstances that now favor the adoption of advanced simulation devices for ab initio use (Brady, 2003).

Embry-Riddle Aeronautical University (ERAU) is now employing FTDs for the primary flight training of private, instrument, commercial and Certified Flight Instructor (CFI) students (Embry-Riddle Aeronautical University, 2003a). ERAU, traditionally an innovative world leader in aviation and aerospace education, is incorporating advanced flight simulation into pilot flight training from the start (i.e., ab initio). This approach to flight training uses new high fidelity Level 6 FTDs with a visual display system (Federal Aviation Administration, 1992). The uses of these FTDs is at the core of the University's flight curriculum and are pending final certification by the Federal Aviation Administration (FAA) under Federal Aviation Regulation (FAR) Part 142 *Training Centers.*

Although the Cessna 172 FTD is a non-motion device, it delivers a flight experience that is realistic visually, both in cockpit and out of cockpit, and tactilely with regard to flight control manipulation. The Cessna 172 FTD does not provide a flight experience that delivers high levels of kinesthetic, proprioceptive and vestibular sensory inputs for the pilot. However, psychophysically, pilots in nonmotion simulators use visual cues to generate sensations of motion. Industry practice and research indicates that physical and tactile considerations are less important than previously believed (Chung, 2000; Hope, 2003; Szczepanski & Leland, 2000). The Cessna 172 FTD with its wide 220-degree visual dome system high fidelity simulation provides a flight environment that is rich with scenery that is conducive to creating the perception of self-motion (Brandt, Wist, & Dichgans, 1975).

Physical Description and Modeling for the Cessna 172 FTD

A Level 6 FTD as defined by the National Simulation Program (NSP) is a non-motion training aid that is aircraft type specific (Federal Aviation Administration, 1992). The FTD addressed in this paper is based upon a Cessna Skyhawk Model 172S. The difference between a simulator and a FTD, as defined by the FAA, is its motion base. Advances in simulator technologies and a cost benefit analysis affected ERAU's decision to adopt a Level 6 Cessna 172 FTD with a wide field of view (FOV) visual display (Brady, 2003). Adaptation of a nonmotion FTD for ab initio flight training presented several challenges to providing positive training for all of the Practical Test Standard (PTS) maneuvers in a light aircraft (Federal Aviation Administration, 1997).

Frasca International Incorporated created the Cessna 172 FTD. The Cessna 172 FTD uses a real cockpit section of a Cessna 172 . The cockpit section was built at Textron Incorporated, Cessna Aircraft Company, Independence, Kansas. This real cockpit is manufactured on the same Cessna 172 production line that produces real and flyable aircraft. From the Cessna 172 FTD's firewall forward, it houses some computer and all flight control loading equipment. Only the two front seats of the cockpit section are present in the Cessna 172 FTD. The cockpit area ends immediately behind the two pilot seats. An instructor's station is located aft of the two pilot seats and incorporates a computer workstation with a graphical interface to monitor and control the simulation.

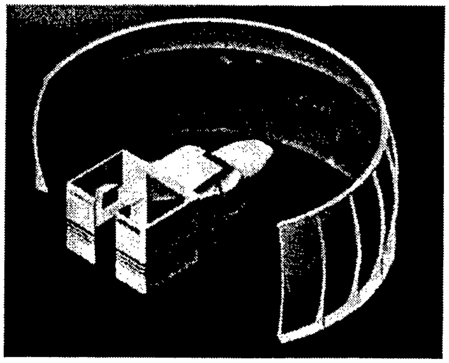


Figure 1. Cessna 172 FTD
(Frasca International Inc., 2003)

Figure 2. Cockpit View Cessna 172 FTD
(Frasca International Inc., 2003)

Selected visual, aural and haptic sensations associated with the real aircraft were incorporated. The air vents blow air on the pilots and the airflow from theses vents change velocity based upon free stream airspeed. The engine, flap movement and stall horn sounds are present. Engine sound varies with RPM and the RPM is dependant upon many factors including airspeed and engine power. The avionics match the ERAU line aircraft physically and functionally. This includes Global Positioning System (GPS), very high frequency omnidirectional range (VOR) and Instrument Landing System (ILS) navigation capabilities. The radios and intercoms function as they do in the real aircraft. In addition, the FTDs have the capability of being networked into a fleet wide simulation. In fleet mode, FTDs can see and hear other FTDs in an interactive simulation environment. Using a two-way transmission of audio over a packet-switched IP network (i.e., Voice Over Internet [VoIP]) methodology, pilots that select the same radio frequency can talk with each other. FTD pilots within visual range can see each other. This simulated flight environment enables training in situational awareness and visual separation.

It was evident from the numerous visual maneuvers -- all flight maneuvers flown early in pilot training are visually based -- that the FTD would need a visual system. Visual systems are not required for even the highest level of FTD. The only NPS prescription for a visual system that is integrated into a FTD is that it does not yield negative training (i.e., a simulated flight experience that does not match real world flight). The Cessna 172 FTD uses a three-projector 220-degree dome visual system. The visual database is based upon satellite imagery of the Daytona Beach area with 10-meter resolution. Local airports are drawn in with a higher degree of detail. The visual system is optimized for flight at altitudes greater than 3,000 feet above ground level (AGL). Below this altitude, a higher resolution to the visual display could yield more realistic simulated scenery. The virtual wings and lift struts obscure the pilot's view of the domed visual system to simulate the real world view out of the cockpit. Even aileron deflections are accurately represented in the visuals and respond in real time to control inputs.

Modeling Capabilities

The modeling of aerodynamics and ground reactions is via a digital computer solving a six-degree of freedom (6-DOF) set of dynamic equations. The aircraft specific data is entered through stability derivative, which are the coefficients of the 6-DOF equations of motion. For many simulations, the simulation occurs in the middle of the flight envelope. The nature of the aerodynamics is such that the coefficients at these low angle-of-attack conditions tend to be linear. Much of the time in training for the PTS, unlike airline training, is at high angles of attack. At high angles of attack, the stability derivatives are non-linear. This makes accurate simulation of high angle of attack flight more difficult that low angle of attack cruise flight. In addition, high angle of attack flight in one G conditions tends to be low speed. At low speeds, the ratio of aerodynamic forces to other forces change. The Cessna 172 is a reciprocating, single-engine propeller aircraft. At low speeds, the effects from the motor and propeller become large with respect to the diminishing aerodynamic forces. Not only do the non-linear aerodynamic coefficients, therefore, have to be modeled, but accurate p-factor, gyroscopic effects, destabilized propeller effects and torque must be modeled.

The first step in determining modeling requirements for the Cessna 172 FTD was to list the required PTS maneuvers and collect flight test data on each maneuver. In addition, flight test procedures were developed to draw out difficult to determine stability derivatives. There are 12 PTS required maneuvers that were modeled for the Cessna 172 FTD (see Table 1).

Table 1: Maneuvers Flight Tested and Modeled for the Cessna 172 FTD

Required Maneuvers	
Lazy-8	Chandelle
Slow Flight	Power Off Stalls
Power On Stalls	Left and Right Turn Spins
Elevator Trim Stalls	Secondary Stalls
Power Off Glides	Steep Turns
ILS Approach	Normal Traffic Pattern

Each of these maneuvers was added to the list of supplemental maneuvers to be flown during the Level 6 flight test program. While the Cessna 172 was in flight test, new models were developed to handle the high angle of attack envelope expansion. The new models necessary to achieve the desired fidelity were: longitudinal and lateral-directional propeller destabilizing effects, longitudinal and lateral-directional gyroscopic effects, p-factor, stall model and an asymmetric wing lift (spin) model.

After incorporating these new models, all of the maneuvers in Table 1 matched the flight test data with one exception, real elevator trim stall characteristics did not match the model and subsequently the simulation. The Cessna 172 when trimmed for landing with flaps down has a dramatic pitch up with the application of thrust. The increase in drag on the flaps is due to propeller slipstream of the flap positioned above the vertical center of gravity (CG). This effect creates a nose up moment. Accounting for this nose up moment required making the aircraft pitching moment a function of flaps and thrust coefficient. Although, the change due to thrust was negligent in the clean configuration, it was significant with flaps down (Embry-Riddle Aeronautical University, 2003b).

After including these new compensated models, the FTD matched flight test data for the maneuvers flight tested (see Table 1). Subjective and qualitative testing of the Cessna 172 FTD by using experienced Cessna 172 pilots showed positive feedback on the modeling and handling qualities. Large improvements were noted in the areas of envelope expansion. The devices were qualified by the NSP and put into service.

Psychophysical Aspects of Flight Training with the Cessna 172 FTD

Pilots in nonmotion simulators (e.g., Cessna 172 FTD) use visual cues as the primary means for generating sensations of motion. Visually induced self-motion and spatial orientation are primarily generated by viewing images in motion located in the periphery of the visual field and in the back ground of the visual scene (Brandt et al., 1975). Auditory cues play a secondary role. In real flight, and to a lesser degree during flight simulation in a motion-based device, the somatosensory system delivers multiple types of sensations from the body (e.g., light touch, pain, pressure, temperature, and joint and muscle position sensations--also called proprioception) that affects the pilot's sense of self-motion. The simulator pilot's spatial orientation and situational awareness are directly linked to how the brain processes these sensory inputs (Szczepanski & Leland, 2000). Longridge, Bürki-Cohen, Go, and Kendra (2001) call for further investigation on the affects of platform motion on transfer of training from FTDs to real flight. Several studies suggest that wide FOV visual systems produce training results equivalent to motion based simulators: the absence of motion does not negatively affect transfer of training from a FTD to real flight (Bürki-Cohen et al., 2000; Longridge, Bürki-Cohen, Go, & Kendra, 2001; Waag, 1981). The use of the Cessna 172 FTD at ERAU highlights issues regarding visually induced self-motion and training to perform flight tasks from the PTS to standard in a FTD with a wide FOV visual system.

The Cessna 172 FTD accurately matches the flight test data obtained during slow flight and stalls with real world Cessna 172s in use at ERAU for student pilot training. In an effort to create a better flight training experience and in the absence of somatosensory and vestibular inputs, the simulation developers modified the original model with regard to slow flight characteristics. As students began flying the Cessna 172 FTD, flight instructors noticed a trend; students demonstrated difficulty performing slow flight in accordance with the PTS . The PTS requires that pilots fly to an angle of attack that if increased at all would result in an immediate stall (Federal Aviation Administration, 1997). While flying a real airplane, the pilot perceives the onset of a stall primarily through proprioception and vestibular sensations. Students demonstrated difficulty determining the stall angle of attack in the Cessna 172 FTD even though all the visual cues and control feels matched the flight test data obtained from the real airplane (Embry-Riddle Aeronautical University, 2003b).

In the Cessna 172 FTD, student pilots repeatedly initiated slow flight well above the stall angle of attack. Students would then slow the aircraft increasing the angle of attack. If the student was unable to determine the correct angle of attack for slow flight, the student would continue to increase the angle of attack well beyond a stalled condition. Both the model and simulation associated with the Cessna 172 FTD delivers all visual and auditory signs of a stall. However, it cannot simulate a true stall; in the real airplane, the "bottom drops out" and the stall becomes physical and obvious. At stall in the Cessna 172 FTD, the nose drops and then pitches up when uncompensated for entering a low power falling leaf. If the student continues to hold back elevator pressure thinking that this condition is still slow flight, the sink rate increases rapidly and the post stall falling leaf condition continues. To correct this problem the student increases the power but not even full power can overcome the drag in this configuration. The aircraft is now in a full power falling leaf. All of the aforementioned models are now strongly governing the dynamics of the aircraft. Without swift corrective action, this full power falling leaf quickly degrades into a full power spin, as does the real aircraft. The comment from a confused student was, "the airplane spins too easy." In fact, this is not true; the Cessna 172 FTD is based upon an accurate model of a real Cessna 172 (Embry-Riddle Aeronautical University, 2003b). The proper pitch attitude is present and the stall dynamics are accurate. Additionally, the stick forces are accurate at stall.

The fact remains, however, that the students have difficulty learning how to perform slow flight. There are two theories that have been developed to explain this problem. First, the flight test data was recorded as a relatively slow rate, 2 Hz. There may be vibration in the stick and movement of the visuals that are higher than 2 Hz that has not been captured in the flight test data as subsequently programmed into the equation of motion. Thus, from a kinematics perspective, the simulation may not exactly replicate high frequency perceptions that may be present in the real airplane.

The second, and likely more tangible theory, is that the feedback-learning loop is not as clear as in the airplane. If slow flight is performed correctly, there should be no need for a motion base as the aircraft is in nearly unaccelerated level flight. If performed incorrectly, however, and the airplane stalls the result is a relatively strong acceleration at the instant of stall. Thus, the student in the real airplane knows immediately that the task has been performed incorrectly and it must be started over. In the FTD the visuals and stick force does not change much from slow through stall to a post-stalled condition. The key indicator that the maneuver has been executed incorrectly, a large acceleration, is not present. Therefore, the student may, continue in a post-stall configuration for some time before recognizing the true state of the aircraft. Once this happens, the student is forced to try to reason when the aircraft stalled so that the conditions just prior can be recognized for the next attempt to perform slow flight tasks. The lack of a significant indicator that the FTD has stalled makes it difficult to learn the conditions of the real aircraft just prior to the point of a real stall. The difficult is most likely a combination of both the lack of high frequency kinematics and difficulty in the feedback of error.

The Need for Further Investigation

Several questions arise and merit further investigation regarding the use of the Cessna 172 FTD in a flight training role, including: Is there a measurable difference in transfer of training to real flight when comparing FTDs with a wide FOV visual display and flight simulators with motion? In initial training, does the student have to accidentally stall several times before being able to determine the correct angle of attack? What, if any, are the high frequency aircraft motions and vibrations associated with slow flight not captured by the flight test data? Does a FTD provide too little feedback that an inadvertent stall has occurred? Would the incorporation of tactile cues during slow flight and stalls positively affect the student's ability to meet the PTS for these maneuvers?

RESULTS

The ever-increasing computational power used to drive ground-based flight simulations and flight training devices (FTDs) is enabling higher levels of fidelity at lower costs while accurately modeling specific aspects of flight. Currently, it is difficult for students to learn slow flight and stalls in an FTD even thought this flight regime was programmed with real flight test data. It is believed that the lack of feedback, at the moment of stall, is the primary reason for the difficulty. In the real aircraft there is a significant acceleration denoting the transition into a stall.

Without this motion feedback the stall condition is misperceived; the other stimuli, visual and proprioceptive, are too small to be distinguished for a new pilot. Researchers should investigate the use of an artificial seat "bump" or tilt to determine the nature and degree of this type of tactile cue's ability to affect a student pilot's perception of the stall transition in an FTD.

REFERENCES

Brady, T. (2003). Level 6 FTDs: The Business Case. On Council on Aviation Accreditation Simulation Symposium. Daytona Beach, FL: Embry-Riddle Aeronautical University.

Brandt, T., Wist, E., & Dichgans, J. (1975). Foreground and Background in Dynamic Spatial Orientation. Perception & Psychophysics, 17(5), 497-503.

Bürki-Cohen, J., Boothe, E. M., Soja, N. N., DiSario, R. D., Go, T. H., & Longridge, T. (2000). Simulator Fidelity - The Effect of Platform Motion. Proceedings of the Royal Aeronautical Society Conference on Flight Simulation.

Chung, W. (2000). A Review of Approaches to Determine the Effectiveness of Ground-Based Flight Simulation (1-10 No. AIAA-2000-4298). Vienna, VA: American Association of Aeronautics & Astronautics.

Embry-Riddle Aeronautical University. (2003a). Embry-Riddle Aeronautical University, Flight Line Daytona Beach. Retrieved November 10, 2003, from http://www.erau.edu/db/flightdb/fleet.html

Embry-Riddle Aeronautical University. (2003b). Flight Test Report for the Cessna 172 (No. Revision Original, Document Number CERT 172). Daytona Beach, FL: Embry-Riddle Aeronautical University.

Federal Aviation Administration. (1992). Advisory Circular (AC) 120-45A, Airplane Flight Training Device Qualification: U.S. Department of Transportation.

Federal Aviation Administration. (1997). Private Pilot for Airplane Single-Engine Land Practical Test Standards. Washington: Aviation Supplies & Academics.

Frasca International Inc. (2003). Cessna 172 FTD. Urbana, IL.

Hope, J. E. (2003). Simulation in Aviation Training: Past - Present - Future? On Council on Aviation Accreditation Simulation Symposium. Daytona Beach, FL: Embry-Riddle Aeronautical University.

Longridge, T., Bürki-Cohen, J., Go, T. H., & Kendra, A. J. (2001). Simulator Fidelity Considerations for Training and Evaluating of Today's Airline Pilots. On The Proceedings of the International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University Press.

Szczepanski, C., & Leland, D. (2000). Move or Not to Move? Continuous Question (No. AIAA-2000-4297): American Association of Aeronautics & Astronautics.

Waag, W. L. (1981). Training Effectiveness of Visual and Motion Simulation (No. AFHRL-TR-79-72). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

WHY ARE ROUTINE FLIGHT OPERATIONS KILLING PILOTS AND THEIR PASSENGERS?

Robert Baron
The Aviation Consulting Group

ABSTRACT

Routine flight operations present pilots with a myriad of latent threats. A scenario is presented that exemplifies how a routine flight operation can end in disaster. The pilot's complex and dynamic psycho-cognitive behaviors are analyzed and show that satisfactory technical training alone does not make a safe pilot.

More emphasis needs to be put on the "human system," the most likely system to fail in flight.Recommendations address the areas where intervention and education may mitigate some of these issues.

Keywords: Pilot Training; Controlled Flight Into Terrain; Routine Flight Operations

INTRODUCTION

The crew had just finished recurrent training. The instructor praised both pilots for exemplary performance in the simulator, and attested to that fact with positive comments on both pilot's grade sheets. Both pilots had thousands of hours of flight experience and thousands of hours of combined time in the particular make and model they were flying. They were back on the line the following day.

Their first leg back on the line proved tragic, as both pilots, and 27 passengers were killed when the aircraft descended prematurely on a non-precision approach at night. As usual, the first question asked was "what happened?" How could such an experienced and well-trained crew commit this type of error, especially the day after they received recurrent training and were commended on their skills?

This is but one example of a routine flight operation gone terribly wrong. The pilots had flown into this airport on numerous occasions, albeit during daylight hours. The weather was reported to be good VFR (Visual Flight Rules), the wind was calm, and the runway was 10,000 feet long. VASI's were available to establish a proper glide angle to the runway threshold. But for some reason, the crew descended below the VASI's prematurely, causing the aircraft to impact the ground a few miles from the end of the runway. Another classic CFIT (Controlled Flight Into Terrain) accident has occurred. A perfectly airworthy airplane, under complete control, was flown unintentionally into the ground without any prior awareness by the flightcrew.

This example shows us, in its purest form, where technical training ends and human factors begin. This type of accident occurs more frequently than one would be led to believe. The pilots *assumed* this was a routine flight. After all, the weather was good and there was nothing wrong with their aircraft just minutes before landing.

As it turns out, the captain, who was the pilot flying, was compelled to attempt a night *visual* approach to the runway, even though the VOR Runway 17 *instrument* approach was briefed and set up earlier. When the first officer queried the captain on this discrepancy, the captain replied that he "wanted to shoot the visual approach since the weather was good and it would save some time." That was the last discussion recorded on the CVR (Cockpit Voice Recorder) before the sound of impact, approximately two minutes later.

In a macro-analysis of this accident, it was concluded that the aircraft impacted rising terrain approximately 2.3 miles from the runway threshold. Additionally, the aircraft was 800 feet lower than it should have been at that point *if* the pilots had executed the VOR Runway 17 instrument approach. For a technically proficient crew, which this crew was, the instrument approach alternative would have been routine, and the outcome would likely have had a more successful result.

WHY?

This scenario might be considered a quintessential example of failure in human performance. A fully trained, experienced, and competent flight crew committed a series of errors that lead to a Controlled Flight Into Terrain accident. **Why?**

"Why," as it relates to aviation accidents, is a very complex and challenging question. The attempt to analyze a pilot's cognitive thought processes extends far beyond the scope of this paper. After all, only the pilot can really answer the question "what were you thinking?" We can however, use deductive reasoning to look at where some of the problems manifest themselves.

For the sake of simplification, we will look at only two distinct areas, (1) Training facility weaknesses, and (2) Psycho-cognitive threats during routine flight. A breakdown in these areas can pave the way for the highest and most undesirable event; an accident.

TRAINING FACILITY WEAKNESSES

Not enough emphasis put on the most unreliable system in the aircraft, (the pilot):
Pilot training on a specific aircraft can last anywhere from a few days, up to a few months, depending on the type of aircraft. Training facilities put a large amount of effort into teaching systems in the shortest amount of time possible. And while the importance of good systems knowledge is undeniably important, the most failure-prone system, the pilot, is often overlooked or disregarded.

Crew Resource Management training is weak or non-existent at many facilities:
Although many training facilities have begun to incorporate a fair amount of CRM training into their programs, some facilities do not have the time or properly trained facilitators to make a significant impact during a normal training period. After a 2 hr training period, a single CRM debriefing comment by the simulator instructor to the affect of "you should speak up more next time," does not adequately address the problem.

Simulator training time is too compressed. Many emergency/abnormal scenarios that are combined to save time are unfounded and are extremely unlikely to occur in real life:
Some facilities, in the interest of time, will combine multiple emergency/abnormal scenarios. It is extremely improbable that a modern airliner or business jet will experience an engine failure *and* a total hydraulic failure at the same exact time, and that the pilots will have to execute a circle-to-land approach with the weather right at landing minimums. Yet, these are the types of scenarios that some facilities are training and testing pilots on.

"Routine" flight operations are under-emphasized. Yet, routine flight operations claim many more lives than non-routine operations:
Inasmuch as the previous topic depicted an overdose of non-realistic scenarios, this topic highlights a relatively untouched realm of training: Routine flight operations. Realistically, engine failures, hydraulic failures, and popped circuit breakers are not killing pilots and their passengers. The largest number of crashes and fatalities occur when nothing is mechanically wrong with the aircraft.

PSYCHO-COGNITIVE THREATS DURING ROUTINE FLIGHT

The next level picks up where the training ends. At this point, the crew has satisfactorily completed recurrent training and is back on the flight line. All incidences referenced from this point forward are considered "in-flight."

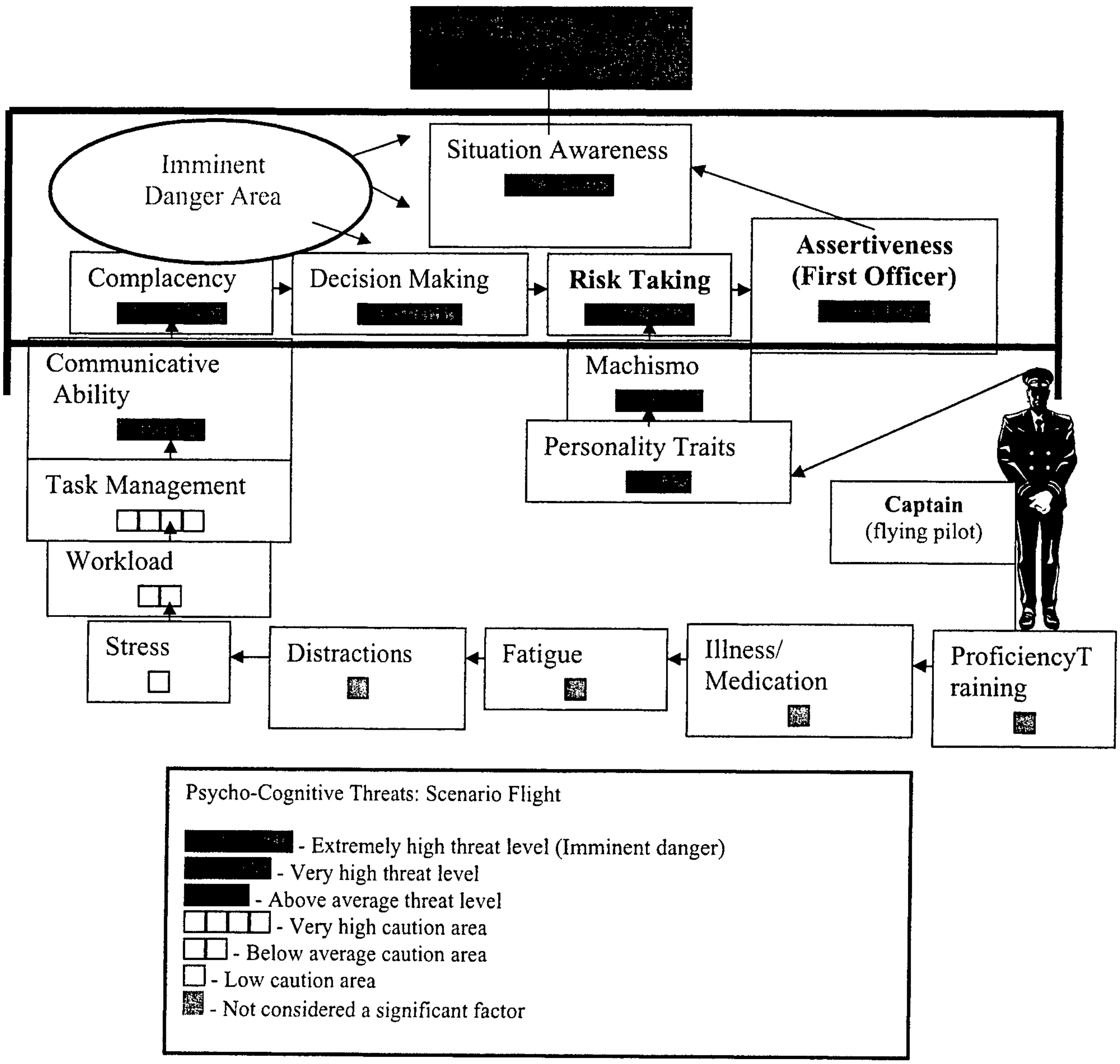
Keep in mind that the scenario accident was due to a failure in human performance, and not a mechanical malfunction. In other words, the problems were not easily identifiable in training, but they became blatantly clear later on.

During flight, the pilot's psycho-cognitive system performs like a computer, inputting thousands of bits of information, with the associated action commands performed as an output. Occasionally, there is a "short circuit" in these processes and the stage is set for problems.

The following items break down the scenario accident into CRM marker clusters, as defined in FAA Advisory Circular 120-51D. The author has incorporated additional clusters for clarity. Refer to the figure on the next page for a graphical flow of the Captain's behavioral patterns.

Proficiency Training- The crew was proficient with no training weaknesses noted.

Illness/Medication- Neither pilot tested positive for alcohol or drugs, including over-the-counter medication.
Fatigue- The crew was well rested
Distractions- Distractions were not considered a significant factor in the accident.
Stress- Stress was low. During the approach phase of flight, stress levels will normally be somewhat elevated.
Workload- Workload was considered routine. During the approach phase of flight, workload will normally be highest.



Task Management- Management of tasks became somewhat ambiguous. A last minute change of the approach procedure by the Captain was a factor.
Communicative Ability- The Captain's decision to change the approach procedure and not re-brief was the beginning of the "red zone."
Complacency- The Captain displayed signs of complacency. He considered this a routine approach and the weather was good. He had also been into that airport many times before.

Decision Making- 'Complacency' likely influenced the 'Decision-Making' misjudgment.
Personality Traits- Ingrained and hard to change. The Captain's personality included a large amount of 'Machismo,' according to pilots who had flown with him in the past.
Risk Taking- This is the area where 'Decision Making' and 'Machismo' converge. The Captain had decided to "take the risk."
Assertiveness- The First Officer may have had the last chance to trap the Captain's bad judgment. However, the F/O did not speak up and challenge the Captain.
Situation Awareness- Due to all the previous unmitigated behavior problems, the crew experienced a loss of 'Situation Awareness.' A perfectly airworthy aircraft was flown into the ground without any prior awareness by the flightcrew.

RECOMMENDATIONS

This accident scenario is a classic example of human error in its purest form. Human performance is a complex and challenging science. More attention needs to be focused on "why pilots do some of the things they do (or don't do)," and what the associated consequences of those actions might be. Recommendations for improving the system should address the following areas:

1. Training facilities must put more emphasis on human performance. This might be accomplished with a stand-alone training module that addresses this area in more detail.
2. CRM training needs to become mandated for *all* flight operations (currently, the FAA does not require Part 135 on-demand charter pilots to have formal CRM training).
3. CRM Facilitators should have some *formal* training on proper training and debriefing methods.
4. Simulator training should concentrate on more realistic flight and emergency/abnormal scenarios and avoid simultaneous unrelated systems failures, compounded by the worst possible weather.
5. During ground school and simulator training, an emphasis should be made that "routine flight operations" can become a significant threat and complacency can exacerbate the problem.
6. Pilot selection, particularly below the airline level (i.e., Part 135 charter and corporate aviation) should implement or expand on the use of psychological testing.
7. All pilots should be required to take a formal (credit or non-credit) course on psychology.

CONCLUSION

In summary, routine flight operations, as benign as it sounds, can and will continue to be a latent threat to flightcrews. Training facilities and pilots need to increase their vigilance of this threat and expand on safeguards and awareness training.

On a research level, both NASA and FAA have stepped up investigation into this area. NASA's research on Cognitive Performance in Aviation Training and Operations, and FAA's AAR-100 Human Factors Division, continue to provide valuable data for incorporation into aviation training programs at all levels.

REFERENCES

Federal Aviation Administration. (2001). Crew Resource Management Training. Advisory Circular 120-51D. Retrieved January 10, 2004, from http://www1.faa.gov/avr/afs/acs/120-51d.pdf
Federal Aviation Administration. AAR-100 Human Factors Division. Available at http://www.hf.faa.gov
National Aeronautics and Space Administration. Research on Cognitive Performance in Aviation Training and Operations. Available at http://olias.arc.nasa.gov/flightcognition/index.html

DESIGN OF LEARNING ENVIRONMENTS FOR COMPLEX SYSTEM ARCHITECTURES:EXPANDING THE 'KEYHOLE' VIA DUAL-MODE THEORY

Daniela Kratchounova, Stephen Fiore, Florian Jentsch
University of Central Florida

ABSTRACT

Current learning environments for complex system architectures, such as aircraft avionics system, seem to be based mostly on memorization of procedures. Pilots are presented with a limited, or "keyhole" (Woods & Watts, 1997) view of the system, where relationships between different subsystems are not apparent. This may cause pilots to become "disoriented" and "lost" in the "space" of multiple subsystems. Successfully integrating what is to be learned into a conceptual framework is a basic characteristic of a training system design, and its implementation into the design of learning environments for complex system architectures is critical (Hutchins, 1992). The proposed design approach integrates theories about spatial knowledge acquisition (e.g., dual-mode theory - Colle & Reid, 1998) to support knowledge acquisition about complex system architectures and provide appropriate conceptual framework for navigating an aircraft avionics system. Ultimately this would result in improved pilot performance and reduced number of automation surprises.

Key words: macro-spatial knowledge acquisition, navigation, complex system architectures, avionics.

INTRODUCTION

In recent years, various industries have experienced the introduction of new complex automated systems. For example, new computer-based flight systems, like flight management systems (FMS), part of the avionics system on an airplane, have been introduced for increased efficiency, precision, and safety. However, with such automation technology a new category of incidents, known as "automation surprise," (Sarter et al., 1997) has been introduced as a result of mismatches between the behavior of the technology and users' expectations (Feary et el., 1997). In the case of an FMS, when pilots don't understand from a conceptual level how the automated flight system works it is easy to be surprised. Hutchins (1992) states that training programs often lack a strong conceptual and theoretical component that could support a better understanding of system behavior, and this shortcoming is due, in part, to increasing system complexity.

Development of training programs for operating complex automated systems relies more and more on interactive computer-based learning systems. Such computer-based learning environments have become a fashionable media for training and hold many promising applications, especially with the advent of powerful computer technology. Yet, current training programs are often based mostly on the memorization of procedures. A trainee is generally presented with a limited, restricted, or "keyhole" (Woods & Watts, 1997) view of the system architecture, where relationships between different subsystems are not apparent. What is directly visible through the "keyhole" view provided by a computer monitor does not reveal the paths, underlying processes, or alternative sequences of action required to navigate through the larger system. This may cause trainees to become "disoriented" and "lost" in a space of multiple systems with high levels of complexity and integration. As complexity of interactive systems increase, more could be hidden from the user making systems training more difficult.

Moreover, there is little support given to the trainee while learning how to carry out the tasks of operating a system (Feary et al., 1997). Hutchins (1992) points out that the learning outcome will be much better when what is learned can be integrated into a conceptual framework. In the case of pilots, an appropriate conceptual framework for navigating an aircraft avionics system may improve pilot performance and reduce the number of automation surprises.

This paper presents a new approach for designing learning environments for complex system architectures such as an aircraft avionics system. Specific emphasis is given to the design of a learning environment for FMS. This new approach is based on the integration of theories and findings from areas such as human spatial knowledge acquisition in real and virtual environments to support knowledge acquisition of complex system architectures. More specifically, the application of dual-mode theory (Colle & Reid, 1998) in the design of both the interface and the instruction content of a learning environment are investigated using a computer-based FMS simulation. The focus is on two major areas: 1) theories about acquisition of spatial knowledge in real and virtual environments and 2) application of such theories into the design of learning environments for complex system architectures.

PROBLEM STATEMENT

Current training programs for complex automated systems, such as the avionics system of an aircraft and more specifically FMS, lack robust conceptual and theoretical frameworks at both interface design and instructional system design levels. As a result when there is a mismatch between the automated system's behavior and the flightcrew expectations, automation surprise may occur. Automation surprises are well documented in the cockpits of advanced commercial aircraft (all equipped with FMS) and several fatal crashes and other incidents are attributed to them (Sarter et al., 1997)

Review of related literature

Learning to navigate in environments such as complex systems architectures has some characteristics similar to the acquisition of spatial knowledge in a real, macro spaces including that the desired objective may not be readily visible. Also, the relationships between global and local views of different 'areas' may not be directly seen by the viewer, and thus appear discontinuous or disconnected. These characteristics can make learning to navigate any complex space difficult.

Different terms have been used in the literature to describe navigation, wayfinding, and route learning, but all of them generally describe how people get from one point to another in a real or virtual environment. Navigation is a process inherently cognitive in nature (Nash et el., 2000) and a good understanding of how people acquire spatial knowledge and use it may prove beneficial to the design of training programs for complex system "spaces". In many cases, users need to be able to locate a site within the virtual space of system architecture and traverse it in order to complete a particular operational or training task. Thus, they must maintain an orientation of important subsystems and be aware of how to "travel" between them.

Maintaining orientation in complex system architectures can be challenging. This is likely due in part to the "keyhole" effect, as described by Woods and Watts (1997). The user has a limited view of the entire 'space' in a similar way as a user has a limited view of a large physical space by looking through the a keyhole. This keyhole provides users with a limited view of the entire architecture and requires that they be able to integrate separate views into an integrated whole. Based on their work, it is herein suggested that the difficulties users experience in navigating or traversing complex system architectures are due to these large "spaces" being presented to users via a narrow keyhole (i.e., the view from the computer screen or the FMS Control and Display Unit). The required integration of the separate views is likely to be complicated even further by what Woods and Watts (1997) describe as the "art museum" effect. This occurs when a user who has examined many items or layers of an interface through a computer "keyhole" becomes overwhelmed. The "art museum" effect acts on both a local and global scale. Users not only lose track of the individual features of the "art" pieces (i.e., subsystems or layers of interface) already seen, but the big picture, the larger, global structure is also lost. It is analogous to getting lost in a museum with many rooms of artwork.

Complex system architectures, such as FMS, are very rich in subsystems and modes. The lack of understanding by users of system's internal architecture may lead to a lack of understanding of what the system is doing or going to do next and why (Billings, 1997). Furthermore, if system architecture is inherently complex and difficult to visualize (i.e. aircraft avionics system), knowledge acquisition may be facilitated by a learning environment that presents users with necessary support tools. In order to develop such support tools it is essential to understand how humans acquire and use spatial knowledge (Colle & Reid, 1998).

Traditionally, the way spatial knowledge is acquired has been described by the Landmark-Route-Survey (LRS) model (Thorndyke & Hayes-Roth, 1982). According to this model there are three levels of spatial knowledge acquisition: landmark, route, and survey knowledge; and each is a reflection of the qualitative and quantitative changes in understanding that take place when an environment is learned. It has been implicit that these representations are acquired in successive stages (Siegel & White, 1975). First, some information about landmarks is acquired. Then, a procedural knowledge about specific routes between those landmarks is developed. Finally, survey knowledge can be constructed.

The LRS model, although very powerful, has not been able to meet some challenges presented in the literature. More specifically, the order of spatial knowledge acquisition expected by the LRS model has not always been found. Colle and Reid (1998) propose a dual-mode model for spatial knowledge acquisition. This model suggests that there are two modes of spatial knowledge acquisition, both engaged in early stages of environment exploration: the gaze viewing mode and route tour mode. The gaze viewing mode is a perceptually-driven mode. Gaze view representations are obtained of objects that are within the spatial span of the observer. By rotating the head and with small movements, observers create a three-dimensional exocentric representation of the local region. In contrast, the route tour mode leads to a more egocentric representation of larger areas. In the route tour mode observers gain knowledge of how to get from place to place in terms of actions that need to be taken to get to destinations. The knowledge that is gained in this mode is in reference to larger areas that are outside of the spatial span and passed through quickly. For these reasons, the spatial information gained in this mode is more cognitively constructed as opposed to perceptually driven. The mode that is evoked depends on both what the user is doing and characteristics of the environment. An important aspect of this model is that the two modes may operate in conjunction with each other and are not necessarily evoked in successive stages.

In the dual-mode model, a distinction is also made between a local region and a distant region. Based on that distinction, Colle and Reid (1998) introduce a concept called "the room effect," which describes a phenomenon in which humans rapidly acquire local survey knowledge from spatial information in a room. The rooms represent local areas as hierarchies to facilitate navigation. The knowledge about these local areas can serve as a vehicle to learn the larger macro space.

When designing complex automated systems and associated learning environments, in order to facilitate the "room effect" the challenge is to consider and develop strategies for instantiating the "room", its contents, and inter-room traversal strategies, which are consistent with the theory and the intended application. In the case of the FMS these considerations may require applying the dual mode theory to the learning environment interface and instructional content. Essentially, this means presenting the user with additional supporting information in the form of other contextually and task relevant cockpit information (i.e. what else is going on within the system) can be viewed as adding "rooms" of information. Users then can evoke the gaze viewing mode of the dual mode theory to gain knowledge within each of the displays that are presented. With the help of relevant contextual information to connect the different displays, users can evoke the route tour mode to tie the knowledge gained from each display into more of a global understanding of what the automated system is doing and it will do next and ultimately avoid automation surprises.

Hypothesis

The implementation of the dual-mode theory into the design of learning environments for complex system architectures, such as the FMS, will support the development of improved knowledge structure about the system by providing a theoretical and conceptual framework for understanding the FMS internal architecture and its interaction with other avionics systems.

METHOD

Participants

Twenty-four undergraduate students from an aeronautical university volunteered to participate in the experiment. There were 13 male and 11 female. The average age of the participants was 21 years. Volunteers were rewarded with extra class credit for participating in the experiment and were treated in accordance with the "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 1992). All participants were recruited from HF315 "Human factors and Automation" class taught at the aeronautical university mentioned above. Participants' current grades in this class were used to determine their level of expertise.

Apparatus

Flight Management System simulation software (Aerosim Technologies Inc. G-IV v 2.0; G-V v 2.0) was used to develop the design of two learning environments. The two environments are referred to as "no context" and "context". Each learning environment consisted of unique versions of the same basic instructional content; a step-by-step instruction of how to perform Lateral Direct-To function of the FMS in a printed form. All the information necessary for the completion of the task was presented in the form of text boxes. The complete procedure was designed using "Gulfstream-V FMZ Series FMS Pilot's Operating Manual" and can be performed by the CDU (Cockpit Display Unit) alone. Thus, both conditions included are fully functional simulation of the CDU and the supporting textual information. Both learning environments contained a still image of the Gulfstream-V aircraft cockpit as a background.

The dual mode model described earlier was used to manipulate the availability of contextual framework in both conditions. Within the "no context" environment, the Lateral Direct To procedure was presented with no other functional instruments available in the simulation for cross-reference except for the CDU. Because of the lack of contextual cues and cross-referencing instruments, this environment is referred to as the "no context" learning environment and represents an environment viewed through a small "keyhole".

Within the "context" environment other cockpit displays or views relevant to the concurrent underlying processes in the system were shown in addition to the CDU, including the Navigation Display and Flight Guidance Controller. Additionally, some important knowledge landmarks about the overall system integration and interaction were included at the beginning of the lesson developed for this condition in the form of text boxes.

Task

Each participant completed the training task under one of the two conditions, while simultaneously providing verbal feedback. Following the training task each participant performed a card-sorting task.

Procedure

Participants were first briefed on the details of the study. They were then randomly assigned to either the control (no context) or experimental (context) group. Participants were instructed that there was a 10-minute time limit to complete the training exercise. The primary objective of the exercise was to learn the steps to perform the FMS task to the point that they could perform the task without any instruction or guidance. Participants were allowed to review the training material as many times as desirable within the time limit.

At the conclusion of testing, participants were asked to perform a card-sorting task. For the card-sorting task a list of concepts and a card-sorting answer sheet were provided. The thirty-six concepts were listed in a random order. The participants were required to use each concept once and to place all of the concepts into one of the four categories: "performance", "interface", "procedure" and "control".

Design

A between-subjects design was used for the study. The independent variable (IV) was learning environment design. There were two levels of the IV: "no context" and "context". There was one dependent variable (DV) in this experiment. The DV was the level of overlap between knowledge structures about the system elicited by an expert and each participant based on their card-sorting task score.

RESULTS

An Analysis of Covariance (ANCOVA) was conducted on card-sorting scores. Analysis was performed using SPSS 10.1 for Windows. Unless otherwise stated, an alpha level of .05 was used for all analyses. The independent variable was group assignment ("context" and "no context"). The dependent variable was the participant's score on the card-sorting task. There was one covariate: level of expertise based on class performance (HF 315 Test 2 scores). Tests of between-subject effects showed significant effect of the covariate, $F\,(2, 24)=20.51$, $p<.005$, $\text{Eta}^2 = .494$. After adjustment by covariate, the adjusted means for group assignment for "context" and "no context" were 42.99, and 46.00 respectively. The results of the ANCOVA indicated no significant main effect for group assignment, $F\,(1, 24)< 1.581$, $p =.454$, $\text{Eta}^2=.027$.

Although the statistical analysis of the card-sorting task showed no significant difference between the group means, the results of the verbal protocol indicated anecdotal evidence that the participants in the "no context" group were looking for additional cockpit information to be provided within the learning environment.

DISCUSSION

Of primary concern for this study were of the effects of the implementation of dual mode model in the design of learning environments for complex system architectures on the development of a knowledge structure about the system. It was anticipated that implementation of the dual-mode theory into the design of learning environments for complex system architectures, such as the FMS, will improve training outcomes by providing a theoretical and conceptual framework for understanding the FMS internal architecture and its interaction with other avionics systems. The results from the card-sorting task showed no difference in the level of overlap between knowledge structures about the system elicited by an expert and participants in either of the two conditions. However, many participants verbalized the need of additional cockpit display information, especially those in the "no context" group.

There are several possible reasons for these results that warrant further study. First, the training task difficulty was not appropriate for the level of expertise of the participants. It was assumed based on the class curriculum for HF 315 "Human factors and Automation" that the level of expertise would be consistent across participants, as they all would have a basic knowledge about automated aviation systems. However, after conducting the analysis, there was a difference in prior knowledge across participants.

The second reason why there was no significant difference found between groups in the card sorting task scores could be the level of treatment. Only two levels of treatment were used in this study, thus there was no clear discrimination between too much, or too little context within the continuum of treatment levels for any given level of expertise.

Third, only one expert's knowledge structure was used to evaluate the card sorting results in this experiment. This may have imposed some limitations on the statistical conclusions coming from fact that participants' scores were determined by the amount of overlap between the knowledge structure elicited by just one expert and the knowledge structure elicited by each of them. Consequently, there were no clearly defined criteria to determine whether this particular expert knowledge structure was the one that would ultimately lead to an optimal trainees' knowledge structure.

Fourth, participants' scores were calculated by only counting the hits, i.e. the correct placement of concepts into card-sorting piles. More precise scoring technique would also include the correct rejections (Fiore et al., 2003).

Fifth, participants' motivation may have also affected the results of this study. This could be due to the fact that merely participating in the experiment earned the participant extra credit. There was no benefit, nor risk related to performance. The outcome could change if participants were to be rewarded for scores over a certain threshold.

Finally, at a system level, there were no specific guidelines or existing implementations where the dual mode, landmark, and expanded keyhole approach has been implemented. The developers of the two learning environment tested in this experiment had expertise in avionics training using only traditional training systems development approaches. The evaluation of the theories and their practical applications and strategies for implementing has not been performed. Therefore, the strategies implemented here are speculative and require validation.

Future Research

The following outlines several directions of future research. First, there are several modifications to the methodology that may lead to different results, these include taking into account previous experience when selecting the sample, providing stronger user centered landmarks in the "context" training environment, conducting the training in a more interactive setting, and testing using a different set of experts' knowledge structures.

Second, more research is needed to create and validate strong metaphorical landmarks that can be used in different training environments to connect between the separate displays of information (rooms). These landmarks should facilitate the user's task of building mental models of the overall system.

Finally, strategies need to be developed for determining the optimal size of the keyhole, which allows for capturing the "Big Picture", without sacrificing local knowledge, thus minimizing the "Art Museum" effect.

REFERENCES

American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. American Psychologist, 47, 1597-1611

Barlow, Todd (2001). Acquisition of route knowledge through navigation in virtual environment. *International journal of cognitive ergonomics*, Vol. 5(3), p 279-295

Billings, C.E. (1997). Aviation automation: The search for a human-centered approach. Lawrence Erlbaum Associates, Mahwah, New Jersey

Bowman, D.A., Davis, E. t., Hodges, L.F., Badre A. N., (1999). Maintaining Spatial orientation during travel in an immersive virtual environment. Presence, Vol. 8(6), p 618-631

Colle, H.A, Reid, G.B.,(1998) The room effect: Metric Spatial Knowledge of Local and Separated Regions. Presence, Vol. 7/2, p 116-128

Darken, R.P., Peterson, B. (2002), Spatial Orientation, Wayfinding, and Representation, in Stanney, K.M. Handbook of Virtual Environments

Feary, M., McCrobie, D., Alkin, M., Sherry, L., Polson, P., Palmer, E., McQuinn, N. (1997). Aiding Vertical Navigation Understanding. Technical Report prepared for NASA Ames Research Center Honeywell ATSD

Fiore, S. M., Cuevas, H. M., & Oser, R. L. (2003). A picture is worth a thousand connections: The facilitative effects of diagrams on task performance and mental model development. Computers in Human Behavior, 19, 185-199

Gamberini, L., Bulssolon, S. (2001) Human Navigation in Electronic Environments. Cyber Psychology and Behaviour, Vol. 4, Number 1

Jansen-Osmann, P. (2002). Using desktop virtual environments to investigate the role of landmarks. Computers in Human Behavior, Vol. 18, p 427-436

Johnson, David M. & Stewart, John E., Jr. (1999). Use of virtual environments for the acquisition of spatial knowledge: Comparison among different visual displays. Military psychology, Vol. 11(2), p 129-148

Koh, G., von Weigant, T.E., Garnett, R.L., Durlach, N.I.,Shinn-Cunningham, B. (1999). Use of Virtual Environments for Acquiring Configurational Knowledge about Specific Real-World Spaces. Presence, Vol. 8(6), p 632-656

Richardson, Anthony E., Montello, Daniel R., Hegarty, Mary (1999). Spatial knowledge acquisition from maps and from navigation in real and virtual environments. Memory and Cognition, Vol. 27(4), p 741-750

Rossano, Matt J., West, Sandra O., Wayne, Michael C., Chase, Richard B. (1999). The acquisition of route and survey knowledge from computer models. Journal of Environmental Psychology, Vol. 19(2), p 101-115

Ruddle, R.A., Howes, A., Payne, S.J., Jones, D.M. (2000). The effects of hyperlinks on navigation in virtual environments. Int. J. Human-Computer Studies, Vol. 53, 551-581

Sarter, N. B., Woods, D. D., and Billings, C. E. Automation surprises. In Salvendy, G., editor, Handbook of Human Factors and Ergonomics. John Wiley and Sons, 2nd edition, 1997

Schlender, D., Peters, O.H., Wienhofner, M., (2000). The effects of maps and textual information on navigation in a desktop environment. Spatial Cognition and Computation 2, p 421-433

Steck, S.D., Mallot, H.A., (2000). The role of global and local landmarks in virtual environment navigation. Presence, Vol. 9(1), p 69-83

Thorndyke, P.W., Hayes, R.B., (1982)Differences in spatial knowledge acquired from maps and navigation, Cognitive Psychology, Vol.14(4): p560-589

Waller, D., Hunt, E., Knapp, D., (1998). The transfer of spatial knowledge in virtual environment training. Presence, Vol. 7(2), p 129-143

Woods, D.D., Watts, J.C. (1997) How not to have to navigate through too many displays. In Handbook of Human-Computer Interaction

MILITARY TRAINING AND SIMULATION FOR THE NINTENDO GENERATION

Pete Muller
Potomac Training Corporation

Dylan Schmorrow
DARPA

Colby Raley
Strategic Analysis

ABSTRACT

The Department of Defense has invested in adapting commercial game technology to the training domain. In one class of games, the First Person Shooter (FPS), there is very little scientific evidence suggesting that any commercial First Person Shooter videogame produce improved training in tactics, techniques, and procedures over more traditional methods of instruction. This paper examines some recent applications of commercial gaming technology and describes some planned research by the Office of Naval Research to address some of these critical issues. This paper will shed some light on the critical differences between entertaining commercial games and military training simulations.

Keywords: Military Training, Simulation, Commercial Games, First Person Shooter

INTRODUCTION

Just as previous generations were forever changed by television, current generations are influenced by the video games they play. The average teenager may not clean up his room, but he can hold a dozen real time instant message conversations while listening to the latest music over the web and playing a video game. None of these technologies were available to his parents. This same teenager is the source of our military recruits and officer candidates and is profoundly different in many ways than previous generations. The DoD Research and Development community was relatively quick to adapt commercial games to a training context. Many games are fairly easy to modify (or "mod" in the gaming community), and there are dozens of games in use by the military. Many games involving strategy and tactics are computerized versions of the board games that have been successfully for years. Surprisingly, there is very little rigorous scientific evidence suggesting that First Person Shooter video games produce improved training in tactics, techniques, and procedures for infantry over more traditional methods of instruction. It is popular to cite anecdotal evidence of improved performance due to various video games, but there are no comprehensive studies that show the types of training that can be improved by various game technologies.

Background

The Marine Corps has led the DoD in adopting commercial gaming technology for infantry training. They evaluated close to thirty games in 1995 for their potential teaching value. None of the games met all of the training needs, but many of them could produce an environment where learning and training could take place. Over the years, this has evolved into the Marine Corps Infantry Tool Kit (ITK), which is a collection of Commercial Off The Shelf (COTS) games, modified COTS games, and custom built games. These games provide an ***environment*** in which the instructor can illustrate training points. Unlike more conventional computer based training programs that have tasks, conditions, and standards, these games are much more free flowing.

Virtual Technologies and Environments (VIRTE)

When the Office of Naval Research (ONR) began the Virtual Technologies and Environments (VIRTE) program in October 2001, one of the goals was to make sophisticated DoD training simulations as simple and intuitive to use as a video game. There is a need to get prototypes to the research community early, so that human performance research can influence fundamental design choices. Unfortunately, when simulation prototypes were previously built, they couldn't be replicated without exorbitant licensing and deployment costs. ONR needed to be able to hand out CDs, license free, to any Department of Defense user who wanted it. We discussed the Navy's needs with the gaming industry, and did a thorough review of available technologies and licensing issues. In addition, we studied whether we could achieve the same level of performance either using Open Source tools or Government Off the Shelf (GOTS) tools. Although non-disclosure agreements with the gaming technology vendors prevent us from sharing our results, we will share some general observations.

Game Consoles vs PC Based Games

Game consoles can be thought of as highly specialized PCs. We wanted to take advantage of the incredible hardware that is available on the game console market. This Christmas, $99 bought a Playstation 2 console with a game. In addition to a low cost, the game console provides a stable and standard hardware platform. This means that programmers don't have to worry about what graphics card is installed, how much RAM is available, etc. This makes developing and testing applications faster and easier.

Unfortunately, we found that all three of the major vendors, Microsoft (X Box), Sony (Playstation 2), and Nintendo (Game Cube) had no interest in supporting DoD training systems. They all lose money on the hardware and make their profit on licensing games. Their business model is simply not compatible with DoD training systems. It is, of course, possible to self-publish and buyout the required number of titles. Although, we considered this option, we didn't think it was a prudent use of our limited resources. We, instead, focused our efforts on making our training systems as "game console-like" as possible, using high end PCs.

Entertainment vs. Military Training

Although an entertaining experience is not impossible in a military training system, it is often at odds with training objectives. If we examine one of the most popular classes of game, the "First Person Shooter" (FPS), the distinctions will become clear. The FPS game is a first person view into the virtual world. Typically, the player looks through a computer monitor into the virtual world with much the same view as he has from his own body. The mission, in most FPS games, is to move about the environment and "kill" as many enemies as possible while not being killed or injured enroute to a goal. If the simulated enemy is realistic and can kill you easily, the game may not be entertaining. Similarly, if you cannot kill the enemy easily, the game may not be fun. Commercial game designers want you to be entertained and they have no qualms with modifying the application of the laws of physics or biology do that. In a military training simulation, it is critical to have realistic physics and human behavioral interactions.

Commercial games have unsophisticated Artificial Intelligence (AI) by DoD standards, although this is changing. Game AI is limited to a fraction of the computational resources that is available on a single personal computer. The game industry works very hard to make their characters appear to have sophisticated Artificial Intelligence, but much of that is done with simple and clever rule sets. DoD has concentrated on rich and complex human behaviors in simulation, often called Computer Generated Forces (CGF) without as much regard to computational resources. The game industry has not used the techniques pioneered by the DoD and the AI research community because they are too processor and memory intensive. While there has been some effort to bridge the gaps between the two extremes, we still have a long way to go.

Shooting a weapon with a keyboard or joystick does not help you become a better shot with an actual weapon. The argument is that playing in the virtual environment improves cognitive skills and can be a mechanism for team coordination of small teams. Spending time thinking about tactics and teamwork in a virtual environment certainly has merit for the warfighter. Is this type of training more effective than physically walking through a building? Is it more effective than looking at a 2D map and marking positions with a pencil? Would a 3D virtual walkthrough without the weapons be just as effective? These are some of the question that we hope to address in our research.

Playing some videogames may improve visual performance

In a study published in *Nature* this year, playing video games such as *Grand Theft Auto* and *Medal of Honor* actually improved performance in vision tests (Green 2003). Interestingly, playing the game Tetris had no effect. Although casual video game playing may seem to have little benefit, it is capable of radically altering visual attentional processing. While this has some interesting applications to military training, more research needs to be done.

VIRTE Approach

The VIRTE program is taking a unique approach to determining the effectiveness of game based training. Rather than modify an existing game, we are developing an entire environment based on game technology. Where game technology is adequate, we are using it. We are concentrating our research on areas that games currently don't address.

Virtual Environment (VE)

We began with an extensive analysis on what environmental features are needed for effective training in our domain. A realistic physical environment is one of the most important factors. Not just realistic from the visual perspective, but realistic in the physical sense. The walls must physically react to weapons' effects based on both the weapon and the composition of the wall. Unlike many games, where you are safe from bullets if you hide behind a gypsum wall, in our environment you will be injured or killed. Furniture is not a static feature of a room. Tables and chairs can be moved to provide obstacles, cover, and concealment. Doors do not open because a button is pushed on a joystick; they open when the appropriate physical force is exerted by the avatar on the virtual door. It is not enough for our environment to be consistent with itself.

Since we are building a networked DoD simulation, we have to share simulation state in real time with a potentially diverse set of simulations. If an artillery simulation destroys the wall of the building in which our infantry simulation is working, the infantry must instantly experience the effects of the artillery. It is not enough to "see" the destroyed wall—the wall's representation must be fundamentally changed so that the infantry can react appropriately.

Head Mounted Displays

While a large monitor is adequate for games and many training tasks, more immersive tasks require a Head Mounted Display (HMD). In addition to their high cost, high quality HMDs require significant bandwidth and this means cables for the near term. We are examining the trade space to see how a lower visual quality wireless HMD compares with a higher quality wired system. Another interesting technical challenge is that the infantry rifle "prop" is brought up to the face and very near the HMD. Proper site alignment is critical to the shooting task.

Locomotion and Tracking

Moving about the VE, or locomotion, is one of the critical tasks that we are examining. Most games use a joystick or a keyboard to navigate in the VE. Joysticks, keyboards, and game controllers are certainly inexpensive, but they have many disadvantages in precisely navigating a VE. One of the unique features of an infantry simulation is that the user always has their weapon ready to fire. The weapon provides a natural platform for both a locomotion and a tracking device. Many systems use a modified joystick mounted on the weapon to provide locomotion in the VE, but this does not eliminate the inherent drawbacks of using a joystick. We will be examining several alternatives to locomotion in a VE. Of course, the most natural way to locomote in a VE is to actually walk through it. This can be accomplished by having a significantly large tracked area and a HMD. Cable and people management is a significant issue, particularly when small teams are involved with rapid movement of large pieces of steel in their hands. A potential solution to this challenge is the Naval Research Labs (NRL) Gaiter system in which the user turns naturally and walks in place to control locomotion. The user is held in place by a harness that also serves to manage cables. Gaiter uses a series of cameras placed around the individual to precisely track the individual's movement and translate that into an avatar. The weapon and upper body movement are sent directly to the avatar, and walking in place is translated to normal walking in the avatar. Simple gestures such moving the leg to the side, translate to side steps and so on. While Gaiter greatly reduces the required footprint, it is still significant for a deployable military system. A new technology known as Strider is being developed at NRL in which the user is seated with their weapon. Like Gaiter, the upper body and weapon is directly transferred to the avatar, but the leg motion will be remapped to control locomotion. As the technologies mature, we will conduct a series of experiments to determine which technologies are best suited for Marines and Seals.

DISCUSION

Videogame technologies form an important of our DoD training arsenal. By leveraging commercial and open source video game technologies, DoD researchers can concentrate on solving real world training problems and exploring technologies that are too expensive and fragile for the mass market.

REFERENCES

Green, C. S. & Bavelier, D. Action video game modifies visual selective attention. Nature, 423, 534 - 537, (2003)

UNMANNED AERIAL VEHICLES

PSYCHOPHYSIOLOGICALLY DETERMINED ADAPTIVE AIDING IN A SIMULATED UCAV TASK

Glenn F. Wilson, Christopher A. Russell
Air Force Research Laboratory

ABSTRACT

Two levels of task difficulty in an uninhabited combat air vehicle simulator were used to manipulate the cognitive workload of subjects performing a target identification task. Psychophysiological data were used to assess operator functional state using artificial neural networks (ANN). Adaptive aiding was provided when the operator's workload was deemed to be high by the ANN. The adaptive aiding improved the hit rate on the targets and the number of times that the weapons release points were successfully met. These results demonstrate that psychophysiologically determined operator functional state estimates can be used in complex operational environments to enhance operator performance.

Keywords: Adaptive aiding, artificial neural networks, psychophysiology, performance

INTRODUCTION

Degraded system performance and errors occur when the cognitive capabilities of the human operator are exceeded. One of the important factors determining the functional state of the operator is the level of cognitive demand placed on the operator by the system/task. If the current functional state of the operator is sufficient to deal with the system demands then the probability of degraded performance and errors is reduced. Coupling system demands with the operator's momentary functional capabilities should improve overall system performance. Numerous factors, in addition to system demands, cause the cognitive capabilities of human operators to fluctuate. Other detrimental factors that contribute to the operator's functional state include fatigue, circadian dysrhythmia and illness (Wilson & Schlegel, in press). System demands and operator functional state typically are not dynamically matched. System demands depend only upon the task and it is typical to assume that the operator has sufficient cognitive capacity to perform the required tasks. If the task demands exceed the momentary capabilities of the human operator then performance may degrade. Operator functional state characteristics can vary from moment-to-moment in response to changing task demands in the context of the internal characteristics of the operator. If the operator's cognitive capabilities do not meet the requirements for system operation then it may be possible to adapt the system demands such that they match the momentary functional state of the operator (Rouse, 1988). For example, if high levels of cognitive task demand exceed the momentary capabilities of the operator then the level of the task demands placed on the operator could be reduced. This could be accomplished by having the system assume some of the required functions or delaying them until the operator is capable of re-assuming the task. For this strategy to work the momentary functional state of the operator must be very accurately assessed. The dynamic nature of the adapting system must not exceed the operator's capabilities or optimal performance will not occur.

This paper describes a project in which the functional state of Uninhabited Combat Air Vehicle (UCAV) operators was assessed using psychophysiological measures, on-line, while they performed tasks having varying levels of cognitive difficulty. Previous research has shown that psychophysiological measures can be used to assess operator functional state on-line (Freeman, Mikulka, Prinzel, & Scerbo, 1999; Wilson & Russell, in press). This information was used to modify the difficulty of the primary task to determine if operator performance could be improved. A complex, simulated UCAV attack scenario was used in which each operator was simultaneously responsible for four vehicles and was required to locate and designate targets using pre-established rules.

METHODS

Five volunteers were trained to stable performance on a simulated UCAV task. The task required the subjects to monitor the progress of four autonomous vehicles as they flew a preplanned bombing mission. When the vehicles reached designated points, radar images of the target area were provided to the subjects. The subjects performed a visual search of the images and using a set of priorities selected six of the targets to be marked for bombing. The vehicles flew a preplanned mission and the subjects determined the order of image presentation from each vehicle. They were required to find and designate six targets in order to complete target selection by a pre-set time. Three categories of targets were used and the subjects were required to use a predetermined set of priorities when selecting targets. If the targets were not selected and/or the weapons release command was not given in time, the bombs from that vehicle could not be released thereby reducing the effectiveness of the entire mission for that vehicle. The complexity of the images was presented at two levels. The more difficult contained a larger number of distracters and required more complex decisions concerning target priority. Simultaneously, the subjects monitored the well-being of each vehicle by observing messages showing potential vehicle problems such as loss of communication. Memory was manipulated by having them keep up to four aircraft-problem combinations in memory until a command was given which signified which one had to be fixed. The subjects then selected the appropriate vehicle from a pull down menu and using other pull down menus found and selected the appropriate fix for the indicated vehicle problem. The easy conditions took approximately 3 to 4 minutes while the difficult conditions took 4 to 5 minutes to complete.

The number of correctly selected targets (hits), the number of designated mean points of impact (DMPI) placed and whether or not the command to release the weapons was executed in time were recorded. These data permit measurement of how accurately the subjects located targets (hits), how many targets were designated for each vehicle and if the subjects were able accomplish target identification and designation in the allotted time. The subjects gave estimates of their mental workload using the NASA TLX. Paired t-tests were used to test for significant differences between conditions for the various variables. One-tailed tests were used and $p < 0.5$.

Five channels of EEG, ECG, vertical and horizontal EOG were recorded. The EEG data were recorded from scalp sites F7, Fz, Pz, T5 and O2 of the 10/20 electrode system. Electrodes attached to the mastoid processes were used as reference and ground. These data were amplified and filtered by a small, subject worn, telemetry device. Our NuWAM software system performed the psychophysiological data reduction on-line. EEG power in five bands, heart rate, and blink rate were calculated from the raw data every second using a five second window with a four second overlap. These reduced data were provided to an artificial neural network (ANN). The ANN was trained by providing examples of psychophysiological data which represented periods of low and high task difficulty. Separate ANNs were trained for each subject. Then during subsequent task performance the ANN provided estimates of the subject's state every second. Three conditions were used. 1) No adaptive aiding during which only subject performance and ANN accuracy were recorded. 2) Adaptive aiding, when the ANN estimates indicated that the subject was in a high state of cognitive workload then the UCAV task was modified such the cognitive demands on the subject were reduced. This was accomplished by decreasing the velocity of the vehicle whose targets were being evaluated thus giving the operator more time to evaluate the images and select the targets. This gave them more time to complete target selection before the weapons release point was reached. 3) Random aiding during which aiding was provided randomly during the trial for a time equal to each subject's aiding time during condition 2. Performance data and subjective workload estimates were also collected.

RESULTS

The ANN accuracy when the subjects were performing the two levels of the task was greater than 70%. This level of accuracy is significantly above chance. The number of hits during target selection was significantly lower for the difficult level than for the easy level for all three task conditions (no aiding, aiding and random aiding, see figure 1). The number of DMPIs placed during the difficult task level was significantly lower than during the easy task level for the no aiding and random aiding conditions. There were significantly more missed weapons releases during the higher difficulty levels for the no aiding and random aiding conditions.

The implementation of adaptive aiding enhanced operator performance by improving target selection during the high difficulty conditions. There were no significant differences for hits when comparing the low difficulty results for the three conditions. However, during the difficult level, the number of hits was significantly higher for the aiding condition than for either the no aiding or the random aiding conditions. The aiding condition for the difficult level showed a mean of 5.2 of a possible 6.0 hits while the no aiding and random aiding hits were 3.8 and 3.9, respectively. The no aiding and random aiding difficult level hits were not statistically different.

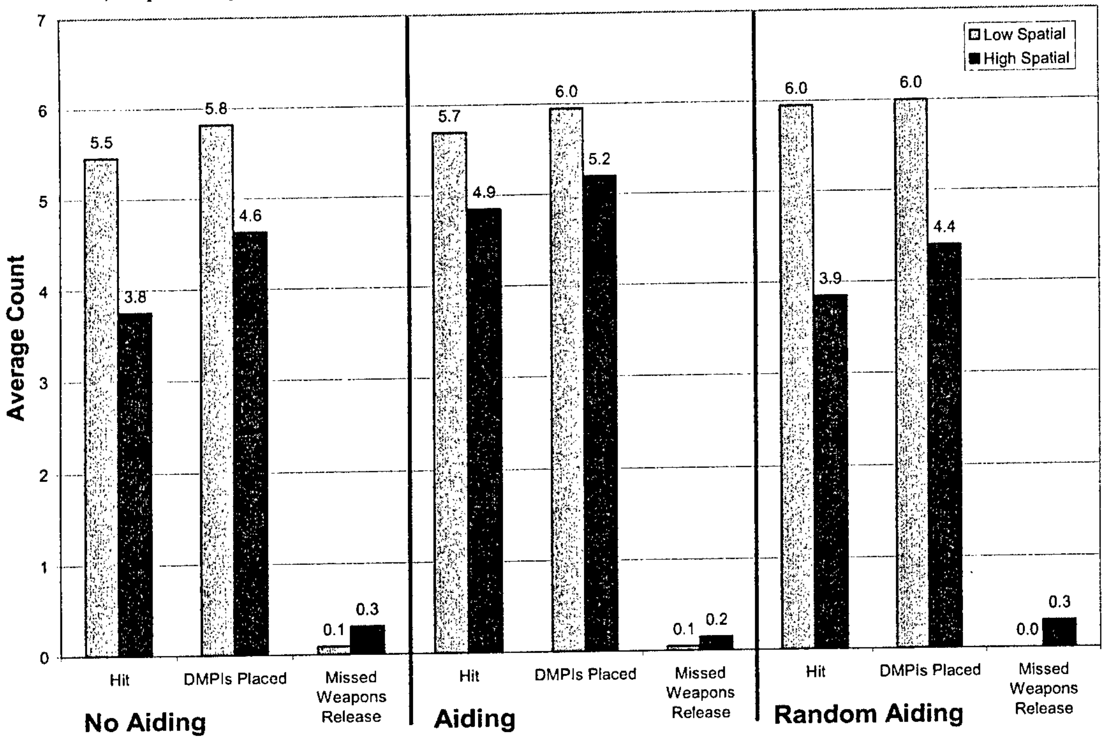


Figure 1. Mean performance data for the group for each difficulty level and condition.

DMPI placement during the easy level for all conditions was essentially the same regardless of whether or not aiding was present (5.9, 6.0 and 5.9 of a possible 6.0). However, during the difficult level there was significant improvement in the number of DMPIs placed during the aiding condition when compared to the other two conditions. The mean DMPI placements for the difficult task level during the aiding condition was 5.2 out of a possible 6, while the mean for the no aiding condition was 4.6 with 4.4 for the random aiding condition.

With regard to the overall mission success, fewer weapons release points were missed when adaptive aiding was implemented. The mean number of missed weapons release points was lower during the aiding condition, 0.2. For the non-aiding and random aiding conditions the missed weapons release point means were higher at 0.3. The difference was significant between the aiding and random aiding condition and marginally significant between the aiding and no aiding conditions. Because missing the weapon release point is such a disastrous event the improvement is highly significant operationally. Every missed weapons release point meant that the mission for the one vehicle was ineffective since it returned to base with all of its weapons. There was a 50% improvement in completing weapon assignments on time when adaptive aiding was implemented. In real world situations these improvements would be highly significant to the conduct of operations.

The subjective data showed that the easy task levels were rated as less demanding than the difficult levels regardless of the type of aiding used, figure 2. However, these differences were statistically significant for only random aiding. Comparisons among the low difficulty level results found only that the aiding condition was significantly lower than the no aiding condition. The differences in the subjective ratings among the high difficulty levels between the aiding and both the no aiding and the random aiding were marginally significant, $p<0.09$ and $p<0.06$, respectively.

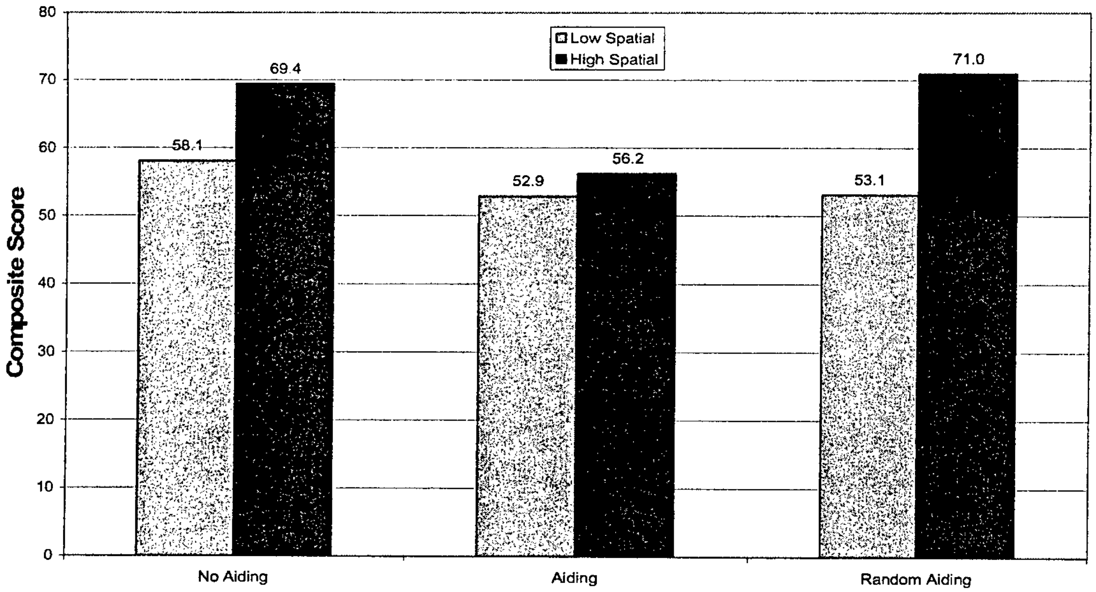


Figure 2. Subjective mean group ratings for each condition and task difficulty level.

DISCUSSION

On-line assessment of operator functional state permitted the recognition of suboptimal states. Further, the subsequent interventions improved operator performance by matching the task requirements to the momentary cognitive capabilities of the operators. These results demonstrate that psychophysiologically determined, real-time, operator state assessment coupled with adaptive aiding improves overall system performance in complex aviation tasks. During the difficult task levels all three performance measures improved with the presentation of adaptive aiding. The number of hits, DMPIs placed and weapon release points missed all improved which increased the success of the operational measures. Another indication of the positive effects of adaptive aiding was the lack of significance differences between the low and high task difficulty levels during the aiding condition. This was the case for both the performance and subject workload estimates. This may be due to the reduced difficulty of the task when the task demands were matched to the operator's functional state by the adaptive aiding.

The strong coupling between cognitive demands and psychophysiologically measures permits the rapid assessment of operator functional state that is necessary for on-line assessment and real-time adaptive aiding. The addition of performance measures and task variables should improve the accuracy and utility of on-line operator functional state assessment and the enhancement to complex task performance (Wilson & Russell, 1999).

These results suggest that adaptive aiding using ANNs with psychophysiological data will have application in actual operational environments. This task was a complex task which required visual search and decision making using specified rules of engagement which are much like actual operational settings. Further, recent advances in physiological sensors and signal processing will provide improved operator functional state assessors to be developed.

REFERENCES

Freeman, F. G., Mikulka, P. J., Prinzel, L. J. and Scerbo, M. W. (1999) Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. Biological Psychology, 50, 61-76.

Rouse, W. B. (1988). Adaptive aiding for human/computer control. Human Factors, 30, 431-443.

Wilson, G. F. & Russell, C. A. (in press). Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks, Human Factors.

Wilson, G. F. & Schlegel, R. E. (in press). Operator Functional State Assessment. NATO RTO Final Report, Neuilly sur Seine: RTO.

Wilson, G. F. and Russell, C. (1999). Operator functional state classification using neural networks with combined physiological and performance features. Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting, 1099-1102.

AUTOMATION RELIABILITY IN UNMANNED AERIAL VEHICLE FLIGHT CONTROL

Stephen R. Dixon & Christopher D. Wickens
University of Illinois at Urbana-Champaign

ABSTRACT

Twenty-four students flew a simulated unmanned aerial vehicle (UAV) through ten mission legs while searching for targets of opportunity and monitoring system parameters. Participants were assisted by automation which provided auditory alerts in response to system failures (SF). The auto-alerts were either 80% reliable or 60% reliable; the latter condition resulted in either a 3:1 ratio of false alarms to misses, or vice versa. Results indicated that the 80% reliable automation exceeded baseline (no automation) performance in the target search task. The two 60% reliable conditions provided no benefits to performance; both false alarms and misses hurt performance in the automated task *and* concurrent tasks, but did so qualitatively differently. Implications for this study suggest that automated aids must be fairly reliable to provide global benefits, and data regarding the relative costs of misses versus false alarms on performance were equivocal.

Keywords: unmanned aerial vehicle, automation, false alarm, miss

INTRODUCTION

Flying a single unmanned aerial vehicle (UAV) includes navigating the UAV, monitoring craft parameters, and searching for possible targets (Dixon & Wickens, 2003). The military currently employs different forms of automation to aid pilots in these tasks; however, very few automated aids are perfectly reliable, and can create different states of overtrust, undertrust, or calibrated trust (Parasuraman & Riley, 1997). It is unclear how unreliable the automation needs to be to cause performance to drop below that of baseline (no automation), and while a 70% "threshold" has been offered (Dixon & Wickens, 2003; Lee & See, in press), there are noted exceptions both above and below that level (e.g. Dzindolet et al., 1999; Rovira, Zinni, & Parasuraman, 2002). Dixon & Wickens (2003) found benefits for an auto-pilot with 67% reliability, but costs for an auto-alerting system at the same reliability level, and reasoned that under conditions of high workload, an operator may rely upon imperfect automation even if the automation is not fully trusted. Such reliance will degrade performance of the automated task itself even as it helps concurrent tasks (e.g. Rovira et al., 2002).

Within the class of automation that guides attention to notice or diagnose a failure (Parasuraman et al, 2000), unreliable aids will create false alarms (alarm with no event) and/or misses (no alarm with an event). False alarms tend to cause distrust in the aid (Meyer & Ballas, 1997), while misses lead to reallocation of visual resources to the raw data in order to "catch" the automation miss (Cotté, Meyer & Coughlin, 2001). Using target recognition automation, Maltz & Shinar (2003) found that increasing false alarm rates caused greater disruption to performance than did increasing miss rates. Dixon & Wickens (2003) also made such a contrast by having pilots perform a high-fidelity UAV simulation under conditions with either no automation, perfectly reliable auto-alerts, or 67% reliable auto-alerts with either false alarms or misses. Results revealed that while the perfectly reliable auto-alerts benefited the automated task, the two imperfect auto-alert conditions equally hurt performance in both the automated task and concurrent tasks.

While Dixon & Wickens (2003) used conditions with *only* false alarms or *only* misses, the current study included an 80% reliable condition with an equal number of false alarms and misses, as well as two 60% reliable conditions with a 3:1 ratio of false alarms to misses and vice versa. We hypothesized that (1) 80% reliability would consistently improve performance above baseline; (2) both 60% reliability conditions would degrade performance below baseline; (3) decrements due to unreliability would be more pronounced on the automated task than on concurrent tasks; and (4) miss-prone automation would disrupt concurrent tasks more than false-alarm prone automation, because of the former's requirement for more continuous visual monitoring of SF status. Please refer to Dixon & Wickens (2004) for a more thorough presentation of the experimental methods

METHOD

Participants and Equipment. Thirty-two students at the University of Illinois received \$8 per hour, plus bonuses of \$20, \$10, and \$5, for 1st, 2nd, and 3rd place finishes, respectively, in their group of eight pilots. Figure 1 presents a sample display for a UAV simulation, with verbal explanations for each display window and task.

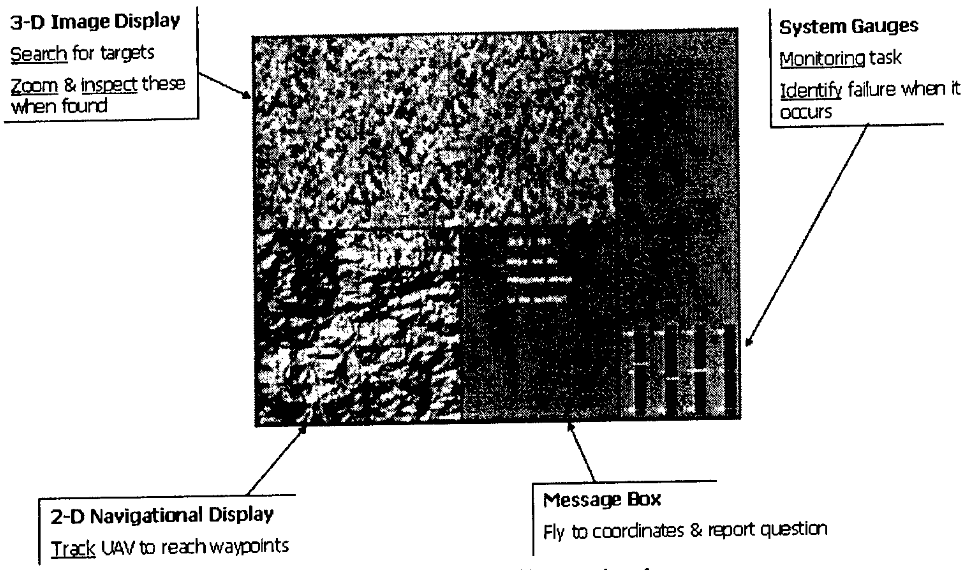


Figure 1. A UAV display with explanations for different visual areas.

Procedure. Each pilot flew one UAV through ten different mission legs, in one of the four experimental conditions, while searching for targets of opportunity and monitoring system parameters. Pilots obtained flight instructions via the Message Box, including fly-to coordinates and a report question pertaining to the command target (CT). These instructions were present for 15 seconds, and pressing a repeat key automatically refreshed the flight instructions for an additional 15 seconds.

CT reports required that pilots loiter around the target, manipulate a camera for closer target inspection, and report back relevant information to mission command. Along each mission leg, pilots were also responsible for detecting and reporting targets of opportunity (TOO), a task similar to the CT report, except that the TOOs were much smaller (1-2 degrees of visual angle) and camouflaged. TOOs could occur during simple tracking (low workload) or during a pilot response to a system failure (high workload).

Concurrently, pilots were also required to monitor system gauges for possible system failures (SF), which were indicated by the white needle moving into a red zone (at the top or bottom of the gauges). SFs were designed to fail either during simple tracking (i.e. low workload) or during TOO and CT inspection (i.e. high workload). The SFs lasted only 30 seconds, after which the screen flashed bright red and a salient auditory alarm announced that the pilot had failed to detect the SF.

Automation aids, in the form of auditory auto-alerts during SFs, were provided for three out of the four conditions. The A80 condition (A = automation; 80% reliable) failed by giving one false alarm (i.e. alarm with no actual SF), and one miss (i.e. a SF with no alarm) during each mission. The A60f condition (f = false alarm; 60% reliable) resulted in more false alarms (3) than misses (1), while the A60m condition (m = miss; 60% reliable) resulted in more misses (3) than false alarms (1). Pilots were told that the automation was either "fairly reliable" or "not very reliable", as well as the bias setting (i.e. more false alarms or more misses). Ratings of subjective trust were given by each pilot at the end of the mission.

RESULTS

3.1 Mission Completion. Tracking error was not affected by condition [$F(3, 27) = 1.24$, $p > .10$]. The number of repeats was affected by condition [$F(3, 25) = 3.56$, $p = .029$]; however, only the A60m condition (mean = 8.5) suffered relative to baseline (mean = 3) condition [$p < .01$].

3.2 Targets of Opportunity (TOO) and Command Targets (CT). For TOO detection rates, only the A80 condition (mean = 93%) improved performance relative to baseline (mean = 76%) [$p < .05$]. For TOO detection times, as shown in Figure 2, an interaction between condition and load [$F(3, 23) = 4.82$, $p = .01$] indicates that the condition effect was only present at high load.

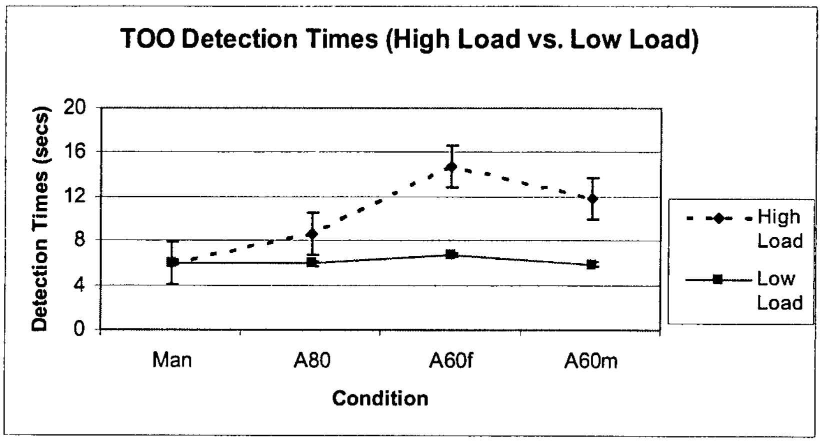


Figure 2. TOO detection times across condition and workload. SE bars are included.

Figure 2 reveals that the penalty for increased load was higher for both the A60f (mean = 14.73) and the A60m (mean = 11.87) conditions relative to baseline (mean = 6.04) [all $p < .05$]. Only the A60f condition differed from the A80 condition (mean = 8.58) [$p < .01$]. For CT detection times, there was a main effect of condition [$F(3, 27) = 6.16$, $p < .01$], and both the A60f (mean = 4.17) and the A60m (mean = 4.11) conditions suffered relative to baseline (mean = 2.45) [all $p < .05$].

3.3 System Failures (SF). For SF detection rates, higher load reduced detection rates [$F(1, 27) = 21.46$]; however, there was no main effect of condition [$F(3, 27) < 1.0$], or interaction [$F(3, 27) < 1.0$]. For SF detection times, as shown in Figure 3, higher load increased detection times [$F(1, 27) = 93.3$, $p < .001$]. The main effect of condition [$F(3, 27) = 3.62$, $p = .026$] can only be interpreted in the context of the interaction [$F(3, 27) = 3.06$, $p = .045$], which reveals that the A60f condition (mean = 19.99) suffered more due to high load than the other conditions.

Figure 3 reveals that the penalty due to high load was approximately 6-9 seconds more for the A60f condition than the other three conditions [all $p < .03$]. We note that each of the 60% condition means is actually composed of two different components: responses when an alert correctly sounded, and those when the alert failed to sound. Table 1 shows the resulting four means, within the high workload condition.

The data reveal the clear slowing for RT when the alarm "missed" the SF event, indicating that in both conditions, pilots had relied heavily upon the automation, and their detection suffered when it failed. Correct alerts were responded to more rapidly with the miss prone automation (mean = 3.96) than the false alarm-prone automation (mean = 13.93) [$p < .05$], reflecting the pilots' immediate *compliance* with the auditory alert (Meyer, 2001) in the former condition, in contrast to the false-alarm prone condition, where pilots were less likely to interrupt target inspection to deal with the alarms. We also infer that greater compliance in the miss condition is coupled with an ongoing greater awareness of the SF gauges, fostered by a reduced *reliance* on that automation, and causing greater disruption to memory recall.

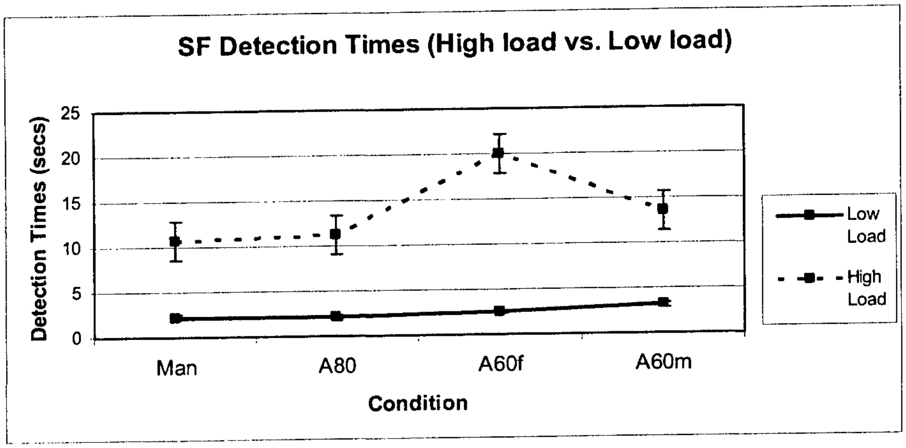


Figure 3. SF detection times across condition and workload. SE bars are included.

Table 1. Component means in the A60f and A60m conditions. SE is in parentheses.

		CONDITION	
		A60f	**A60m**
EVENT	**Miss (failure)**	26.05 sec *(1.83)*	23.29 sec *(2.77)*
	Alarm (correct)	13.93 sec *(4.85)*	3.96 sec *(1.17)*

3.4 Subjective ratings of trust. Pilots were surprisingly accurate in their overall assessment of the automation reliability [A80 = 82%; A60f = 54%; A60m = 56%], in contrast to Dixon & Wickens (2003), who concluded that pilot trust in the automation was poorly calibrated when they did not receive any prior information as to reliability levels or bias setting.

DISCUSSION

The A80 condition (80% reliability) supported a significant increase in concurrent task performance, confirming our first hypothesis. This indicates that the automation, while imperfect, still allowed pilots to save visual and cognitive resources, which they could reallocate to the concurrent target search task (Rovira et al, 2002).

At 60% reliability, neither the false alarm nor miss conditions (A60f and A60m) provided any benefits, and in some instances performance was well below baseline during high workload conditions, thereby confirming hypothesis 2. In general, however, the costs of imperfection were as heavily born on the concurrent tasks as on the SF task itself, a pattern inconsistent with hypothesis 3.

Finally, regarding hypothesis 4, the false alarm condition (on average, across performance measures) resulted in slightly poorer performance in the SF detection task, than did the miss condition. On the one hand, the miss condition degraded CT memory (requiring more repeats) to a greater extent than did the false alarm condition, supporting hypothesis 4. That is, more continuous monitoring of the raw system data was required in the miss condition. On the other hand, the false-alarm condition (in high workload) appeared to delay detection of a TOO that became visible while the failure was present, more than the miss condition. This difference we attribute to pilots' need, when an alarm sounds in the A60F condition, to double check the raw data (visual system gauges) to assess its consistency with the auditory alert (a distrust, or reduced compliance). Thus the two types of automation imperfection had opposing effects on the concurrent tasks, both replicating prior findings of Dixon & Wickens (2003).

With regard to SF performance itself, figure 3 and table 2 clearly indicate reduced costs for the miss condition than for the false alarm condition at high workload, a pattern at odds with that reported by Dixon &

Wickens (2003). We can account for the current pattern in terms of the greater compliance with, and lesser reliance on, the imperfect automation in the miss than in the false alarm condition (Meyer, 2001). Compliance is increased because of the belief that if an alarm sounds, it is quite likely to be true. Reliance on the alert is reduced because of the subjects' knowledge that it may frequently fail to signal a true system failure. The reason for the discrepancy of the current pattern of results with those of Dixon and Wickens requires further research.

The implications of this study are that higher reliability automation in necessary to facilitate improvements in overall performance relative to baseline, and that false alarms may be more detrimental to overall alerted task performance than misses.

ACKNOWLEDGMENTS

This research was sponsored by a subcontract # ARMY MAAD 6021.000-01 from Microanalysis and Design, as part of the Army Human Engineering Laboratory Robotics CTA, contracted to General Dynamics. David Dahn and Marc Gacy were the scientific/technical monitors Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Army CTA. The authors also wish to acknowledge the support of Ron Carbonari and Jonathan Sivier (in developing the UAV simulation), and of Dervon Chang for assisting with data collection.

REFERENCES

Cotté, N., Meyer, J., & Coughlin, J. F. (2001). Older and younger driver's reliance on collision warning systems. Proceedings of the 45th Annual Meeting of the Human Factor Society (pp. 277-280). Santa Monica, CA: Human Factors and Ergonomics Society.

Dixon, S. & Wickens, C.D. (2003). Imperfect Automation in Unmanned Aerial Vehicle Flight Control. (AHFD-03-17/MAAD-03-1). Savoy, IL: University of Illinois, Aviation Research Lab.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (1999). Misuse and disuse of automated aids. Proceedings of the 43rd Annual Meeting of the Human Factors and Ergonomics Society (pp. 339-343). Santa Monica, CA: Human Factors and Ergonomics Society.

Lee, J. D., & See, K. A. (in press). Trust in automation: Designing for appropriate reliance. Human Factors.

Maltz, M., & Shinar, D. (2003). New alternative methods in analyzing human behavior in cued target acquisition. Human Factors, 45, 281-295.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. Human Factors, 43, 563-572.

Meyer, J., & Ballas, E. (1997). A two-detector signal detection analysis of learning to use alarms. Proceedings of the 41st Annual Meeting of the Human Factor Society (pp. 186-189). Santa Monica, CA: Human Factors and Ergonomics Society.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39, 230-253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, & Cybernetics, 30(3), 286-297.

Rovira, E., Zinni, M., & Parasuraman, R. (2002). Effects of information and decision automation on multi-task performance. In Proceedings of the 26th Annual Meeting of the Human Factors and Ergonomics Society. (pp. 327-331). Santa Monica, CA: Human Factors and Ergonomics Society.

A "PLAYBOOK" FOR VARIABLE AUTONOMY CONTROL OF MULTIPLE, HETEROGENEOUS UNMANNED AIR VEHICLES

Christopher A. Miller, Harry B. Funk, Robert P. Goldman and Peggy Wu
Smart Information Flow Technologies

ABSTRACT

Human interaction with complex and highly capable automation, such as robots and Unmanned Air Vehicles (UAVs), will profit from more flexible forms of user control. A particularly powerful method of interacting with and "controlling" other autonomous agents—specifically, other humans—is delegation. We describe the attributes of delegation relationships and then propose an approach for implementing a similar relationship between humans and multiple, heterogeneous UAVs. We call our approach a "playbook" since it permits humans to delegate tasks to automation via very rapid, pre-compiled commands which leave substantial interpretation open to the automation, or to "drill down" and refine the delegated task if time permits and/or need requires. While the playbook approach promises particular relevance for highly flexible control of UAVs, the fact that it places the operator in charge of determining how much and what kind of automation to use when makes it applicable to a wide range of complex automation types.

Keywords: Delegation, automation, playbook, unmanned air vehicles, robots, human-automation interaction, human-robotic interaction, variable initiative control, adaptive autonomy

INTRODUCTION

As robots become more prevalent, there is an increasing need for models and methods of human interaction with them that provide the kinds of control that users need and desire without undue additional workload. This need is currently most pressing in the control of unmanned robotic vehicles for military purposes, particularly Unmanned Air Vehicles (UAVs) which lead the way in their near-term availability and complexity of operations.

As UAVs become more common in operations, several core challenges are emerging. First, current operations employ multiple operators to control and operate each vehicle. This approach is rapidly becoming unacceptable and a host of research and development efforts are underway to reduce or even reverse those ratios—enabling a single individual to control multiple UAVs. Second, current practices of providing a dedicated, special purpose workstation for each vehicle or vehicle type will also be unacceptable in applications where individuals must interact with multiple, heterogeneous vehicles. Third, some types of UAVs and their concepts of employment will imply radical differences in the way in which they are operated. For example, as UAVs become available to lower echelons, new usability and training requirements are imposed. Vehicles such as the U.S. Marines DragonEye, DARPA's Organic Air Vehicle, etc. are small, human-portable and -launchable UAVs for small field units during their operations. With such vehicles, it is unreasonable to demand that all operators spend months training to operate a particular vehicle class, nor can they devote full attention to vehicle management. Instead, UAVs must be controllable with much less training and while the user is engaged in many other activities—perhaps even while taking fire.

In recent work, we have advocated a "delegation" approach to human interaction with intelligent automation (Miller, 2003; Miller and Parasuraman, 2003). Delegation is clearly a form of supervisory control (Sheridan, 1984), but highly flexible, adaptive delegation approaching the power of effective human-human delegatory relationships extends Sheridan's concept and requires an explicit vocabulary with which to communicate about goals, plans, constraints, stipulations and priorities/values. We have been developing a "playbook" that provides such a communication mechanism and are now implementing it in a tool for the control of multiple, heterogeneous UAVs. Our most challenging current application for this "Playbook" is a control interface for small unit operators of multiple, heterogeneous small UAVs during urban operations. In this paper, we will present the rationale for a delegation approach to controlling UAVs and, indeed, many forms of robots and automation, and will present our initial design concepts for a Playbook Interface to the control of heterogeneous, UAVs.

Delegation Interfaces for High Level Control

Humans have been striving to retain control and produce efficient outcomes via the behavior of other autonomous agents, within the limitations of their own cognitive and attentional resources, for millennia. It so happens that those "agents" have been other humans. Not surprisingly, we have developed many useful methods for accomplishing these goals, each customized to a different domain or context of use. When we have some degree of managerial authority over another human actor and yet will not be directly commanding performance of every aspect of a task, we call the relationship (and the method of commanding performance) delegation. Delegation allows the supervisor to set the agenda either broadly or specifically, but leaves some authority to the subordinate to decide exactly how to achieve the commands supplied by the supervisor. Thus, a delegation relationship between supervisor and subordinate has many requirements:

1. The supervisor retains overall responsibility for the work undertaken by the supervisor/subordinate team and retains commensurate authority.
2. The supervisor can interact flexibly and at multiple levels. When and if the supervisor wishes to provide detailed instructions, s/he can; when s/he wishes to provide only loose guidelines and leave detailed decisions to the subordinate, s/he can do that also—within the capability limits of the subordinate.
3. To provide useful assistance, the subordinate must have substantial knowledge about and capabilities within the work domain. The greater these are, the greater the potential for the supervisor to offload tasks (including higher level decision making tasks) on the subordinate. Among other things, the subordinate must be able to make reasonable decisions and tradeoffs among the various courses of action within the space of authority (see 6 below) delegated to him/her. It may be helpful if the subordinate can interact with the supervisor before taking action in order to improve course of action selection, and can explain after the fact why a given course was chosen. Item 5 below will facilitate these interactions.
4. The supervisor must be aware of the subordinate's capabilities and limitations and must either not task the subordinate beyond his/her abilities or must provide more explicit instructions and oversight when there is doubt about those abilities.
5. There must be a "language" or representation available for the supervisor to task and instruct the subordinate. This language must (a) be easy to use, (b) be adaptable to a variety of time and situational contexts, (c) afford discussing tasks, goals and constraints (as well as world and equipment states) directly (as first order objects), (d) minimize undesired ambiguity and (e) most importantly, be shared by both the supervisor and the subordinate(s).
6. The act of delegation will itself define a window or space of control authority within which the subordinate may act. This authority need not be complete (e.g., checking in with the supervisor before proceeding with specific actions or using some resources may be required), but the greater the authority, the greater the workload reduction on the supervisor.

Items 4 and 6 together imply that the space of control authority delegated to automation is flexible: the supervisor can choose to delegate more or less "space," and more or less authority within that space (that is, range of control options), to automation. Item 5 implies that the language available for delegation must make the task of delegating feasible and robust—enabling, for example, the provision of detailed instructions on how the supervisor wants a task to be performed or a simple statement of the desired outcome.

In essence, delegation is the process of instructing an intelligent subordinate in what the supervisor wants to occur and how (within what constraints)—expressing intent. As such, it is implicit in Sheridan's (1984) notion that, as a part of supervisory control, the supervisor would have to *instruct* automation in the ways it should behave. At the time of his writing, however, Sheridan seems to have envisioned primarily instructing simple assembly line automata in how to execute their movements. Today's automation permits much more complex behaviors, but also provides much more intelligence about how to organize and plan those behaviors—thereby making more complex (and more abstract or "higher level") delegation interactions feasible, as we will describe below.

A Playbook Approach to Delegation

Delegation approaches can be configured along the various methods of expressing intent (cf. Shattuck, 1995; Klein, 1998) Miller (2003) describes five components of delegation that can be composed and reused in different combinations for different styles of delegation appropriate to different contexts and domains:

1. Stipulation of a goal to be achieved—a desired (partial) state of the world.
2. Stipulation of a plan to be performed—where a plan is a series of actions, perhaps with sequential or world state dependencies.
3. Provide constraints in the form of actions or states to be avoided.
4. Provide "stipulations" in the form of actions or states (i.e., sub-goals) to be achieved.
5. Provide an objective function or other guidelines that enables the subordinate to make informed decisions about the desirability of various states and actions

We have emphasized an the first four styles of interaction in the Playbook architecture for small unit control of heterogeneous UAVs described above.

Figure 1 presents the basic architecture for our Playbook. Playbook consists of a user interface (UI) and an automated analysis and planning component that communicate via a shared model of the tasks that can be performed in a domain. This task model is both hierarchically and sequentially organized allowing the stringing together of tasks or "plays" in commandable sequences and/or drilling down within a given play to select alternate performance methods. These components form the Playbook itself, but Playbook must communicate with a control environment (e.g., UAV controllers) if it is to accomplish behaviors in the real world.

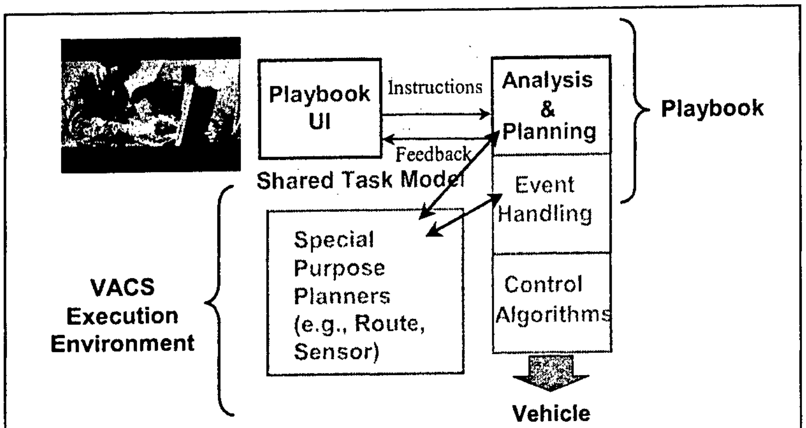


Figure 1. General Playbook Architecture.

Operators can interact with automation in highly sophisticated and flexible ways via Playbook. Like the quarterback of a football team, a PVACS operator can command a very complex "play"—even one involving a heterogeneous mix of actors (vehicles)—by accessing a simple label and trusting the individual actors to enact that play appropriately in the current context. Also like the quarterback, the operator can issue more specific constraints on or stipulations about finer-grained behaviors of individuals, and can even (in principle, but not yet in implemented practice) compose entirely novel plays, albeit spending more time in the process.

In each case, this flexibility is enabled by the interaction between a UI based on recognizable tasks and a smart planning component that understands those tasks. The Playbook planning component evaluates the feasibility of alternate methods of performing commanded plays. When given a high-level play, the planning component selects among various applicable methods, issues instructions to the execution environment and monitors for necessary revisions during performance. When given lower-level, more detailed commands, the planning component reviews them for feasibility and either (a) reports when commanded actions are infeasible, (b) passes 'validated' commands to the execution environment and monitors their performance, or (c) fleshes out operator commands to an executable level within the constraints imposed.

Playbook Control of Heterogeneous UAVs

We are developing a prototype Playbook for variable initiative, play-like control of multiple heterogeneous UAVs by a single individual in urban combat operations. Our initial demonstration scenario involves a platoon commander who must coordinate multiple UAVs for sustained surveillance of a fixed location (e.g., an intersection) while simultaneously securing nearby buildings. Since the commander cannot devote sustained attention to managing the UAVs, they must operate largely autonomously. Furthermore, the commander might have little time to convey his/her intentions. S/he can task the UAV team through the Playbook by "calling" a single, simple *Overwatch* play and providing a single parameter (the target area). The conceptual structure of *Overwatch*, as understood by both the Playbook system and the human operator is illustrated in Figure 2.

Note that this description is decomposed both functionally (with only the darkened nodes being expanded in progressively deeper layers in this illustration) and sequentially. This representation of *Overwatch* permits a wide variety of specific implementations. Not only could a variety of vehicles with different sensors and flight capabilities be used to satisfy the scanning portions of the plan, but the specific routes to be flown, whether or not to launch a vehicle, etc. are all alternatives available under this general, functional task description. In practice, in order to be executable, a specific selection must be made among each set of these alternatives (though each selection may alter the set of later alternatives available) at some point before that portion of the task/plan can be executed. In keeping with Vicente's (1999) recommendation to allow the user, in context, to "finish the design" Playbook allows those decisions to be made by a mixture of the human user and the UAV(s) themselves at various points up to actual execution. The strength of the Playbook approach, however, comes from the fact that the human user can stipulate as much or as little of those decisions as s/he wants and has time to do.

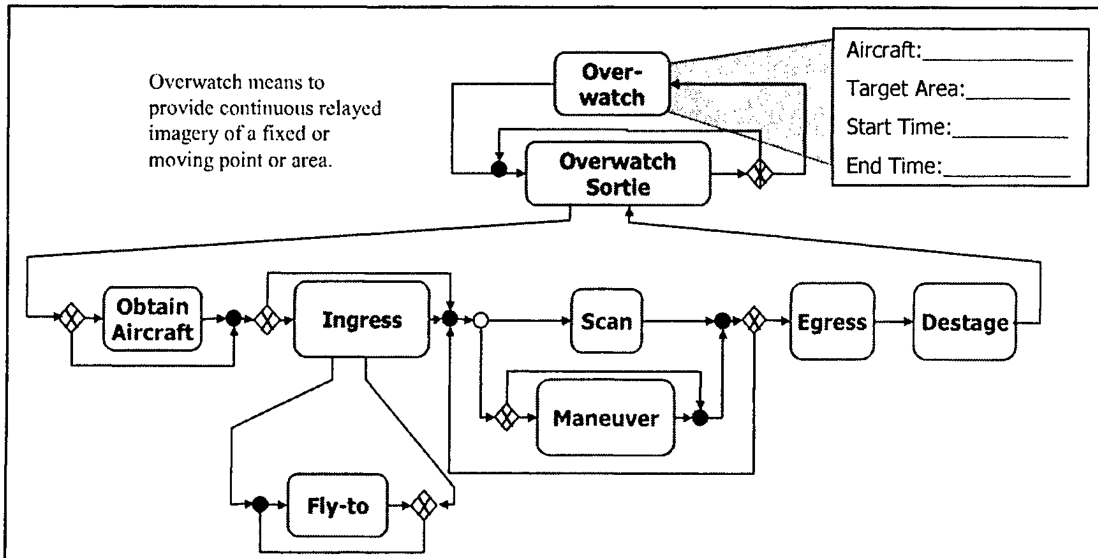


Figure 1. Shared Task Description of Overwatch.

Our Playbook is currently capable of commanding specific Overwatch sorties—the second level in Figure 2 above. When an *Overwatch* sortie for a fixed target is commanded, Playbook's planning component expands the definition of that play and seeks available vehicles for it to request or command. Playbook will automatically decide, for example, that if *Overwatch* is commanded at night, UAVs without night vision cameras are unacceptable. If no satisfactory vehicle is available, Playbook will report this fact to the user and allow him/her to revise the play. Otherwise, Playbook (in this high tempo operational environment) will begin execution of the plan within the operator's constraints.

We are extending this Playbook so that users will be able to simply specify that they would like *Overwatch* (i.e., the top level task in Figure 2) performed over a particular area. Instead of needing to call for a specific *Overwatch* sortie (for a particular vehicle, with specified start and end times), they will be able to specify a duration that a particular area must be watched. If no single vehicle can provide the desired degree of sustained surveillance (either because of area or flight time constraints), Playbook's planning component will meet the task specification with a set of Overwatch sorties. In the process, Playbook will automatically coordinate the behaviors of multiple, heterogeneous UAVs to satisfy the demands of the play as called by the operator.

Our demonstration will illustrate Playbook's ability to coordinate multiple, heterogeneous vehicles in a high fidelity simulation by showing a situation in which available vehicles have different capabilities, different arrival times and different loiter capabilities. Using Playbook's control capabilities, we will coordinate the behaviors of at least three different UAVs (notionally, a Dakota fixed-wing aircraft, a GT-MAX helicopter and an ducted-fan Organic Air Vehicle) to provide sustained surveillance, all from a single, initial 15-second operator command sequence. We will also illustrate Playbook's capability to provide feedback on mission performance. Future enhancements will illustrate the ability for the operator to shape the mission via more complex interactions with the Playbook.

ACKNOWLEDGEMENTS

This program is funded by a Phase II Small Business Innovation Research grant from DARPA IXO, administered by the U.S. Army Aviation & Missile Command, contract number DAAH01-03-C-R177. We would like to thank Dr. John Bay for his oversight and advice during the investigation. Geneva Aerospace personnel, especially Billy Pate, deserve thanks for their work supporting the execution environment for the Playbook demonstration described above.

REFERENCES

Klein, G. (1998). *Sources of Power*: How People Make Decisions. Cambridge, MA; MIT Press.

Miller, C. (2003). Delegation Architectures; Playbooks and Policy for Keeping Operators in Charge. In Proceedings of the NATO Research and Technology Organization Special Workshop on Uninhabited Military Vehicles, June 10-13; Leiden, Holland.

Miller, C. and Parasuraman, R. (2003). Beyond Levels of Automation: An Architecture for More Flexible Human-Automation Collaboration. In Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society, October 13-17; Denver, CO.

Sheridan, T. (1987). "Supervisory Control," in Handbook of Human Factors. G. Salvendy, Ed. New York: John Wiley & Sons, pp. 1244-1268.

Shattuck, L. (1995). Communication of Intent in Distributed Supervisory Control Systems. Unpublished dissertation. The Ohio State University, Columbus, OH.

Vicente, K. (1999). Cognitive Work Analysis.

INTELLIGENT AGENT SAVES LIVES

Dick Steinberg
Schafer Corporation

ABSTRACT

Designing human computer interfaces for rapid command and control decision making displays has unique challenges. Displays for Air Traffic Control, Military Operations, and Emergency Management require an interface which optimizes performance while minimizing errors. In addition to performing the emergency functions, maintaining operational readiness of equipment is a task where errors and time delays could cause loss of human life. Interaction efficiency is critical to avoid operator fatigue and minimize operator error rates that could cost lives. In emerging emergency management systems and concepts, the user typically acts as a manager by exception while the majority of system activity is computer automated. While direct interaction of the user with the system is minimal, an inaccurate action by the user can have catastrophic consequences. There is clearly a need for decision aiding systems to help focus users on the most important information. **This paper describes a case study which demonstrates the use of human engineered intelligent agent.** The results of this effort suggest as much as a 40% reduction in casualty risks while using the intelligent agent and strong positive preference ratings by operators by all operators who used it.

KEYWORDS: Human Engineering, Intelligent Agent, Fuzzy Logic

INTRODUCTION

An intelligent agent is a process than acts as an assistant to the user by displaying suggested actions and information which may need to be evaluated. The most well known agent is the Microsoft Office Assistant. However, even the best software algorithm is useless if the user interface is poor. While intelligent agents are not new (Coury and Semmel, 1996), this case study emphasized the use of human engineering methods in the design and implementation of the agent. It is important to note, that the method of displaying information, directly affects the action taken by an operator (Wickens, 1987). The actions taken by an operator may be erroneous if the decision aiding display method is not compatible with the cognitive functions of the operator or intrusive to the operator's task. Without the appropriate incorporation of Human Factors Engineering integrated into a decision aiding display, the most sophisticated intelligent algorithm is made useless.

Additionally, many of the decisions that operators of rapid command and control emergency systems are required to make are based on uncertainty in measured data and predicted future events. The problem solutions are based on a complex set of relationships of factors. There is clearly a strong need for decision aiding systems coupled with advanced Human Computer Interaction (HCI) methods to minimize risk of erroneous actions by operators/commands. This effort demonstrated the use of human engineered intelligent tools to satisfy this need using the Military Command and Control domain. The intelligent tool adapted the display presentation to the anticipated needs of the user based on existing external factors and habits of the user or group of users.

THESIS

Successful development of intelligent tools built with Human Factors Design methods for emergency command and control display, will enhance the ability of a operator to successfully execute a mission minimizing casualties.

DISCUSSION OF FUZZY INFERENCE ENGINE

Fuzzy Inference rule sets have been used effectively to build intelligent tools when the conditions of decision-making are not absolute thresholds. Fuzzy rule base tools have demonstrated value when information is intrinsically imprecise or uncertain. Fuzzy tools help to smooth data transitions when uncertainties in data are high. Consider this example of a fuzzy rule set for making a management decision on a target in a military command and control application. Using subject matter expert interviews, the following statement was generated.

If there is a Weapons Poor Resource Situation and there is a high ratio of weapons allocated in relation to the lethality of the target and the target is threatening relatively fewer people, then this is a candidate to override system processing to lower the number of weapons allocated to negate it.

This statement is difficult to model using discrete mathematics. For example, how many weapons expended will it take to make the situation Interceptor Poor? How many casualties at risk does it take to make a target threatening to relatively fewer people? This logic lends itself to fuzzy rule sets. A Fuzzy Inference Engine was constructed using criteria mentioned in the example. Following the generation of the rule set, membership function curves were constructed based on cognitive task analysis with users.

HUMAN ENGINEERING DESIGN

The design approach used was an Object Oriented Task Analysis (OOTA) integrated with cognitive task analysis and screen walkthroughs. The importance of using cognitive task analysis is emphasized by the fact that an action taken by an operator may be erroneous if the decision aiding display method is not compatible with the cognitive functions of the operator (Hammond, 1988). The displays were tested using a "Think Aloud" Protocol (Armstrong, Brewer, Steinberg, 2001) and revised based upon usability tests and cognitive walkthroughs.

DATA GATHERING

Due to the high degree of expertise required for the task, manpower and training schedules, only three subjects were available to gather the objective measurements. Each operator was asked to think aloud as they worked, acting as a play-by-play-announcer describing the details of a sporting event. The operator's behaviors were observed to ascertain confusing areas of the interface. Most of the testing was recorded on video tape with five hours of tape collected. Most of the operator actions were logged and assessed to see if any patterns or frequencies of actions could be used to re-design the displays. Following each scenario, the expected casualties were determined. The following graphic shows the results of the testing for the three subjects.

RESULTS

The first subject made decisions which decreased the likelihood of casualties 40%. The second subject made decisions which decreased the likelihood of casualties 25%. The third subject, who by their own assertion, felt they could out guess the system, performed only %5 better using the intelligent agent. (Figure 1)

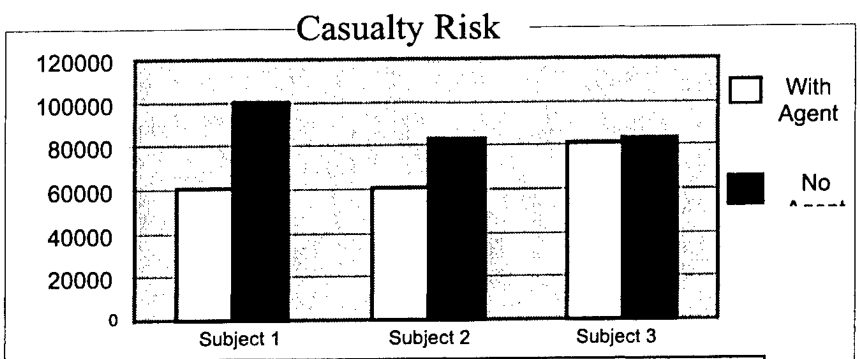


Figure 1. Intelligent Agent Reduced Casualty Risks

SUBJECTIVE DATA

Seven subjects participated in the subjective data gathering. All seven subjects gave high accolades for the intelligent agent.

Subject 1. SGF. "The Agent is Fantastic."

Subject 2. MAJ. "It helps you focus on information"

Subject 3. SGF. "It helps you see information you might otherwise overlook. It should also alert you to information you may have filtered out.
Subject 4. 1LT. "It helps you focus on the right decision"
Subject 5. MAJ. "The agent would be very useful"
Subject 6. COL "It is a much better tool than is available now"
Subject 7 CW3. "The concept is great"

CONCLUSION

While there were too few subjects to make a statistical conclusion, the human engineered intelligent agent supports the thesis showing tremendous promise for diminishing casualty risks for command and control display operators.

ACKNOWLEDGMENTS

This research was performed under a Small Business Innovative Research Contract to the Army Space and Missile Defense Command.

REFERENCES

Armstrong, S., Brewer W., and Steinberg R., (2001) Usability Testing. In O' Brien and Charlton, Handbook of Human Factors Engineering Methods, (Erlbaum and Associates).

Coury, B.G., and Semmel, R.D. (1996). New Directions in the design of intelligent user interfaces for supervisory control. In M. Mouloua and R. Parasuraman, Automation and Human Performance: Theory and Applications (Hillsdale, NJ. Lawrence Erlbaum)

Hammond, K.R. (1988), Judgment and decision making in dynamic tasks. Information and Decision Technologies

Wickens, C.D. (1987). Information processing, decision-making, and cognition. G. Salvendy (Ed.), Handbook of human factors. New York. Wiley

EXPLORING AUTOMATION ISSUES IN SUPERVISORY CONTROL OF MULTIPLE UAVS

Heath A. Ruff
Sytronics, Inc., Dayton, OH

Gloria L. Calhoun, Mark H. Draper, John V. Fontejon
Air Force Research Laboratory

Brian J. Guilfoos
Sytronics, Inc., Dayton, OH

ABSTRACT

An evaluation was conducted on a generic UAV operator interface simulation testbed to explore the effects of levels-of-automation (LOAs) and automation reliability on the number of simulated UAVs that could be supervised by a single operator. LOAs included Management-by-Consent (operator consent required) and Management-by-Exception (action automatic unless operator declines). Results indicated that the tasks were manageable, but performance decreased with increased number of UAVs supervised and reduced automation reliability. Performance with the two LOAs varied little and did not show a consistent trend across measures. Analyses indicated that participants typically did not utilize the automation. A follow-on study was conducted that employed shorter LOA time limits. Results showed participants' workload and confidence ratings were less favorable for the shorter limits and they still exercised the automation rarely, although more frequently. Further research is needed to explore the complex relationship between LOAs, time limits, perception of workload, vigilance effects, and confidence.

Keywords: Level of Automation; Supervisory Control; UAV; Reliability; Multi-aircraft control

INTRODUCTION

The majority of present day Unmanned Air Vehicle (UAV) systems require multiple operators to control a single UAV. Reducing the operator-to-vehicle ratio would reduce life-cycle costs and serve as a force multiplier. Thus, automation technology is under rapid development. The envisioned system involves *multiple* semi-autonomous UAVs being controlled by a *single* supervisor. These UAVs will have the capability to make certain higher-order decisions independent of operator input and predefined mission plans. This capability of the UAV 'to decide' constitutes an entirely new tasking on the operator to rapidly judge the appropriateness of decisions/actions made by the automation and assess their impact on overall mission objectives, priorities, etc. The number of systems to monitor will increase and it will be more of a challenge for operators to maintain situation awareness (SA) through long periods of nominal operations, interjected with short periods of time-sensitive contingency operation.

Unfortunately, it has been documented in studies of manned systems that increasing the use of automation can cause rapid and significant fluctuations in operator workload and can result in loss of operator SA and performance. In fact, there are numerous issues associated with automation management such as task allocation between operator and system, human vigilance decrements, clumsy automation, limited system flexibility, mode awareness, trust/acceptance, failure detection, automation biases, etc. (Parasuraman, Sheridan, & Wickens, 2000). Innovative methods are required to keep the operator 'in the loop' for optimal SA, workload, and decision making. One method that may enhance supervisory control is multiple levels-of-automation (LOAs), whereby each level specifies the degree to which a task is automated. Thus, automation can vary across a continuum of levels, from the lowest level of fully manual performance to the highest level of full automation. Use of higher LOAs might allow for more vehicles to be controlled by a single supervisor. Unfortunately, these high LOAs tend to remove the operator from the task at hand and can lead to poorer performance during automation failures. In contrast, an intermediate LOA that involves both the operator and the automation system in operations may preclude multi-UAV control due to increased operator task requirements. However, it has been hypothesized that an intermediate LOA can improve performance and SA, even as system complexity increases and automation fails. Some research supports this hypothesis (e.g., Ruff, Narayanan, & Draper, 2002) and other results (e.g., Endsley & Kaber, 1999) suggest that there are factors that can impact the benefit of a LOA (e.g., whether task involves option selection versus higher-level cognition). Such results demonstrate the need for more research comparing LOAs in different task environments.

The Air Force Research Laboratory is conducting supervisory control human factors research utilizing a multi-UAV synthetic task environment. The present paper will focus on initial studies examining operator performance and SA with different LOAs and system reliabilities while supervising multiple simulated UAVs.

STUDY ONE

METHOD

Experimental Design

Two LOAs were evaluated. In Management-by-Consent (MBC), the operator had to explicitly agree to suggested actions before they occurred. The automation proposed route re-plans and target identifications, but required operator consent before acting. In Management-by-Exception (MBE), the system automatically implemented suggested actions after a preset time period, unless the operator objected. The settings for the MBE LOA (time limit until override) and the low/high reliability levels were: image prosecutions: 40 sec, 75/98%; route re-plans: 15 sec, 75/100%. The experiment employed a mixed design: 1 between-subjects variable (automation reliability, low/high) and 3 within-subjects variables (number of UAVs, LOA, and monitor arrangement). (Monitor arrangement will not be addressed here due to space restrictions.) The LOA variable was blocked and counterbalanced. UAV number (2 or 4) and monitor arrangement (horizontal or vertical) were also counterbalanced. After completion of training on the displays and all tasks/variables, each of the 16 participants completed 8 experimental trials, one sixteen-minute trial with each combination of independent variables.

Multi-UAV Synthetic Task Environment

The MIIRO (Multi-Modal Immersive Intelligent Interface for Remote Operation; Tso, et al., 2003) testbed was utilized, consisting of two monitors, a keyboard, and mouse (Figure 1). One monitor (Figure 2, left) presented the Tactical Situation Display showing the color coded UAV routes, suggested route re-plans, waypoints, targets, threat rings, and any unidentified aircraft. As each UAV passed a target, its camera took images and these appeared in the queue at the bottom of the Image Management display (Figure 2, right). The image in the top row of the queue was displayed. Suspected hostile targets within the image were highlighted by the automatic target recognizer (ATR) with red squares.

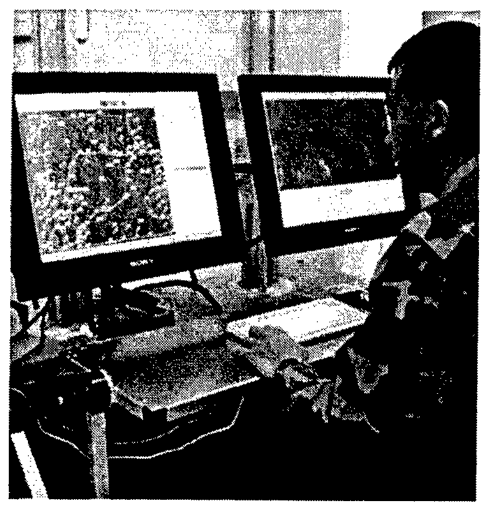


Figure 1. Multi-UAV Task Environment.

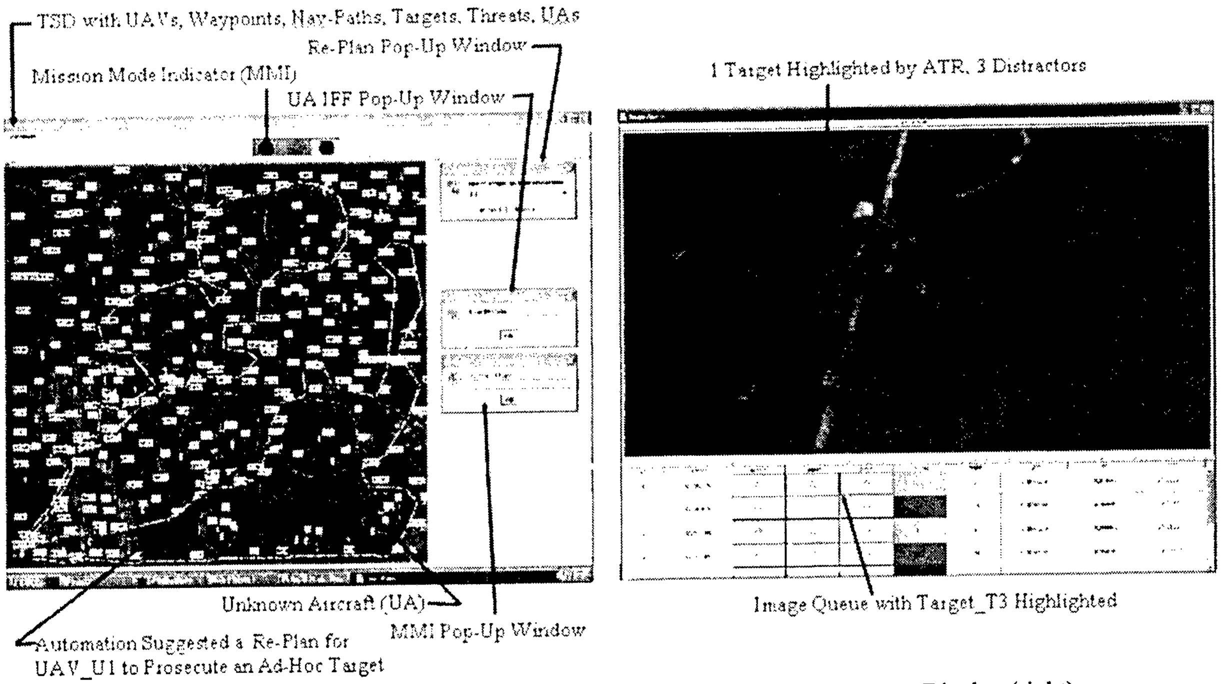


Figure 2. Examples: Tactical Situation Display (left) and Image Management Display (right).

Mission/Operator Tasks

Participants were required to respond to several types of events, listed in order of priority:

- Unidentified Aircraft (2 per mission). This task emulated having a highly unexpected, non-routine, high-priority event occur during a mission. When participants saw a red airplane icon appear, the response was to click on the symbol and enter a code in a pop-up window.
- Route Re-Plans (16 per mission). When alternate routes were suggested by the automation in response to ad-hoc targets and threats, participants were required to inspect the alternate route and make a decision to accept or reject the re-plan in a pop-up window, based on whether the re-plan crossed another threat or another UAV's route.
- Image Prosecutions (per mission: 34 (2 UAVs), 66 (4 UAVs)). Participants were required to view the image in the top window and verify that red boxes were only around targets (versus distractors). Participants could add or delete boxes by clicking on the items, if there were errors. Then participants made an accept/reject decision by clicking the appropriate box.
- Mission Mode Indicator (MMI) (per mission: 16 (2 UAVs), 32 (4 UAVs). This secondary monitoring task was used to represent the various contingency management panels that will likely exist in future stations. The panel's green light meant everything was operating normally. When this light extinguished and either the yellow or red light activated, then participants' response was to click on the panel and make an entry in a pop-up window.

RESULTS/DISCUSSION

Data recorded included time and accuracy in responses to: 1) image prosecutions, 2) proposed re-plans, and 3) system state changes and unknown aircraft. Workload, SA, and trust ratings were also collected. Results indicated that the tasks were manageable, but performance and subjective ratings decreased with:

- Increased number of UAVs: For image prosecutions, route re-plans, and MMI tasks, participants' average completion times were faster with 2 UAVs than 4 (all $p < .01$) and less time was spent in threat zones ($p < .05$). With the 2 UAV condition, participants were also more likely to respond before the automation acted

($p < .01$). The subjective ratings indicated that participants viewed the 4 UAV condition as higher workload, more difficult, and less trustworthy (all $p < .01$).

- Reduced automation reliability: Fewer images were prosecuted and more errors were made ($p < .01$) in the Low Reliability level compared to the High level. The subjective data also indicated that the participants had less trust when Reliability was Low ($p < .05$).

Performance between the two LOAs varied little and did not show a consistent trend across measures. The design dictated that trials with the MBC automation never timed-out. With MBE, participants typically responded rather than let the action automatically occur. In fact, image prosecution time averaged 12 seconds for both LOAs, much shorter than the criterion time limits employed. Thus, the results pertaining to LOAs are questionable, as the automation was not utilized as designed. Rather, the results suggest that the time criterions employed in the LOAs should be shortened significantly, to determine whether automation is a benefit in this simulated task environment. A follow on study was conducted to evaluate this change.

STUDY TWO

METHOD

Two of the three variables were the same as the first study: Automation Reliability (low/high; between-subjects) and LOA (MBC/MBE; within-subjects). A third (within-subjects) variable was Time Limit for the LOA ("short/long"). The short/long time limits to override were: image prosecutions (15/40 sec) and route re-plans (10/15 sec). The LOA variable was blocked and the order counterbalanced across subjects. The order of the Time Limit levels was counterbalanced within each LOA block. For all trials, there were 4 UAVs and the monitors were arranged horizontally. After training, each of 16 participants completed 4 experimental trials, one sixteen-minute trial with each combination of the independent variables.

All other procedures were the same as that used in the first study, except for how the route re-plan task was implemented. In Study One, participants were only required to inspect whether the re-plan crossed the path of another UAV or a threat zone. To better simulate the cognitive effort anticipated in operational missions, Study Two's re-route task required participants to view three readouts in a pop-up window that gave two fuel levels and the UAV's "resources" (low/medium/high). The accept/reject criteria was based on a mathematical relationship between these variables (e.g., if Fuel A plus .5 Fuel B is greater than 5 *and* Resources = Low, then Re-route should be accepted).

RESULTS/DISCUSSION

The efficacy and flexibility of the testbed were demonstrated by the successful change in the route re-plan task. (Average completion time, with longer Time Limit, was longer in Study Two (by 2.2 sec), presumably reflecting the changes in this task to increase its cognitive difficulty.) Also different in Study Two, only one measure showed a significant effect of Reliability: the percentage of images correctly prosecuted was less for Low, compared to High ($p < .01$). In regards to LOA, there were no significant differences in the performance and subjective measures, except as a function of the Time Limit variable. Participants' difficulty and workload ratings were similar for the two Time Limits for MBC LOA. With MBE, however, their ratings indicated the shorter limit was higher workload (Figure 3, left) and more difficult (both measures, $p < .05$). The participants' ratings may reflect the fact that their average time to complete image prosecutions was faster with the shorter time limit in MBE ($F(1,14) = 5.256$, $p < .05$; Figure 3, right) than the other three combinations of LOA and Time Limit. These findings may be related to the participants' ratings of less confidence with the shorter time limits ($p < .01$) and the nature of the LOA. In MBE, if the participant didn't respond to images before the time limit, they were automatically prosecuted. The fact that an erroneous action could occur, and more likely with the shorter time limit, may have pressured participants to respond faster and view it as higher workload. Thus, although MBE was hypothesized to be a workload reducer, it actually appeared to add to perceived workload.

Time Limit was also key in terms of the frequency in which the automation was exercised. Both image prosecution and route re-plans were more likely to activate automatically in trials with the shorter limit (e.g., 12.4% of the image prosecutions were automated in trials with the shorter limit, 1% with longer limit, the latter similar to Study One that employed a similar time limit). Yet, most re-plans and image prosecution tasks were completed manually, in less time (7.2 and 11.7 sec, respectively) than the available Shorter Time limits (10/15 sec).

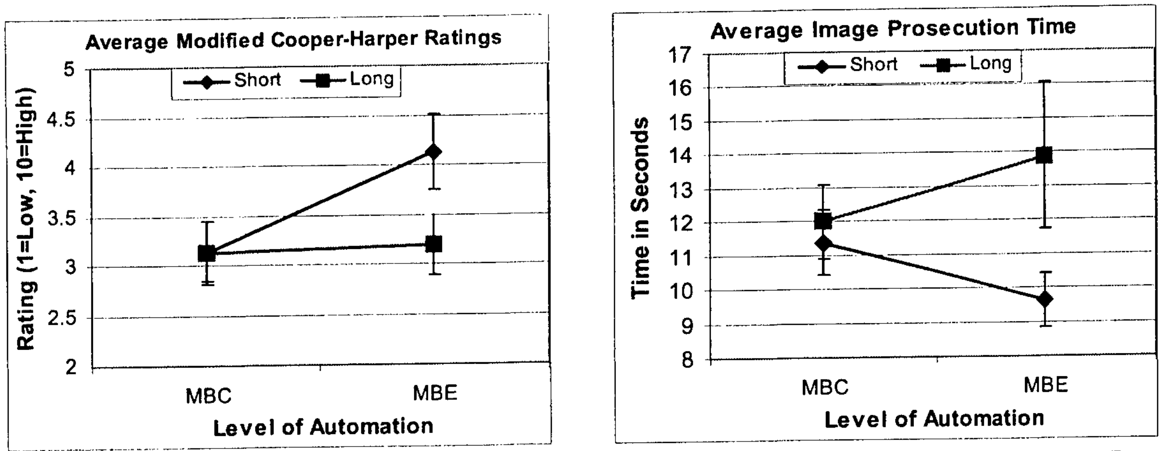


Figure 3. For each LOA (Management-by-Consent and Management-by-Exception) and Time Limit (Short/Long): Average Modified Cooper-Harper Rating for Workload (left) and Average Image Prosecution Time (right) with Standard Error of the Mean.

CONCLUSIONS

The rarity of automated actions, together with the increased workload and decreased re-plan and image prosecution times and confidence ratings with the shorter time limit, suggests that the participants preferred to respond manually rather than rely on the automation. At the very least, these results illustrate the complex relationship between LOA, time limits, and perception of difficulty and confidence. Moreover, participants' inclination to exercise the automation may increase in longer trials where vigilance effects are more likely to occur. Further research is needed before an optimal operator system design can be determined for supervision of multi-UAVs. This research will also explore the utility of additional LOAs that are: 1) contingency/task specific and 2) changeable during a mission, to better explore the utility of context-sensitive automation and decision aiding in UAV supervisory control.

REFERENCES

Endsley, M. R. and Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics, 42(3), 462-492.

Parasuraman, R., Sheridan, T. B., and Wickens, C.D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 30(3), 286-297.

Ruff, H. A., Narayanan, S., and Draper, M. H. (2002). Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned air vehicles. Presence: Teleoperators and Virtual Environments, 11(4), 335-351.

Tso, K. S., Tharp, G. K., Tai, A. T., Draper, M. H., Calhoun, G. L., & Ruff, H. A. (2003, October). A human factors testbed for command and control of unmanned air vehicles. Digital Avionics Systems 22th Conference, 8.C.1-1-12.

MILITARY APPLICATIONS

INTEGRATING CONTEXT FOR INFORMATION FUSION: AUTOMATING INTELLIGENCE PREPARATION OF THE BATTLEFIELD

Robin Glinton, Joseph Giampapa, Sean Owens and Katia Sycara
Carnegie Mellon University

Charles Grindle, Michael Lewis
University of Pittsburgh

ABSTRACT

Terrain information supplies an important context for ground operations. The layout of terrain is a determining factor in arraying of forces, both friendly and enemy, and the structuring of Courses of Action (COAs). For example, key terrain, such as a bridge over an unfordable river, or terrain that allows observation of the opposing forces line of advance, is likely to give a big military advantage to the force that holds it. Combining information about terrain features with hypotheses about enemy assets can lead to inferences about possible avenues of approach, areas that provide cover and concealment, areas that are vulnerable to enemy observation, or choke points. Currently, intelligence officers manually combine terrain-based information, information about the tactical significance of certain terrain features as well as information regarding enemy assets and doctrine to form hypotheses about the disposition of enemy forces and enemy intent. In this paper, we present a set of algorithms and tools for automating terrain analysis and compare their results with those of experienced intelligence analysts.

Keywords: terrain analysis, intelligence preparation of the battlefield, information fusion, GIS

INTRODUCTION

The particular type of terrain on which ground operations are conducted is a key determining factor of the types of operations and arraying of forces both for friendly and enemy forces, Terrain provides important context for analysis of sensed data as well as for guiding the tasking of data collection assets. The importance of the study and analysis of terrain has been recognized for hundreds of years in military science. Currently, such analysis is called the Intelligence Preparation of the Battlefield (IPB). IPB is a process that starts in advance of operations and continues during operations planning and execution. It provides guidelines for the gathering, analysis, and organization of intelligence. The purpose of this intelligence is to inform a commander's decision process during the preparation for, and execution of a mission.

The resulting products of IPB are identification of various areas of the battlefield that affect Courses of Action (COAs). Such distinctive areas include engagement areas, battle positions, infiltration lanes, avenue of approach etc. For example, an unfordable river is an obstacle, i.e. a terrain feature that impedes or prevents the maneuver of forces. Identification of such terrain features is invaluable since it allows the commander to make inferences about possible enemy avenues of approach and degree of vulnerability of his own force to enemy attacks. Such information, combined with information about possible enemy assets and force structure, e.g. tank platoon, or company or battalion, provide measures of *ease of movement (trafficability)* of forces throughout the terrain.

Key terrain is any location whose control is likely to give distinct military advantage to the force that holds it. Key terrain examples include road intersections that connect with a force's line of communication; a bridge over an unfordable river; or terrain that affords observation of the opposing force's line of advance. Key terrain areas cannot be defined by geographical features alone. The evaluation of terrain features must be fused with information about weather, enemy asset types, friendly and enemy range of fire, enemy doctrine and type of operation (e.g. defensive or offensive). For example, if an enemy tank company has been observed on the move towards an unfordable river, the presence of that river is not necessarily an obstacle if the company has an associated corps of engineers who could easily construct a bridge to allow passage. Hence the presence of the corps of engineers is a key element in a commander's threat assessment and evaluation. It is crucial for a commander to know whether enemy forces have occupied or are about to occupy key terrain. Therefore, key terrain areas identify areas where intelligence collection efforts should be focused.

An analysis of *concealment* provides areas that offer protection from observation and an analysis of *cover* identifies areas that offer protection from fires. The analysis of the terrain's suitability for providing concealment and cover result in the identification of *defensible terrain.* Fusing information about ranges of weapons with information on areas that provide poor concealment and cover identifies *engagement areas*: such areas are to be avoided by an attacking force, whereas they are potential engagement areas for a defending command. Therefore, the identification of defensible terrain and engagement areas is an important component supporting adversarial intent inference. To this end, engagement areas indicate areas where it is very useful to concentrate activity of collection assets.

Currently, IPB is done manually by intelligence officers using hardcopy maps on which they notate various significant areas, such as key terrain or defensible terrain. This manual process suffers from a number of inefficiencies: First, the hardcopy maps do not allow variable zooming in and out to obtain desired level of detail in an integrated, fast and consistent manner. Second, manually annotating the maps is time consuming. Third, notations on maps get cluttered with the risk of being misread, especially in the stressful times during operations. Fourth, depending on the experience and ability of individual intelligence officers and due to cognitive overload, various pieces of information could be disregarded or not used effectively in the process of the Intelligence Preparation of the Battlefield. Therefore, decision support tools that automate part of the process are highly needed.

Development of such decision support tools faces many challenges. First, computational algorithms must be developed to transform low level terrain information, e.g. soil types, vegetation, elevation slopes to higher level notions such as maneuverability of a force, engagements areas, defensible terrain etc. Second, appropriate cost schemes must be developed to allow expression of degree of strength of particular concepts of interest, for example degree of concealment that is afforded by a particular area. Third, since the IPB process is ongoing, spanning pre-operational activity and continuing throughout an operation, the computational algorithms must be efficient. Fourth, effective rule bases must be developed to allow combination of different pieces of terrain-based information with information about assets, weather, doctrine and results of sensors. Fifth, a user-friendly and flexible GUI must be developed for user interaction.

In this paper, we present a set of representation schemes and algorithms developed for automated terrain analysis and compare their conclusions with those of experienced intelligence analysts.

Automating MCOO development

IPB is a cyclical process that continues throughout the planning and execution stages of a mission. The goal of IPB is to guide the collection, organization and use of intelligence. IPB products identify areas in the terrain where intelligence collection efforts should be focused in order to discern the intent of the opposing forces commander. Terrain analysis is performed in order to identify the potential effects of terrain in the operation of friendly or enemy forces. The initial product of the analysis is the Combined Obstacle Overlay (COO). Combining the COO with Key Terrain, Defensible Terrain, Engagement Areas, and Avenues of Approach results in the Modified Combined Obstacle Overlay (MCOO). The features in the MCOO are high level terrain-based concepts of crucial tactical significance.

Trafficability

Fig. 1 shows separate overlays, each of which depicts untrafficable terrain due to vegetation and soil type, weather and surface drainage, slopes, minefields, trenches, and bodies of water. These are combined to form an overlay that shows all obstacles. We use as our terrain representation the Compact Terrain Database (CTDB) format used by the OTBSAF simulation software. The CTDB format gives us access to a grid of elevation values as well as an associated soil type for each grid cell. We use the elevation grid to calculate both slope and surface configuration. Surface configuration refers to whether a grid cell lies on a flat surface, a concavity like a hill, or a convexity like a trench. This calculation allows us to judge the effects of precipitation on a certain grid cell. Rain, for example, is much less likely to affect the trafficabilty of a region that lies on top of a small hill than it would a previously dry riverbed. The grid surface is smoothed and these regions identified as shown in Figure 2.

Vegetation in OTBSAF's CTDB database is limited to tree canopies so at this point the tree spacing is assessed to determine if it is sufficient for the given vehicle type to pass. Next the slope of the grid cell under consideration is compared to the maximum trafficable slope for the given vehicle type. If the slope is less than this value, the slope is passed on to a vehicle speed calculation where it is used as a multiplier for the base vehicle speed. The base vehicle speed is the vehicle's maximum speed on flat terrain for the given soil type. The

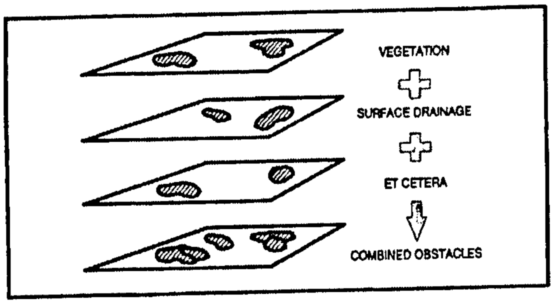


Fig. 1. Obstacle overlays combined to form COO.

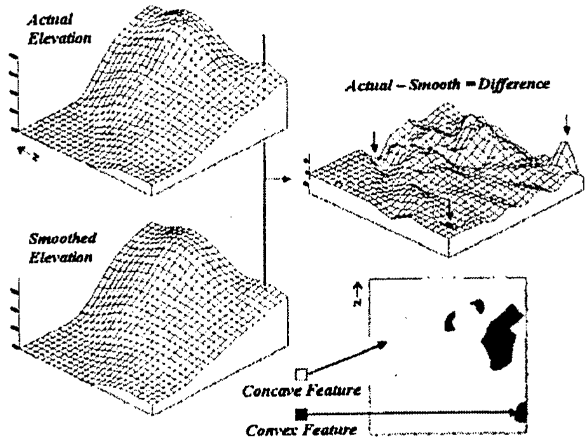
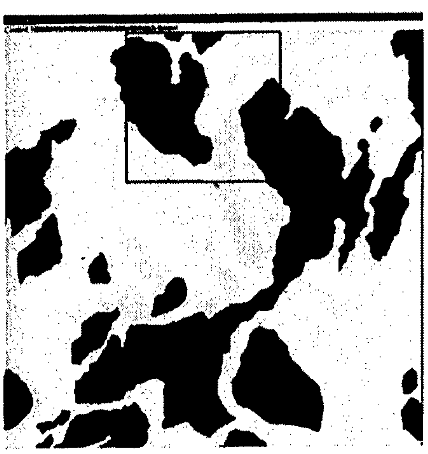


Fig. 2. Surface configuration calculation

speed also takes into consideration weather and surface configuration. If the surface is concave and there is precipitation then the speed calculation uses the wet soil type value. Otherwise the dry soil type value is used. The result of the trafficability calculation is shown in Fig 3. Computational details for determining surface configuration and other aspects of automated terrain analysis are presented in (Glinton, et al. in press).



Fig, 3. Result of trafficability calculation

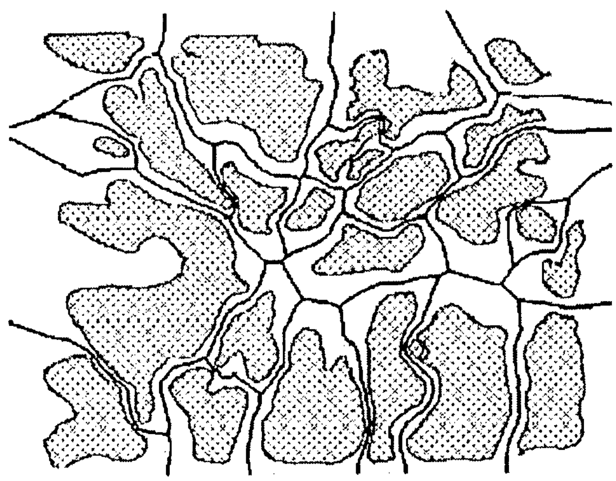


Fig. 4. Generalized Voronoi diagram of NO-GO regions

The COO tells us at a glance the ease of movement for a given vehicle type through a certain grid cell on a terrain. If a corridor is too narrow to support travel in formation, however, the unit must change formation. The reduced speed and dispersed forces caused by narrow corridors or canalizing terrain makes units more vulnerable to attack. Our automated terrain analysis uses *configuration spaces*, a technique commonly used in path planning for mobile robots to identify these features. The Voronoi diagram, a common tool from computational geometry (de Berg et al., 2000) is then used to express the topology of unrestricted regions. Fig. 4 shows a generalized Voronoi diagram (GVD) (Choset, et al., 2000) calculated using the NO-GO regions of a heavily restricted COO. Notice how GVD edges correspond with mobility corridors through the terrain while GVD vertices occur in enclosed regions. These properties lend themselves to automating the identification of avenues of approach, defensible areas, and other important tactical features of terrain. By treating paths through this network as a circuit posing resistances through restrictive terrain and weapons emplacements defensive analysis becomes a study of what areas best provide resistance to an encroaching enemy while an offensive analysis aims to find the weak points in the enemy's ability to apply resistance.

Engagement Areas

The army field manuals instruct the terrain analyst to consider cover and concealment and favor enclosed regions in choosing engagement areas. The GVD vertices are prime candidates because they only occur in enclosed regions. A line of sight analysis between the location of such a vertex and its surroundings are used to assess the amount of cover and concealment available providing a first ranking. To choose among the many candidate engagement areas a circuit analysis is then used considering enemy movement along an expected axis such as from the SE corner to the NW corner of the operational area. By considering possible defensive manning allocations and the resulting resistance the engagement areas most disruptive to enemy movement can be identified.

Avenues of Approach

An *avenue of approach (AA)* is a route that an attacking force can use to reach an objective. Features that must be considered in the evaluation of AA's are

- Degree of canalization (presence of choke points)
- Sustainability (access to a line of sight)
- Availability of Concealment and Cover
- Obstacles

Avenues of approach are found using a technique similar to that used to find engagement areas. In this case the resistance of identified candidate engagement areas are increased. The mobility corridors with the highest current flow are then chosen as components of the avenues of approach.

Named Areas of Interest

Named areas of interest (NAIs) are areas of terrain that have particular tactical significance because they overlook potential engagement areas or canalized avenues of approach allowing the force that controls them early observation of enemy movements. While cultural features such as bridges can also qualify as NAIs, our approach is based on analysis of elevation and lines of sight to choose patches of ground that offer the broadest coverage of possible avenues of approach and engagement areas.

METHOD

Two subject matter experts (SMEs) with field experience in intelligence analysis were videotaped and provided think aloud verbal protocols while filling in MCOO overlays for a map generated from CTDB data. Their instructions for the portions of the task presented in this paper were:

> You are the S-2 of 1-22 Infantry battalion. Your battalion is located in the North West corner of this map. Your battalion is to seize an objective that is located in the Southeast corner of this map. You begin the Military Decision Making Process (MDMP) by doing terrain analysis and developing your MCOO. Please annotate the following: a) Slow-go/No-go terrain, b) Identify enemy engagement areas and potential defensible terrain (and the size of force that he could defend with), c) Named Areas of Interest (NAIs) (given that you will be moving from the Northwest to the Southeast), d) Display with a double arrow the path with the least terrain resistance and display with a single arrow an alternate path.

RESULTS

Figure 5 depicts the major annotations made by SME-1 on the MCOO overlay. The double headed arrow indicates the primary avenue of approach. Single headed arrows denote the secondary avenues of approach. The boxes represent engagement areas and the smaller boxes with lines indicate named areas of interest (NAIs). The results of the analysis completed automatically by our terrain analysis algorithms are shown along side in Figure 6. The regions marked with an X represent engagement areas. An arrow with a solid head denotes the primary avenue of approach while an arrow with a clear head denotes the secondary avenue of approach.

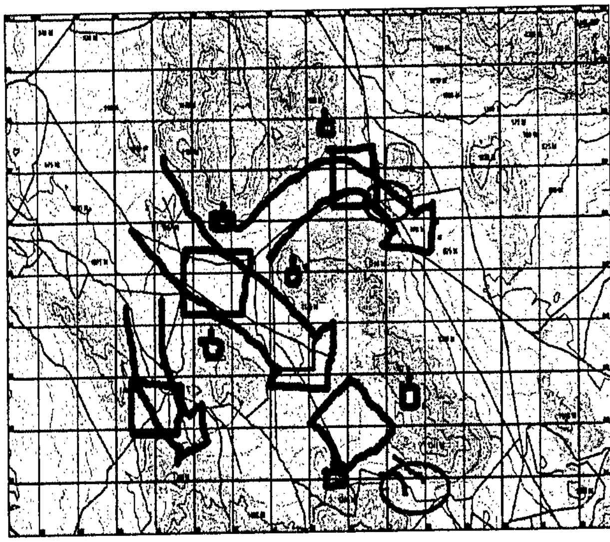


Fig. 5. MCOO produced by SME-1.

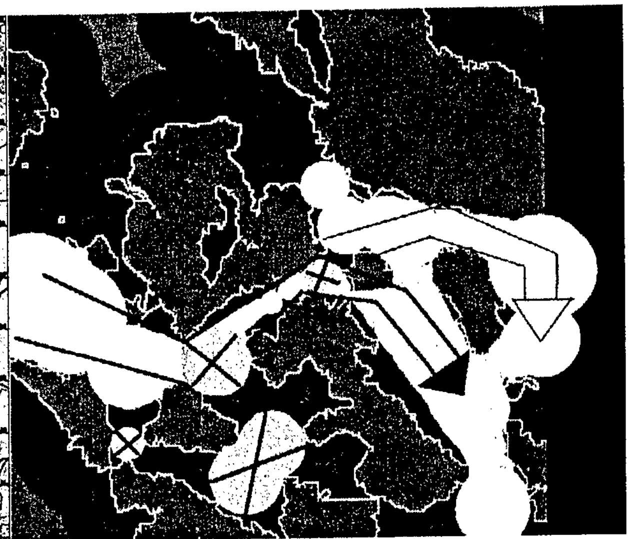


Fig. 6. MCOO produced by automated terrain analysis.

Our analysis chose the same primary avenue of approach as SME-2. This avenue of approach coincided with SME-1's choice as a secondary AA. This discrepancy between the program and SME-2's choice of the "Eastern route" and SME-1's choice of the more direct "Southern route" appears to lie in the SMEs' prior command experiences. Of the two paths circled in Figure 5, the one closest to the bottom of the map is the most canalizing. SME-1 indicated that although this made the path more dangerous, the shorter path to the objective made the added risk acceptable. This reasoning was not available to the program because path length is considered only indirectly through its affect on resistance in determining ranking. The agreement between the program and SME-2 shows, however, that even its current stage of development our automated terrain analysis identifies avenues of approach within the range of variation among human SMEs. We hope to include facilities to allow users to interactively adjust cost functions to express such value judgments in the next version of our software. A solution as simple as a slide bar with safety on one end and speed on the other would allow the user to indicate the desired balance by positioning the slide bar to modify the weight given path length in path resistance calculations.

There is good correspondence between our selections of NAI's with those of the SMEs. However, the SME is limited by the granularity of the map. A physical map cannot be "zoomed in" to find some feature that does not appear at the resolution used for printing it. Our algorithms, however, can calculate line of sight between engagement areas and their surroundings with high precision from high-resolution elevation data. For this reason our algorithms also produce more candidate NAI's. The NAI's selected by our algorithms are shown in Figure 6. Of the eight NAIs identified, three were found by both SMEs and the program, two were identified jointly by SME-1 and the program, one was identified by both SMEs but not the program, and two singletons were found, one by SME-1 and the other by the program. The program again fell well within the range of variation of the SMEs matching more of the NAIs identified by SME-1 then did SME-2.

There is an exact correspondence between SME-1's choice of engagement areas and our algorithm's top 3 selections. The algorithm's 4th selection, the closest to the bottom of Figure 6, is positioned slightly differently from this expert's final choice. This is because our program currently tries to pick candidate regions for engagement areas so that they control as many approaches as possible. The SME realized that two of the three paths entering this region had already been covered by previous engagement area choices. This suggests that we should consider topology in the selection of candidate engagement areas. Currently topology is only considered for culling the candidate engagement areas. SME-2 chose a single engagement area which was among those chosen by SME-1 and the program. The discrepancies in SME-2's overlay

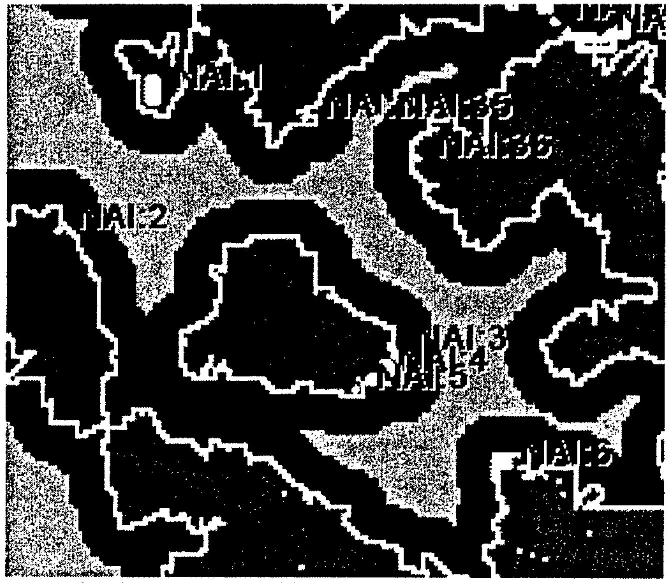


Fig. 6 NAI's selected by automated terrain analysis.

seem to stem from an early choice of an extreme Eastern path as a secondary route. Because the "Southern route" was not chosen, NAIs and engagement areas along its path were considered less closely.

DISCUSSION

Our work in terrain analysis is ultimately meant to inform high-level information fusion. Only by capturing the context within which targets are identified and tracked can we attribute intent to their actions and guess at what else may be out there that we have not yet seen. Our early success in automating the MCOO process has exceeded our expectations and we are now extending the informal comparisons presented here with a full-fledged validation effort using a larger sample of SMEs with varying levels of experience and a larger collection of terrains.

ACKNOWLEDGEMENTS

This research was supported by AFOSR grant F49620-01-1-0542. We wish to thank our subject matter experts LTC John Richerson and MAJ Ronald Bonomo for their ungrudging help, assistance, and guidance without which, this research could not have been conducted.

REFERENCES

Glinton, R., Grindle, C. Giampapa, J., Lewis, M., Owens, S. and Sycara, K. (to appear) Terrain based information fusion and inference, Proceedings of the Seventh International Conference on Information Fusion, Stockholm, Sweden, July 28-July 1.

Mark de Berg, Marc van Kreveld, Mark Overmars and Otfried Schwarzkopf. Computational Geometry Algorithms and Applications. Springer, Berlin, 2000.

FM 5-33 Terrain Analysis, Headquarters, Department of the Army, July 1990.

FM 34-130 Intelligence Preparation of the Battlefield, Headquarters, Department of the Army, July 8,1994.H.

Choset, H.,Walker, S., Eiamsa-Ard, K. and J. Burdick, J. Sensor-based exploration: incremental construction of the hierarchical generalized voronoi graph. The International Journal of Robotics Research, Vol. 19, No. 2, February, 2000, pp. 126 - 148.

AN EVALUATION OF SPEECH RECOGNITION TECHNOLOGY IN A SIMULATED AIR BATTLE MANAGEMENT TASK

Nicole M. Guilliams, Annie B. McLaughlin, Michael A. Vidulich, W. Todd Nelson, Robert S. Bolia, Brian P. Donnelly

Air Force Research Laboratory

ABSTRACT

Air Battle Management (ABM) is a complex and demanding activity that involves numerous tasks performed at the ABM workstation, which currently includes panels of toggle switches, knobs and dials, a trackball, a keyboard, and numerous reconfigurable pushbuttons. Although functional, such workstations are manually-intensive, may require extensive training, and could subject operators to unacceptable levels of workload. The primary goal of this research was to evaluate the appropriateness of speech recognition technology for workload reduction in ABM work domains. A simulated Batterfield Air Interdiction (BAI) mission was employed. Results indicated significant advantages for the speech control interface with respect to performance efficiency and perceived mental workload. In addition, when given a choice, operators preferred to employ speech inputs over manual inputs for a variety of control functions. These findings are discussed in terms of the appropriateness of speech control technology for ABM applications.

Keywords: AWACS; Air Battle Management; Speech Recognition; Command and Control; C2

INTRODUCTION

Air battle management (ABM) responsibilities involve directing the implementation of the air tasking order (ATO) and controlling the execution of the associated air-to-air and air-to-ground operations. ABM also involves the monitoring and manual control of up to eight communication channels, including radios and intercoms. Performing these tasks requires tracking an enormous amount of information, such as the position, heading, altitude, and speed of both friendly and hostile aircraft, and the fuel and armament status of friendly aircraft (Fahey et al., 2001). The ABM workstation includes panels of toggle switches, knobs and dials, a trackball, a keyboard, and numerous reconfigurable pushbuttons. Although functional, such workstations are manually-intensive and require extensive training. During periods of low to moderate air traffic, an ABM can comfortably manage the tasks. However, during periods of heavier air traffic, operators are likely to reach unacceptable levels of workload, which may negatively impact performance efficiency and mission effectiveness.

Speech Recognition

One possible way to reduce the high manual demands placed on air battle managers is to implement speech recognition into the ABM environment. Automatic speech recognition is a reasonably mature control technology that has been under development for almost three decades. It has been widely accepted and used in the commercial world in systems such as telephone call handling, telephone dialing technology, speech-based control of numerous appliances and devices for physically-disabled users, and telephone-based banking and credit card systems (see McMillan, Eggleston, & Anderson, 1997 for review). It has also been tested in various experimental military platforms including simulated F-16 cockpits and UAV ground control stations, as well as in theater air planning systems. Research indicates that speech-based control may be particularly effective when used in conjunction with complex control tasks that would normally require manual input (Williamson & Barry, 2000). The primary goal of the research described in this paper was to extend the evaluation of speech recognition technology by assessing its appropriateness for application in complex ABM work domains.

METHOD

Participants

Twelve active-duty Air Weapons Officers (AWO), eleven male and one female, served as participants. All were trained in basic ABM procedures, but had varying levels of experience from Basic Qualified to Mission Ready.

Experimental Design

A 2 CONTROL MODALITY (speech, all-manual) × 3 BLOCK (I, II, III) within-subjects design was employed. Specific tasks were completed by using only the speech or all-manual interface. Participants completed six trials, alternating between the speech and all-manual interfaces, and one preference trial, in which they were free to select which interface or combination of interfaces they wanted to use to complete the tasks.

Apparatus

Data collection was executed in a medium-fidelity simulated Airborne Warning and Control System (AWACS) operating environment comprising six PC-based operator workstations, a spatial audio intercom system, and a speech recognition system. Spatial separation of the communications channels was achieved using AuSIM, Inc. spatial intercom technology. Speech recognition was achieved using Nuance 8.0, a commercial-off-the-shelf speech recognition system. A modified version of the Solypsis Tactical Display Framework 3.7 Prototype AWACS Display was employed at each of the operator workstations to support the three primary experimental roles: the AWO, the Senior Director, and the Strike Lead. Finally, background white noise (approximately 85 dB) was generated to simulate an ambient noise environment comparable to that of an AWACS or other airborne ABM platform.

During the experiment, participants controlled Batterfield Air Interdiction (BAI) scenarios under speech or all-manual conditions. Following each BAI scenario, operators provided ratings of perceived mental workload using the NASA Task Load Index (NASA TLX; Hart & Staveland, 1988).

BAI Scenario

A BAI mission was employed throughout the experiment. As defined by the Department of Defense (JP 1-02, p.21), BAI missions involve "air action by fixed- and rotary-wing aircraft against hostile targets that are in close proximity to friendly forces and that require detailed integration of each air mission with the fire and movement of those forces." BAI missions have the characteristic of being communications-intense, high workload missions, which require operators to monitor multiple channels of communication while simultaneously performing a set of airborne command and control tasks.

The AWO's communications workload was increased by having them perform an adapted version of the Coordinate Response Measure (CRM; Bolia, Nelson, Ericson, & Simpson, 2000) throughout the entire mission. The CRM requires listeners to respond to short phrases comprising a call sign followed by a color-number combination (e.g., "Ready Baron, go to Blue Five now."). Participants were instructed to listen for a specific call sign ("Baron") and, if detected, to enter the color-number combination contained in the phrase into a keypad.

Figure 1 illustrates the events and tasks associated with the BAI mission. As can be seen in the figure, the trial began with the set-up phase, in which the AWO marked the controllers for each aircraft, set an initial bulls-eye, and sorted the ATO list for easier manipulation. These tasks were completed with speech commands or all-manual inputs. The ingress phase began upon completion of the set-up phase, at which time a package of fighter aircraft entered the area of responsibility and began to *check in*. At that point, the AWO also began making *picture calls* to alert the aircraft of threats, and monitoring radio frequencies for critical call signs as part of the CRM task. The retargeting phase began with the Senior Director passing the first target change, in the form of *nine-lines,* to the AWO. Once received, the AWO notified the strikers, passed the changes to them, and retargeted the aircraft on his or her own display. Throughout the retargeting phase, the AWO also continued to make threat calls and listen for critical callsigns (CRM task).

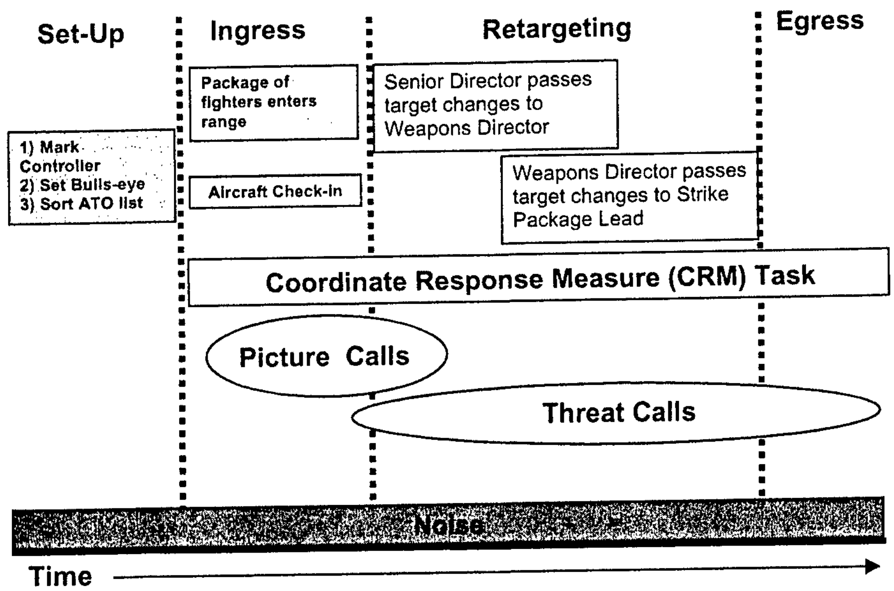


Figure 1. Schematic overview of the Batterfield Air Interdiction mission and associated experimental tasks.

Egress served as the final phase, which started after all of the nine-lines were passed to the strikers.

RESULTS

Mission Performance Efficiency

Mission performance efficiency refers to the participants' speed in performing the basic tasks that they would normally do as part of a real mission. Three measures of performance efficiency were especially relevant to the investigation of control modality: (1) *Set-up Phase Duration* – the time required to complete the configuration of the interface in the set-up phase of the trial; (2) *Nine-Line Transmission Time* – the time required to receive the four nine-line transmissions from the Senior Director and to transmit them to the Strike Package Lead; and (3) *Strike Package Repairing Times* – the time required to update the pairings information to correspond to the new missions.

Analysis of these data with separate 2 (CONTROL MODALITY) × 3 (BLOCK) repeated measures ANOVAs revealed significantly faster performance with the speech control as compared to the all-manual condition for the *Set-up Phase*, $F(1, 11) = 6.97, p < .05$; *Nine-Line Transmission Time*, $F(1, 11) = 5.50, p < .05$; and the *Strike Packages Repairing Times*, $F(1, 11) = 11.03, p < .05$. These main effects are illustrated in Figure 2.

Workload Ratings

Mean overall workload ratings were submitted to a 2 (CONTROL MODALITY) × 4 (PHASE) × 3 (BLOCK) repeated measures ANOVA, which revealed a significant Control Modality × Phase interaction, $F(3, 33) = 2.99, p < .05$. As can be seen in Figure 3, the significant interaction can be explained by noting that although workload ratings associated with the speech control were lower than the all-manual condition across all phases, the most pronounced effects occurred in conjunction with the retargeting phase.

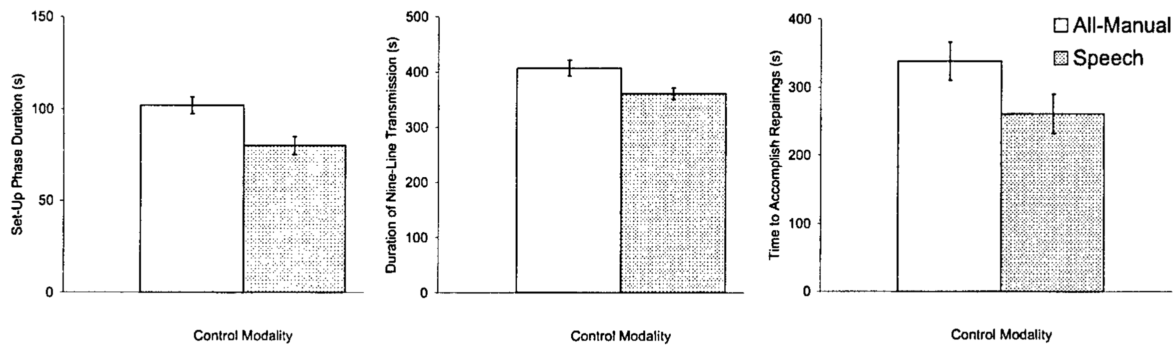


Figure 2. Control modality significant main effects for mission performance efficiency. Speech control was found to produce significantly shorter durations for Set-Up, Nine-Line Transmission, and Strike Package Repairings.

Preference Trial Summary

For the preference trial, participants were instructed to use their choice of the speech interface, the all-manual interface, or any combination of the two to complete the set of tasks required by the scenario. Preference trial data comprised the percentage of participants that employed speech control inputs while completing the preference trial. These data are presented in Figure 4, which shows the mean percentage of participants who used speech inputs for each of the 13 speech-enabled functions. Inspection of the figure reveals that most of the participants chose to employ speech inputs for a majority of the tasks. In fact, for 10 of the 13 speech-enabled tasks, 50% or more of the participants chose to use the speech interface over the all-manual interface. More impressive was the finding that, for three of the tasks, all 12 participants chose to use speech over the all-manual control.

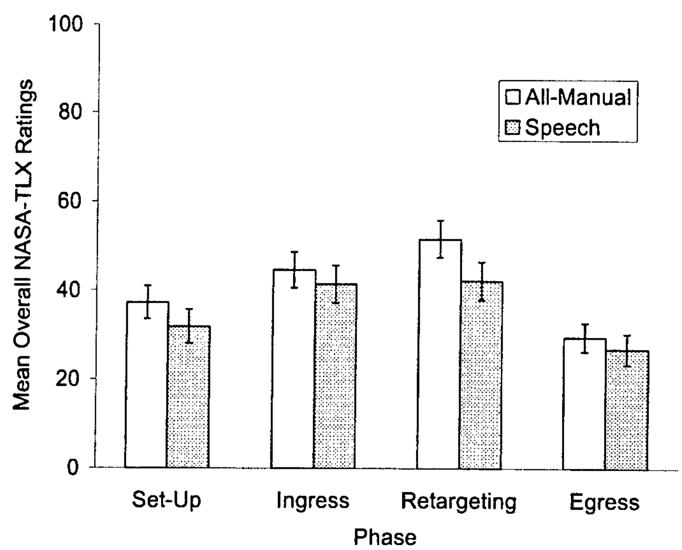


Figure 3. Operator workload ratings by phase.

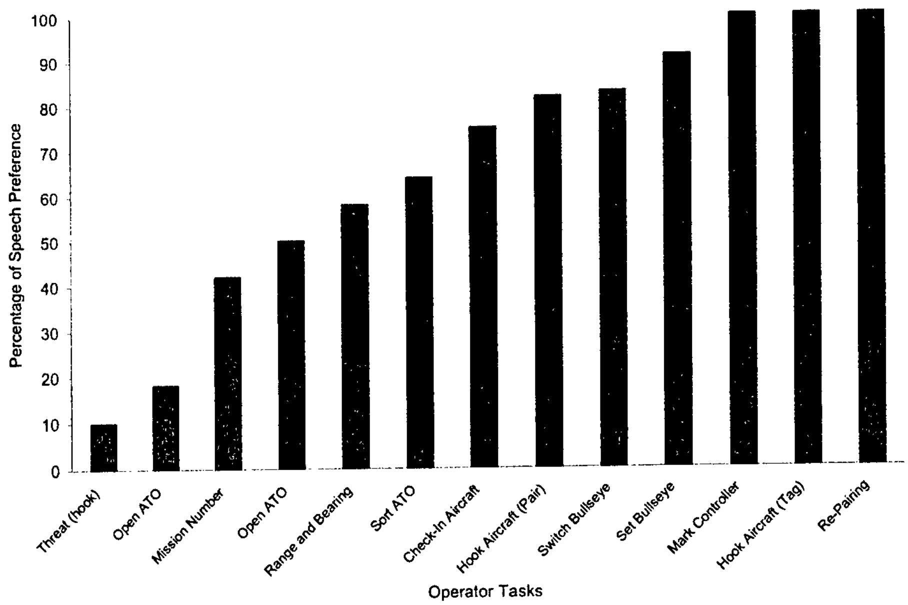


Figure 4. Operator preference data depicting the percentage of tasks completed using speech commands.

DISCUSSION

The purpose of this study was to evaluate the appropriateness of speech recognition technology for application in the ABM work domain. The results of this study indicate that speech recognition may be effective for use in this type of environment. A performance efficiency advantage was seen through task completion times. Participants completed the set-up, nine-line transmission and repairing tasks significantly faster using the speech interface than with the all-manual interface. Operator workload ratings also greatly favored the speech interface – i.e., workload was rated lower across all mission phases, with the greatest difference occurring in the retargeting phase. In addition to the performance efficiency and workload findings, data from the preference trials indicated that operators found the speech interface intuitive and easy to use, electing to complete most of the tasks using speech commands.

The results of this study clearly support the implementation of a speech interface into the ABM environment. However, there are several limitations of the current study. First, the speech vocabulary that was employed was very restricted, comprising approximately 30 words and commands. Given the complexity of real ABM task environments, it may be necessary to utilize hundreds of words and/or commands to fully speech-enable the ABM interface. Accordingly, it will be important to develop more comprehensive and robust vocabularies and grammars, and assess their appropriateness in representative ABM scenarios. As with many simulated environments, the complexities and challenges of the real environment are difficult to mimic. This is especially true in the simulation employed in the experiment described herein, which limited the number of operators to three. Furthermore, this investigation only assessed the utility of speech input for a BAI scenario, which may not generalize to other ABM missions. Despite the limitations, the present study was able to replicate and extend the findings of Nelson and his colleagues (2003), who assessed the utility of a similar speech interface in a non-interactive ABM task environment. Given the results of the present study, there is strong support for further exploration and development of advanced speech recognition interfaces for ABM work environments. Additional research is clearly warranted and should focus on the development of more robust vocabularies, as well as the utility of this technology in different ABM scenarios.

REFERENCES

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multi-talker communications research. Journal of the Acoustical Society of America 107, 1065-1066.

Dictionary of Military and Associated Terms (JP 1-02). Department of Defense. <<www.dtic.mil/doctrine/jel/new_pubs/jp1_02.pdf>>. 8 Dec 2003.

Hart, S. G., & Steveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.), Human Mental Workload (pp. 139-183)

Fahey, R. P., Rowe, A. L., Dunlap, K. L., deBoom, D. O. (2001). Synthetic task design: cognitive task analysis of AWACS weapons director teams (Tech. Rep. No. AFRL-HE-AZ-TR-2000-0159). WPAFB, OH: Air Force Research Laboratory.

McMillan, G. R., Eggleston, R. G., & Anderson, T. R. (1997). Nonconventional controls. In Salvendy, G. (Ed.), Handbook of human factors and ergonomics *(2nd ed.)*. NY, NY: John Wiley & Sons.

Nelson, T., Vidulich, M., Bolia, R., McLaughlin, A., Guilliams, N., & Donnelly, B. (2003). Designing speech interfaces for command and control applications. Proceedings of the Sixth Annual International Australian Aviation Psychology Symposium. Victoria, Australia: Australian Psychology Aviation Association.

Williamson, D.T., & Barry, T.P. (2000). The design and evaluation of a speech interface for generation of Air Tasking Orders. Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting, 750-753. Santa Monica, CA: Human Factors and Ergonomics Society.

A QUALITATIVE REDEFINING OF MILITARY UTILITY IN TERMS OF HUMAN FACTORS

David G. Smith, Lieutenant Colonel, USAF
Air Force Flight Test Center and Embry-Riddle University

ABSTRACT

The term "military utility" is often addressed in the acquisition community, yet it lacks clarity as a measurable and statistically relevant concept that can provide fact-based assessment of a systems worth to the warfighter. This research will support a redefining of the term "military utility" to one specifically addressing human factors and human factor integration. Methodology for this assessment will include a review of the current concept of military utility, analysis of three relevant proposed concepts for this term and their associated impacts, support for the focus on human factors as the critical element in military utility that the tester and acquisition professional should be concerned with, and ultimately concluding by tying the evidence to the solution for redefinition of the broad term military utility to a precise and measurable one.

Keywords: Military Utility; Cost as an Independent Variable; Survivability; Maintainability; Lethality; Parametric Analysis; Dimensional Analysis; Systems Analysis; Key Performance Parameters

INTRODUCTION

Efforts to improve the products produced for the warfighter through the acquisition community are not new. Recently a focus on "military utility" has become a driver in the determining of capabilities that need to be evaluated before a system is procured and delivered. Unfortunately, the term "military utility," while appearing to posses tangible benefits, is far to broad and loosely defined to allow for quantifiable exploitation. There is a need for clear data in support of evaluation and decision making. The current definition is not accepted across the acquisition community as one with sufficient fidelity to make decisions. This research will demonstrate that a focused definition will enhance results, thereby directly supporting the warfighter.

Military Utility Today

Defense Acquisition University (DAU), from their glossary, defines military utility as "The military worth of a system performing its mission in a competitive environment including versatility (or potential) of the system. It is measured against the operational concept, operational effectiveness, safety, security, and cost/worth. Military utility estimates form a rational basis for making management decisions."

Up until recent review of the DoD 5000 series guidance, not even this definition (just a void) existed. The term was used commonly in and around the test and acquisition community, but was used without regard to a precise understanding. This created situations where the message became clouded as a deliverer of a communication using the term military utility may be faced with a receiver that does not share the same concept. (Rajadhyaksha) This lack of fidelity directly impacted the B-52 Avionics Midlife Improvement (AMI) program as concepts relating to military utility were desired by Air Force Flight Test Center engineering personnel but understood differently by the test operators as well as the Boeing designers and test report writers. (Farrell) Had a more focused and understood definition been accepted, the initial design of the B-52 AMI test plan would have been streamlined. This plan was intended to be benchmarked as also creating the template for the generation of a Test Report. This "outcome based" document is required by AFFTCI 99-3 65 days after completion of the test program, but is historically late. Insurance of timely report completion through early design was the goal, unfortunately the misunderstandings resulted in considerable discussion eventually resulting in a more conventional plan to address the report process.

Human Factors Integration

The integration of human factors has been increasing considerably over the past several years. Defining aerospace achievements in engineering terms only is incomplete at best. Early engineering efforts placed a low priority on the tailoring of systems to the needs of the operator. Wiener points out that human factors has grown from an unrecognized byproduct of design to a specific discipline in the 1950s. Today human factors is considered a "core technology" evaluated in much the same way as powerplants, navigation, and communications. (Orlady) The DAU definition is clearly lacking in addressing this "core technology," and appears mired in the past concepts addressed by Wiener as placing a low priority on the needs of the operator.

Interestingly, the new DoDD 5000.2 has a two page enclosure devoted to "Human Systems Integration." This new directive clearly signals and enforces the need to manage the integration of human factors into the acquisition process. The Program manager is now given clear direction that issues such as human factors engineering and survivability must be addressed, but the guidance still fails to fully embrace an integrated human element. Instead, guidance directs that the system be "built to accommodate the characteristics of the user population."

METHOD

Proposal One: the Status Quo. It may be argued that the military community has in the past produced outstanding systems with phenomenal capability, but this does not justify an absence of change. General Jumper, the Chief of Staff made clear the need for building a culture of "continual transformation." Past successes do not relieve the military of the need to continually improve. The United States Army recognizes this in their concept of operations for the National Training Center. The 11th Armored Cavalry Regiment serves as an opposing force of exceptional fidelity. They train our soldiers against the most capable and lethal threat that they may ever face, not just a projected threat.

The tester of new systems, weather military or civilian, is faced with a requirement to produce data. The current definition is far too broad to produce useable data to support system evaluation. Although there have been strong improvements in addressing this concept, the criteria is still unwieldy and unquantifiable. Statistical quality control is defined as "the application of statistical techniques in all stages of an operation in order to meet established standards of quality in the most economical manner." (Braverman) Further, he states that data must be interpreted numerically, and related to batches of sufficient size to represent a population. This suggests that the broad definition of military utility currently in use would make statistical analysis difficult at best.

Proposal Two: Cost, Schedule and Content. When asked for a product delivery that was cheep, fast and good (cost, schedule and content), an often accepted concept in the civilian world is "pick any two." This may be good enough for a new product launch, but will not sustain the long term viability of a product, and clearly may be contrary to military use. The concept of cost as an independent variable is a fact of life, (DoDD 5000.1, 2003) but it should not be the only driver used to assess system viability. Schedule is important, but should not result in faulty design, and quality is essential, but not the only feature considered in the total solution. All three critical features must be considered as part of the overall system.

The cost of the system is defined by DoD in terms of it's entire life cycle. This cradle to grave mentality is designed to include all features relevant to research, acquiring and disposing of a potential system. The Glossary for Acquisition Terms shows that Cost as an Independent Variable (CAIV) methodologies are used to acquire affordable DoD systems by setting aggressive but achievable life cycle costs. This is accomplished by trading off performance and schedule, as needed, to meet pricing demands balancing mission needs against projected year-out resources.

It may seem obvious that systems are needed in relation to a time variable. What is not so obvious is the impact of schedule on performance. Secretary Rumsfeld stated in a Senate Armed Services Committee Briefing: "Too many weapons take too long to reach the battlefield because of requirements like the congressional rules that systems must pass operational tests before they can be fielded. That requirement means that some useful weapons never get the stamp of approval." (DAU TST-301, 2003) The National Defense Authorization Act supports rapid acquisition and deployment for products under development or available through the commercial sector, or articles urgently needed to counter a threat; but, they must include an operational assessment in accordance with Developmental and Operational Test and Evaluation (DOT&E) guidance. These two statements appear contradictory as assessment is required in one (DOT&E) while being underplayed in Secretary Rumsfeld's comments.

The content of the system is the "meat and potatoes" of the process--what is there and what can it do. The new approach to acquisition as demonstrated in the rewrite of the DoD 5000 series regulations addresses a focus on technology development and risk reduction as well as the need for rigorous exit criteria before program commitment. The goal of this rewrite is to deliver advanced technology to the warfighter faster, with reduced ownership costs.

Survivability, Lethality, Maintainability

The "buzz words" of content are survivability, lethality and maintainability (Glossary, 2001). Survivability is the ability of the system to avoid or withstand a man-made hostile environment while being able to accomplish its mission. Lethality is the probability that a weapon will engage and destroy a target. Maintainability is the ability to keep, or restore, a system to a specific condition. These three concepts comprise content.

It is apparent that using the concept of cost, schedule and content to define military utility may increase the ability to manage data to particular areas, but there is constant interaction and interface between these three critical parameters. This interface would make management of data still awkward and unwieldy. At what point would cost demand concession from schedule? When would lethality demand a compromise of survivability? These issues demand hard data to support statistical analysis. Unfortunately, if addressed under the guise of military utility they still provide too broad of an assessment arena for fact based study.

Proposal Three: the Future. Our weapons systems have a clear commonality that can't be avoided. Each of our systems requires an interface with a human element. Recent development of "unmanned" systems do not delete this interface, they simply relocate it away from the system or theater. The interface still clearly exists. There is still an operator in the loop. The following is a diagram (figure 1) provided by NASA's Langley Research Center of a "pilot-in-the-loop" system. This model is useful in defining both the conventional system for aircraft with a pilot on board, as well as unmanned systems (pilot displaced logistically).

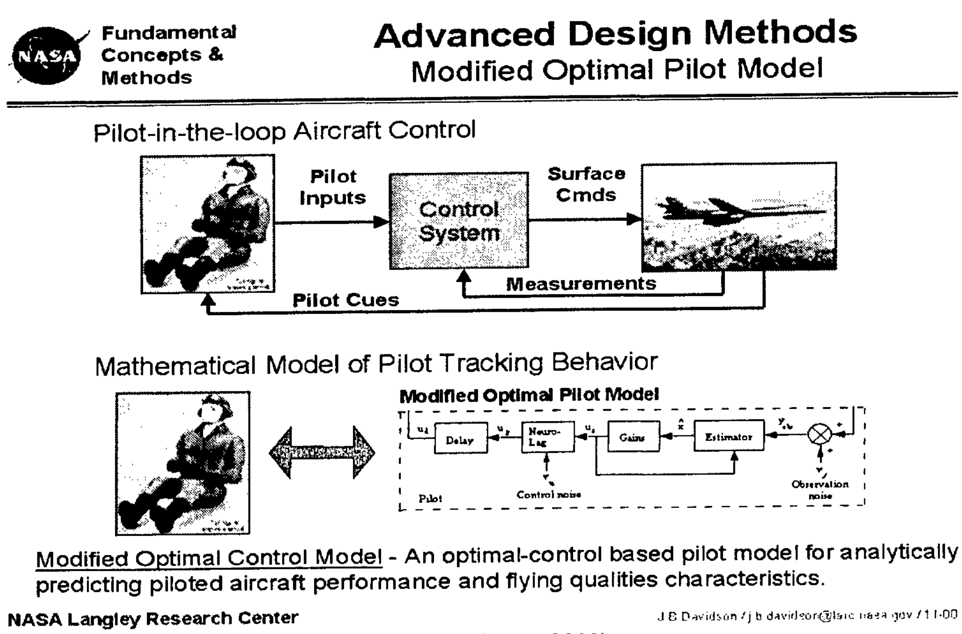


Figure 1. Modified Optimal Pilot Model (Davidson, 2000)

In either case (pilot on board or operator/pilot displaced) there is clear interface between the operator and the system. This case stands for other systems as well. A tank requires a crew as does a military ordnance disposal robot. The crew is simply not in the same location. Unfortunately our new definition of military utility fails to recognize that this interface is critical to system operation and evaluation.

NASA studies have quantified these characteristics in what is referred to as the OCM or Optimal Control Model. Research has supported a mathematical assessment of every step in the process based on measured stimuli (input) and reaction (output). (Davidson, 1992) For this study a dissection of the calculus is not of value, but the relevance of quantifiable data is. In another words, although this study will not present the exhaustive math that the NASA study did, the fact-based and scientifically sound processes clearly exist.

Hawkins describes that before one can react to a given situation information about that situation must have been sensed. Here there lies a first source of potential error. The input may be misunderstood, it may be processed based upon erroneous memory data, it may be incorrectly managed, it may be managed in an untimely fashion, and/or it may be processed correctly. All of these options can be addressed as a variable and quantified. This quantifiable feature would be the source of empirical data that would have constructive impact on system assessment.

Option one, although sponsored by an era of phenomenal weapons performances is unlikely to be flexible enough to fully function in an environment of acquisition reform. Option two has demonstrated considerable improvement, but quantifying data is blurred by the need to trade off performance in one area for performance in another. Option three addresses the critical human element, a factor of the system not previously effectively addressed. The human element is evident in every system in place today, and projected for the future. The data demonstrates that redefining the term to one incorporating human factors is a plausible, fact based solution.

RESULTS

Assuming that the above logic is valid, a new definition must be offered. Considering our initial definition from the Defense Acquisition University as incomplete, verses simply wrong, the following definition is proposed:

"The ability of an operator to perform as planned in or with a designed system maximizing lethality, survivability and maintainability in a competitive environment including versatility (or potential) of the system. It is measured against the operational concept, operational effectiveness, safety, security, and cost/worth." This definition moves the point of measurement from an undefinable term, "military worth," to a definable concept, "operator performing as planned." This concept allows for exploitation of measurement tools and techniques that are too precise to be used in the original definition. This new definition embraces the concept of human factors as a "core technology" as previously addressed, and allows for technical interpretation.

Toolbox Analogy

If all aircraft operate just as designed, and the design is effective, the need for test data would be limited at best. Specifications become the yard stick to insure that the system performs as designed. If a system does not operate as specified, or the specifications themselves are inadequate, then the toolbox of test techniques must be opened. This toolbox contains the techniques and procedures used to determine the extent and corrective action of the problem. Specifications may not cover every aspect of the mission, and they do not handle new technology well; but they do address the quantifiable features that require assessment. (National Test Pilot School, 2001)

The tester must interpret data based on open or closed loop testing principals. Open loop testing is simply putting in inputs and watching. It is loose in design and execution. Closed loop testing addresses specific points. A detailed plan will be followed and data obtained for specific points. Experimental (or Developmental) test is often closed loop to allow for precise control. Operational test is usually open loop to allow for operator assessment. Whichever process is followed, there must be a level of control on the input of measurable data to produce data that is addressable to particular situations.

Dimensional analysis is a tool used by engineers to limit the number of variables when testing a particular problem or phenomena. This facilitates the study of the interrelationships of systems and models of systems. It can also be the source of data in support of modeling to insure the representation is faithful to the original concept.

Parametric analysis allows the interpretation of data by changing one input at a time and evaluating the change to the system. Equations demonstrate the interrelationships of variables, and by controlling all but one variable the system impact of a single change can be easily quantified.

System Analysis

System analysis involves the use of either or both of these techniques (dimensional and parametric) and can easily integrate the human factor issue as a variable. In essence, a system perspective is a way of breaking some selected issue into definable pieces and observing how they interact. (Wiener, 1988) The concern comes now to the problem of measuring the interface between the human and the system. In either process (dimensional or parametric) the outcome can be measured and traced back to the human interface. Value may be additionally found if the human interaction could be measured as part of an entering argument. The B-52 AMI Data Analysis Plan integrated several tools as methods for quantifying human factors data.

Very Unsatisfactory	Unsatisfactory	Marginally Unsatisfactory	Marginally Satisfactory	Satisfactory	Very Satisfactory
1	2	3	4	5	6

Table 1 AFFTC 6-Point System Adequacy Rating Scale (Test and Evaluation Master Plan)

This simple table allows the evaluator to assess a system with certain subjective criteria and extract a numerical value useable to the test and evaluation program. Operator assessment using this and programs such as the Bedford Ten Point Workload Scale and the Cooper Harper Scale provide quantifiable data. These provide additional assessment techniques and have even been used by NASA as part of the Space Shuttle cockpit redesign program (Hilty).

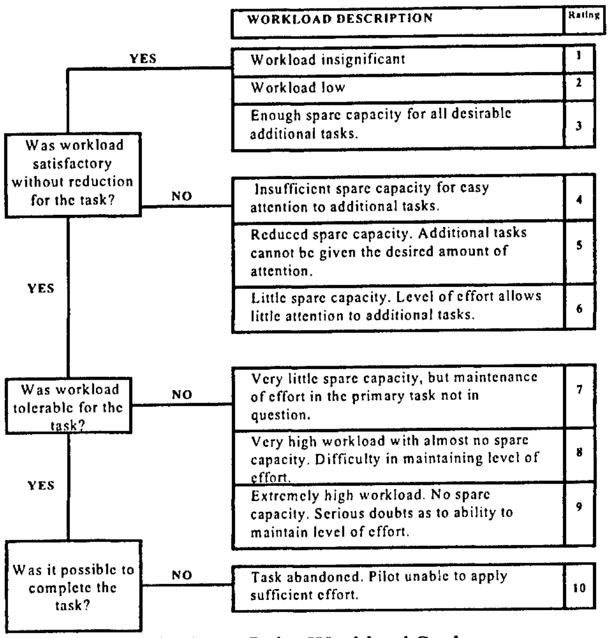


Figure 2. Bedford Ten Point Workload Scale

DISCUSSION

Use of these and other tools to evaluate human performance in various situations provides empirical data that is of greater decision making value that a vague military utility definition might suggest. With various operators "scoring systems" based on the presented criteria a fact-based assessment of the user-system interface becomes evident. (National Test Pilot School, 2001) These rating systems provide the tools needed to empower the new definition integrating human factors as the key concept in military utility. This alignment of the test and evaluation system with current technological support for the human system integration issue makes for the ideal and quantifiable solution to a redefining of military utility.

To restate the proposed definition: "The ability of an operator to perform as planned in or with a designed system maximizing lethality, survivability and maintainability in a competitive environment including versatility (or potential) of the system. It is measured against the operational concept, operational effectiveness, safety, security, and cost/worth." This definition provides the fact based opportunities to assess a projected system and would be essential to a reengineered acquisition process.

The human has always born the brunt of combat. Does it not make sense that the tools employed to test the systems that soldiers, sailors, airmen and marines operate be assessed based on the human interaction? Does it not also follow that the combat and combat support systems fielded must assess their ultimate utility against the ability to serve the tactical needs of the operator?

CONCLUSION

The primary concern of any test or evaluation enterprise in the acquisition process is to produce data. Data production should be focused on the essential elements needed to employ the system in development in its combat or combat support role. Defense Acquisition University (2001) stresses that this evaluation should be based upon key issues, often referred to as Key Performance Parameters (KPP). These KPPs are the capabilities that the system must demonstrate to meet the warfighters needs.

Unfortunately, the past has provided loose guidance that does not support our current fiscally restrained era. The guidance in the use of the term military utility was at best broad and of limited use to the test and acquisition community. A redefining of this term is essential to insure that the warfighter is blessed with the best possible systems that are operable under the stresses of combat. The acquisition community is also faced with the reality that as stewards of the public's money, they must keep an eye on fiscal issues as well as purely functional ones.

The only way to insure this performance is to evaluate systems on their human factor interface. Lethality is nothing except for the integration of the operator. Similar can be said for survivability and maintainability. The key variable is the human not the technology. Integrating the human element in a system based solution is the only logical option to insure the finest in fielded weaponry.

REFERENCES

AFFTCI 99-3, AFFTC Technical Report Program. (2002). Air Force Flight Test Center, Edwards Air Force Base CA.

Braverman, Jerome D. (1981). Fundamentals of Statistical Quality Control. Lawrenceville, NJ: Prentice Hall Inc.

Davidson, John D., and Schmidt, David K. (2000). Modified Optimal Control Pilot Model. Retrieved June 15, 2003 from NASA Langley Research Center: http://dcb.larc.nasa.gov/DCBStaff/jbd/OnePagers/MOCM_Sumry.html.

Davidson, John D., and Schmidt, David K. (1992). Modified Optimal Control Pilot Model for Computer-Aided Design and Analysis, NASA Report Number TM-4384, NASA Langley Research Center, Hampton VA.

Defense Acquisition University, (Courseware). (2001). ACQ-201B, Intermediate Systems Acquisition Course. DAU

Defense Acquisition University, (Courseware). (2002). TST-202, Intermediate Test and Evaluation Course. DAU.

Defense Acquisition University, (Courseware). (2003). TST-301, Executive Test and Evaluation Course Student Guide. DAU.

DoDD 5000.1, The Defense Acquisition System, Chg *1*. (2001). Department of Defense, Washington: GPO.

DoDD 5000.1, The Defense Acquisition System. (2003). Department of Defense, USD (AT&L), Washington: GPO.

DoDD 5000.2, Operation of the Defense Acquisition System. (2003). Department of Defense, USD (AT&L), Washington: GPO.

Farrell, Edward, J. Avionics Engineer. (2003, July 21). (Interview). Edwards Air Force Base, CA.

Federal Aviation Administration Human Factors Team Report, The Interface Between Flightcrews and Modern Flight Deck Systems. (1996). FAA.

Glossary, Defense Acquisition Acronyms and Terms. (2001). Department of Defense, Washington: Defense Acquisition University Press.

Gourley, Scott R. (2003). From Hunters to Killers. Armed Forces Journal, July 2003, 38-40.

Hawkins, Frank H. (1995). Human Factors in Flight, Brookfield VT: Ashgate Publishing Co.

Hilty, Bruce and Lindsey, Steve. (n.d.). Crew Workload/Performance Margin Assessment. Retrieved June 15, 2003 from NASA at http://vision.arc.nasa.gov/AFH/Archives/Space-Shuttle2.pdf.

Jumper, John, General. (2003). USAF, Chief's Sight Picture, Technology-to-Warfighting: Delivering Advantages to Airmen.

National Test Pilot School, (Courseware). (2001). Introduction to Performance and Flying Qualities Flight Testing. Mojave, CA.

National Training Center, 11th Armored Cavalry Regiment Mission Statement. (n.d.). Retrieved 29 June 2003 from Fort Irwin at http://www.irwin.army.mil/11acr/visitors/mission.htm.

Orlady, Harry W., and Orlady, Linda M. (1999). Human Factors in Multi-Crew Flight Operations. Brookfield, VT: Ashgate Publishing Co.

Rajadhyaksha, Dinesh, V. B-52 Armament Engineer. (2003, July 21). (Interview). Edwards Air Force Base, CA.

Test and Evaluation Master Plan, B-52 Avionics Midlife Improvement Program, AMI Human-Systems Interface Data Analysis Plan. (n.d.). Edwrads Air Force Base, CA.

Wiener, Earl L., and Nagel, David C. (1988). Human Factors in Aviation. San Diego, CA: Academic Press.

A HUMAN PERFORMANCE MODEL OF DRIVING GROUND VEHICLES

Josephine Q. Wojciechowski
U.S. Army Research Laboratory

ABSTRACT

The Human Research and Engineering Directorate of the U.S. Army Research Laboratory developed a model of the tasks and workload associated with driving a ground vehicle. The human performance modeling tool, Improved Performance Research Integration Tool (IMPRINT), was used to simulate the driving tasks. Perception, cognition, and motor control were represented in the IMPRINT driving model. Human processing, attention, and response were simulated as concurrent discrete events.

Subsequently, the driving model was incorporated into other IMPRINT models used to investigate crew size and function allocation in Future Combat System (FCS) conceptual ground vehicles. Driving is a primary crew function in FCS ground vehicles. The results of this study indicated that a dedicated driver was required in combat vehicles. In all configurations tested, the driver was consistently the crewmember with the most and greatest workload peaks.

As expected, results of simulation runs were consistent with research on driving and distraction. Structural and output validation of the model was completed through literature review. Driving by itself is a high mental workload function. The human processing capacity is fully engaged in tasks when one is driving, with the primary load being in perception and cognition. Literature shows that performance will start to degrade if additional tasks are attempted during driving, especially if the tasks are highly perceptual or cognitive.

This model provides an efficient means to represent the driving function and can be used for investigating any system where driving is important. For FCS, this will include direct driving and indirect driving. Several additional validation studies are planned.

Keywords: Driving; Task network modeling; Human performance modeling; Mental workload

INTRODUCTION

Driving is a fairly routine function for most of us. As we become experienced drivers, the tasks become "automatic." In today's society, driving has almost become secondary to other tasks. We are eating, talking on the cell phone, navigating, and performing many other tasks while driving.

Likewise, in the U.S. Army, transformation of the force is changing the roles of the Soldiers. With new technologies and force structures, the changing roles of the Soldiers depend on our ability to understand how the Soldier can function in the new roles. To fully understand the mental demand associated with the tasks involved in driving, a task-network model of driving was developed from a human information processing point of view. These driving tasks were subsequently used in conjunction with other tasks performed in a combat vehicle.

A study was completed on function allocation and crew size with this set of combined driving and military tasks. Through the application of this driving model, the criticality of driving in a military vehicle was recognized.

The purpose of this model was to measure the mental demand associated with driving. These driving functions can then be used in conjunction with any other set of tasks for other investigations. To that end, it was important to validate this driving model for that purpose. This paper describes an attempt to validate both the structure and the output of the driving model.

Description of the Model

The model was originally built to represent all aspects of the human information processing model and how they relate to driving (Wojciechowski et al., 2001). A human information processing (HIP) model developed by Wickens and Hollands (2000) is shown in figure 1. The only part of the HIP model that is

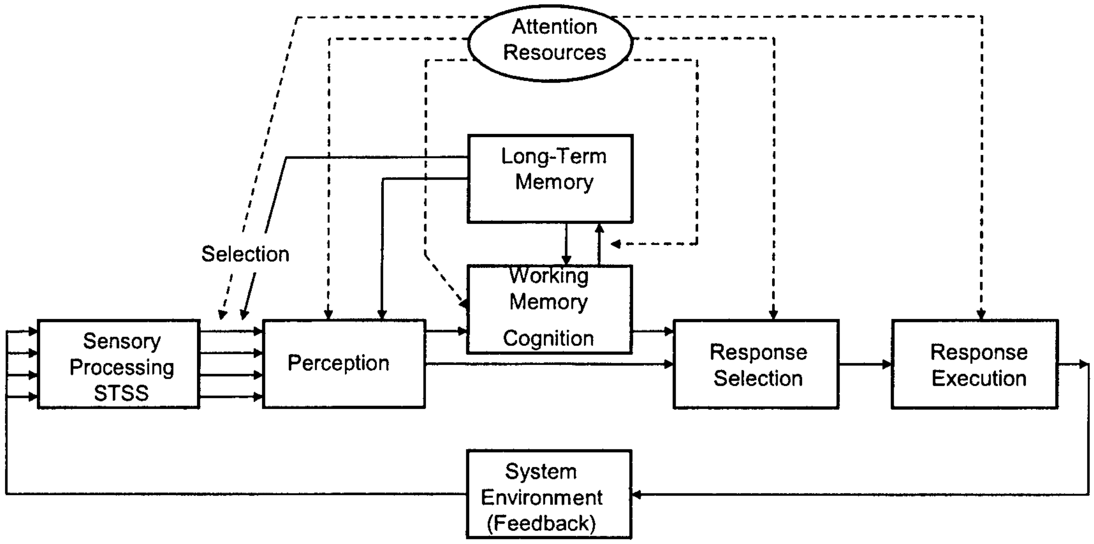


Figure 1. Human Information Processing Model (Wickens and Hollands, 2000).

not included in the driving model is the feedback loop. The driving model uses probabilistic inputs to represent the feedback into the perceptual process.

The model was built with a human performance tool developed by the U.S. Army Research Laboratory (ARL) called Improved Performance Research Integration Tool (IMPRINT) (ARL, 2004). IMPRINT allows the analyst to determine the mental demands of tasks that are programmed to represent the functions of interest (in this case, driving). Human performance algorithms built into IMPRINT will calculate the mental workload of the tasks as they are performed and report the mental demand over time. These mental demands are represented by the "attention resources" in the HIP model.

The driving tasks were grouped into three main functions. These can be described as the psychomotor function, the perceptual and cognitive function, and the kinesthetic and vestibular function. IMPRINT does not measure physical workload so the output of this model is the mental demand associated with the tasks in the functions listed. The tasks included in each of these functions are described.

The psychomotor function represents the "response execution" included in the HIP model (see figure 2). The beginning task in this function is an initial acceleration. A looping branch that includes an acceleration task, a deceleration task, and a coast task follows this initial task. These are continuously looping with a probabilistic determination as to whether the operator will respond by accelerating, decelerating, or maintaining a constant speed. Additionally, after the initial acceleration task, a second loop is entered that alternately has the driver steer or maintain course. The acceleration-deceleration-coast loop and the steering loop run concurrently. The times of these tasks can be varied to represent different terrains.

The perceptual and cognitive tasks, shown in figure 3, represent "sensory processing, perception, working memory, long-term memory, and response selection" in the HIP model. The initial task in this function is "scan sector." A landmark may or may not be perceived (probabilistic) and then a cognitive process is initiated. The process includes three tasks: recognizing the path being traveled, determining the distance to the objective, and comparing that information with what is known. These tasks are performed simultaneously. A decision is then made as to the path and speed to be traveled. This process is then begun again at scan sector. Note that the highly visual tasks of "scan sector" and "perceive reference" include some cognitive demand. Accordingly, the highly cognitive tasks of recognizing the path being traveled, determining the distance to the objective, and comparing that information with what is known include some visual demand. This is how the continuous input of a process such as driving can be represented in a discrete-event simulation tool such as IMPRINT.

The kinesthetic and vestibular function represents the mental demand associated with the physical actions when one is riding in the car and are displayed in figure 4. These tasks also represent "sensory

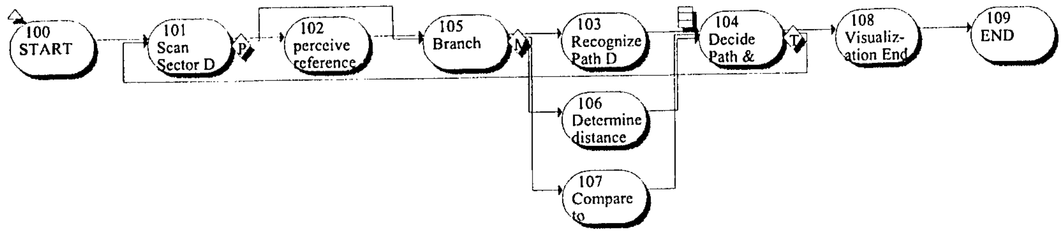


Figure 2. Tasks modeled in the psychomotor function.

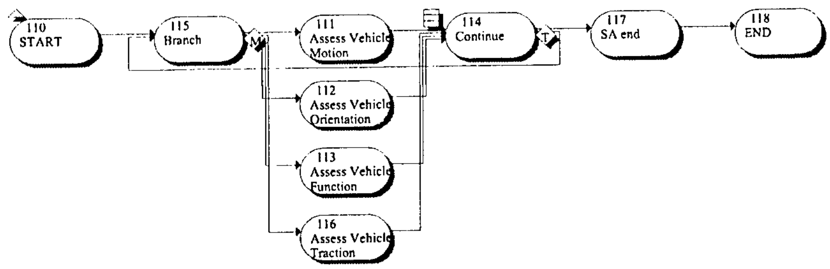


Figure 3. Tasks modeled in the perceptual and cognitive function.

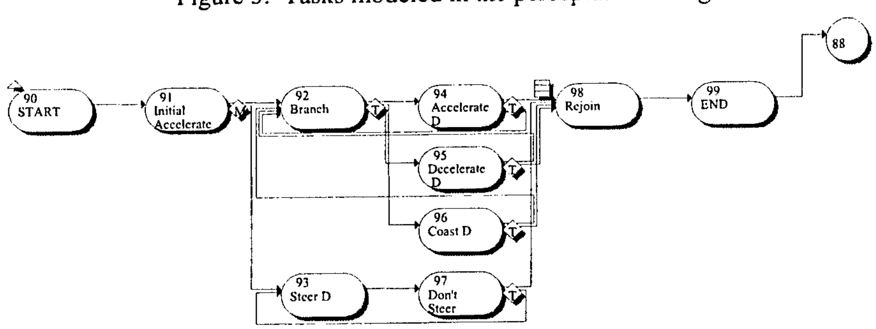


Figure 4. Tasks modeled in the kinesthetic and vestibular function.

processing, perception, working memory, long-term memory, and response selection" in the HIP model. They are assessing the motion, traction, orientation, and function of the vehicle. They, too, loop continuously so that there is a constant mental demand for this function.

An Application of the Model

The transformation of the Army has brought about a desire to reduce crew size and vehicle weight. Along with this, advancing technology has given the perception that each soldier will be capable of performing an increasing number of tasks. Previous work performed at the Human Research and Engineering Directorate of ARL, showed that it would be difficult to reduce crew size in a combat vehicle and maintain satisfactory performance (Mitchell, 2003). As a result, a study was initiated to investigate the mental demand associated with the tasks performed in a combat vehicle. Military functions from the previous work were combined with this driving model to examine allocation of function in a two-person versus three-person combat vehicle (Mitchell et al., 2003).

The tasks were grouped into three primary functions: driving, gunning, and commanding. Four separate IMPRINT models were built. Three of the configurations represented a two-person crew and the fourth configuration was a three-person crew. The first configuration had a commander-driver and a gunner. The second configuration included a gunner-driver and a commander. The third configuration was a commander-gunner and a driver. The three-person configuration had one crewmember performing each of the functions, a commander, a gunner, and a driver.

Results from this study concluded that a three-person crew was necessary for a combat vehicle because of the high mental demand required to perform these tasks. The first configuration showed the highest workload condition of all. The commander-driver was consistently in a high mental demand situation. The second configuration was a little better, but with a gunner-driver, shooting while moving would be impossible. However, this is a necessary survivability tactic. The last two-person configuration was the most desirable. The commander-gunner was able to perform most of his tasks without an excess of mental demand. However, this configuration prevents the hunter-killer philosophy of the commander identifying a target and continuing to scan for others while the gunner fires. The best configuration of all was the three-person crew. This configuration ensures that two crewmembers are available for scanning and it allows the hunter-killer philosophy.

There was interesting result from this study. In all four configurations, the crewmember with the highest mental demand was the one responsible for driving. Even in the three-person configuration, the driver had many instances of high mental workload. As a result of this, a validation study was undertaken to ensure that the driving model was an adequate representation of driving tasks.

Validation of the Model

The validation effort consisted of two parts. As per Army Regulation 5-11 (Department of the Army, 1997 & 1999), structural validation of the model itself was necessary. Additionally, output validation of the model results was required. The structural validation was to be achieved by comparison with other driving models, and output validation was initially to be achieved through comparison to study data.

In order to achieve the structural validation, one must review the assumptions and architecture of the model. Through discussion with colleagues and researchers in the field of driving, several driving models were identified that could be used to compare with this IMPRINT model. It is important to note, however that each of these models was built for a specific purpose. The purpose for each model varied, but the structure of the models included all the same aspects. The purpose of our model was to represent the components of driving in such a way that mental workload and performance, in terms of mission completion and time, could be determined. Our model was actually built to determine the attentional demands that are controlled in Levison's procedural model (1993) described below. We do not represent the feedback loops with the vehicle. The model is a stochastic model that is used to look at the different combinations of driving tasks that may happen concurrently. This provides us with the ability to identify which concurrent tasks will overload a driver's mental demand and therefore identify areas for potential performance degradation. Salvucci, Boer, and Lui (2001) use a cognitive architecture to model driver behavior. They characterized their model in terms of three primary components: control, monitoring, and decision making. The control component accounts for perception of control variables and motor control. The monitoring component accounts for monitoring the environment. The decision-making component is the cognitive process of determining if a lane change is necessary or safe.

In 1993, Levison described a "Driver Performance Model," which has since been used as a basis for other driving models. The processes represented in Levison's model include perception, cognition, control actions, and decision-making. This model is actually two models combined: a driver/vehicle model and a procedural model. The driver/vehicle model is a continuous feedback model between the driver's actions and the vehicle reactions. The procedural model looks at how the driving tasks determine task selection along with simulating the in-vehicle auxiliary tasks. The procedural model represents the regulation of attention. These components are all represented in our driving model.

Brown, Lee, and McGehee (2000) described a driver model for a rear-end collision warnings. The results are a time history of the driver's response in avoiding a rear-end collision. It contains three major components. The first is a representation of the attention to the roadway, based on the uncertainty of the driver. The second component describes the decision process for braking or travel. The third component describes the driver's response.

Biral and Da Lio (2001) surveyed driver models in literature. They suggest that good driver models are required to predict vehicle performance. Their investigation revealed three main types of driver models. First, some models are based on conventional continuous control such as proportional integral derivative (PID) and generalized predictive control (GPC). The second type of driver models that exist are fuzzy logic or neural network based controllers. Fuzzy logic controllers are popular for representing human behavior and neural nets are often used to represent a human's capability to learn. The final class of driver model that the authors found was called hybrid and hierarchical models. These employ the other two types. Of the driver models identified, Biral and Da Lio determined that for models to represent realistic driving behaviors, they must functionally consider the following components: perception, cognition, decision, and motor process of the human.

All the human information processes represented in each of the other models are also represented in our model. The representations are different, but that is expected because the purpose of each of the models is different.

Output validation was the next critical step to validate our model. This type of validation compares the output of the model to the perceived real world. Initially, it was thought that the best means of validating the model was to run a field study and compare the output. Validation by direct comparison of output from the model to field data is complicated primarily because of the difficulty in measuring workload. Most measures of workload are indirect or subjective. Additionally, many previous studies have shown that additional distracters to driving would result in performance errors. The errors can range from lane maintenance to vehicle accidents. These studies could validate our findings.

Strayer, Drews, and Johnson (2003) did a series of experiments that showed that talking on a hands-free cell phone while driving causes what they label "inattentive blindness." The experiments ranged from looking at driving performance errors to determining that drivers do not recall billboards that they fixated on while driving and talking on the cell phone.

Direct Line Motor Insurance (2002) has shown that reaction times for drivers were on average 30% slower when the driver was engaged in a cell phone conversation while driving as compared to when the driver was legally over the limit for alcohol consumption and driving. Furthermore, the reaction times for drivers talking on a mobile phone were 50% slower than when they were only driving.

In the *New England Journal of Medicine*, Redelmeier and Tibshirani (1997) used an epidemiological method to look at the risk of accident attributable to cell phone use. They state that the accident risk quadruples during cell phone use while driving.

Tijerina (2000) reports that predicting costs and benefits of the driver distraction associated with in-vehicle technology is very complex and difficult. However, driver behaviors and operational problems with the technology can be evaluated. There is no doubt that crash data and driver distraction are related. There are, however, so many variables that it is difficult to predict what level of distraction would cause an accident.

The conclusion that can be drawn from these studies is that driving is a high mental demand function. Performance errors are indeed likely with distractions to driving. The output from these studies validate the high mental demand results from the IMPRINT driving model.

Future Work

Additional work is planned to further validate the driving model and further validate the finding that a combat vehicle driver should not be required to perform additional tasks unless driving is fully and reliably automated. Two separate studies are planned. The first study will use the driving tasks from this model to represent teleoperation. The driving tasks will not change but the workload will be different because of the modality and attentional demands of the task. The revised model will then be used in a "model-test-model" approach to predict performance in a study by Hill, Tauson, and Stachowiak (2003) Model predictions of performance and test results will be compared and the model will be adjusted to better represent the actual teleoperation. The model output will then be validated with test results from an additional study by ARL.

The second planned study being considered is a validation of the workload threshold predicted by the model. This study will use an actual vehicle on an outdoor course. The driver will be required to operate the vehicle separately while completing secondary tasks. Secondary tasks will mimic typical tasks that are performed while one is driving both in the civilian world and the military, e.g., talking on the radio, talking to other individuals, looking for hazard indicators. The expectation is that each of these distractions will cause a decrease in performance. This study is still being developed.

This model appears to be an acceptable representation of driving for determining the mental demand associated with driving. The results of the two studies should give further validation and credibility to the model.

REFERENCES

Biral, F. and Da Lio, M. (2001). Modelling Drivers with the Optimal Manoeuvre Method [On-line]. http://www.dinamoto.mecc.unipd.it/articoli%20moto/30articolo.pdf

Brown, T.L., Lee, J.D., and McGehee, D.V. (2000). "An attention-based model of driver performance in rear-end collision situations." Transportation Research Record, Vol. 1724, pp. 14-20.

Department of the Army. (1997). Management of Army Models and Simulations (AR 5-11). Washington, DC.

Department of the Army. (1999). Verification, Validation, and Accreditation of Army Models and Simulations (DA Pam 5-11). Washington, DC.

Direct Line Insurance (2002). The Mobile Phone Report [On-line]. http://info.directline.com/xxx/news.nsf/64125738690474fe00256a6f003a151b/bec9c738833c7fb180256b84002dec5f/$FILE/Mobile%20Phone%20Report.pdf.

Hill, S.G., Tauson, R.A., and Stachowiak, C. (2002). An Evaluation of Human Performance on an Unmanned Ground Vehicle Driving Simulation Task During Operations on the Move (Human Research and Engineering Directorate Protocol ARL 20098-03007). Aberdeen Proving Ground, MD: U.S. Army Research Laboratory. Unpublished manuscript.

Levison, W.H. (1993). "A Simulation Model for Driver's Use of In-Vehicle Information Systems," Transportation Research Record, Vol. 1403. pp. 7-13.

Mitchell, D.K. (2003). Advanced Improved Performance Research Integration Tool (IMPRINT) Vetronics Technology Test Bed Model Development. ARL-TN-0208. Aberdeen Proving Ground, MD: Army Research Laboratory.

Mitchell, D.K., Samms, C., Henthorn, MAJ T., and Wojciechowski, J.Q. (2003). Trade Study: 2- vs. 3-soldier crew for the Mounted Combat System (MCS) and other Future Combat System platforms. ARL-TR-3026. Aberdeen Proving Ground, MD: Army Research Laboratory.

Redelmeier, D.A. MD and Tibshirani, R.J. MD. (1997). Association Between Cellular-Telephone Calls and Motor Vehicle Collisions. The New England Journal of Medicine, 336(7), 453-458.

Salvucci, D.D., Boer, E.R., and Lui, A. (2001). Toward an integrated model of driver behavior in cognitive architecture. Transportation Research Record, Volume 1779, pp. 9-16.

Strayer, D.L, Drews, F.A. and Johnston, W.A. (2003). Cell Phone Induced Failures of Visual Attention During Simulated Driving. Journal of Experimental Psychology: Applied, 9, 23-33.

Tijerina, L. (2000). Issues in the evaluation of driver distraction associated with in-vehicle information and telecommunications systems [On-line]. http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/PDF/3.PDF

U.S. Army Research Laboratory. (2004). IMPRINT Web-site [On-line]. http://www.arl.army.mil/ARL-Directorates/HRED/imb/imprint/imprint.htm

Wickens, C. and Hollands, J. (2000). *Engineering Psychology and Human Performance.* Upper Saddle River, NJ: Prentice Hall.

Wojciechowski, J.Q., Kogler, T.M., and Lockett, J.F. (2001). Examining Direct, Teleoperated, and Semi-Autonomous Off-Road Driving From a Human Information Processing Perspective. Unpublished manuscript.

3-DIMENSIONAL ENHANCEMENTS FOR VISUALIZING LANE NAVIGATION PERFORMANCE

Stephanie A. Myrick and Chad A. Steed
Naval Research Laboratory

ABSTRACT

Research results are typically reported using 2-dimensional (2D) methods that include tables, figures, and charts. With the availability of 3-dimensional (3D) visualization applications, based on the Virtual Reality Modeling Language (VRML) and Extensible 3D (X3D) graphics, the Naval Research Laboratory (NRL) has employed alternative methods of information presentation. These 3D applications are displayed with viewer software on conventional Internet web-browsers and may be effectively used in oral presentations and for separate viewing on the Internet. This paper describes 3D applications that were developed to visualize Marine Corps Amphibious Assault Vehicle (AAV) navigation performances during field demonstrations and augment the 2D performance data. They depict steering patterns used to avoid surface waves, how well the drivers negotiate lane turning points, and a vehicle's vulnerability to mines and other dangers (e.g., subsurface rocks) when steered outside the cleared navigation lanes.

Keywords: 3D Visualization; Amphibious Assault Vehicle; Navigation

INTRODUCTION

Amphibious landing operations conducted in a mined environment require assault lanes that are either cleared of mines or designed to avoid mined areas. Lane width is largely determined by the ability of AAVs to precisely navigate within lanes. Therefore, assault vehicles with more accurate navigation capabilities support reduced lane clearance requirements. To this end, NRL was tasked to develop, test and demonstrate a prototype moving-map system that facilitates lane navigation improvements for AAVs and subsequently report its findings to sponsoring program offices. NRL proposed that a moving-map would improve crew situational awareness and communications, compared with using conventional navigation methods, thereby improving precise lane navigability (Gendron, Myrick, Edwards, & Mang, 2002). Several demonstrations were performed over the past two years, notably Fleet Battle Experiment Juliet in July 2002 and Transparent Hunter in January 2003 (TH03). Comparisons in navigation performance were measured in terms of cross-track error for vehicles using the moving-map system and the same vehicles using no moving-map as they navigated through a designated course (Lohrenz, Edwards, Myrick, Gendron, Trenchard, 2003). NRL has developed 3D visualization applications to enhance its reporting of these demonstration results. These applications are displayed with Cortona VRML Client viewer software on conventional Internet web-browsers (e.g., Netscape and MS Explorer); many other viewers are free and available for download on the Internet. Each visualization depicts a beach and ocean scenario with an animated 3D AAV model navigating through a planned course using actual track data that was recorded as a series of latitude and longitude points.

METHOD

The visualizations were designed to augment 2D data that were collected during TH03 demonstrations. Test runs that could reveal significant navigation issues (e.g., to compare navigation performance using different navigation aids) were selected for 3D visualization. Latitude and longitude coordinates were originally recorded every second during navigation runs. However, since AAVs typically travel at 6 knots or less, these data sets tended to be rather large and subsequently required long application initialization times. With such a high collection rate, it was possible to downsample the original data and still maintain essential visual information. Consequently, every fourth coordinate set was used during downsampling, resulting in an AAV position displayed for every 4 seconds of original run time. Data set sizes were reduced 75% and initialization times were reasonably brief.

3D military models have been developed using X3D graphics at the Naval Post-Graduate School's Scenario Authoring and Visualization for Advanced Graphical Environments (SAVAGE) group and are available through their website. NRL selected the SAVAGE AAV model and modified it to include a windowed driver's hatch (Fig. 1). The SAVAGE group Waypoint Interpolator code was modified and used for AAV animation. The NRL visualization software includes downsampled test run data and modified SAVAGE code to create re-enactments of actual navigation performance during TH03 demonstrations.

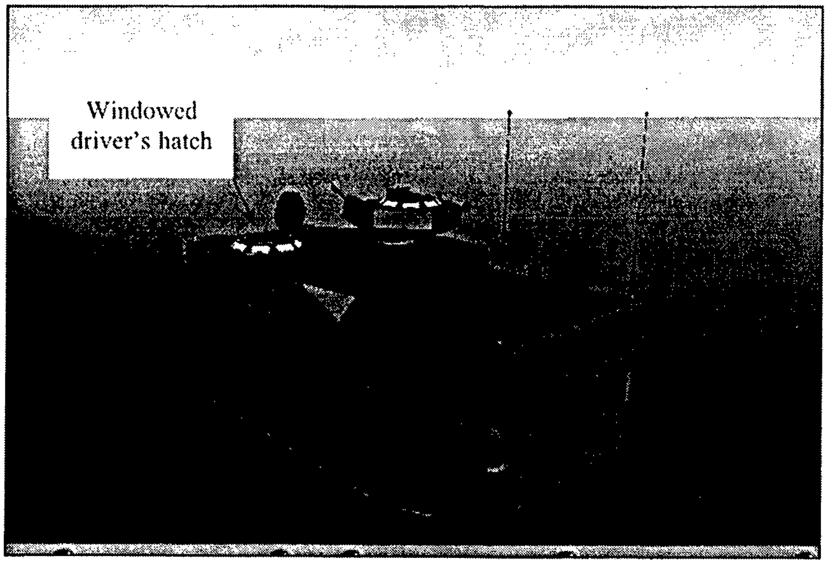


Figure 1. The AAV model modified to include a windowed driver's hatch.

The re-created demonstration area is deliberately depicted with simple beach and ocean regions since the visualizations are intended to focus solely on AAV navigation performance. These regions were created using rectangular objects with texture overlays. During the test runs, AAV drivers were instructed to navigate along a predetermined route; the 3D visualizations include this route drawn in white and the AAV's actual course drawn in red. During animation, downsampled latitude and longitude data are used to depict the AAV traveling on its actual course.

In VRML, viewpoints can be created to provide different perspectives on the scene of interest (Ames, Nadeau, Moreland, 1997). Two different full-scene designs were produced for these visualizations. The default viewpoint is an exocentric perspective view, which gives an impression of looking at the scene from a raised and angled distance (e.g., Fig. 2 and Fig. 3). A second viewpoint looks directly down on the course from above (i.e., plan view, Fig. 4). Two additional viewpoints designed as part of the original AAV model include riding from the rear of the AAV and riding on the front of the AAV.

RESULTS

Navigation runs that illustrate significant navigation problems or interesting observations were selected for visualization. For example, figure 2 depicts a typical back-and-forth steering pattern used by drivers to avoid submersion of the AAV under surface waves and also shows how well this particular driver negotiated lane turning points. The run in figure 2 was navigated with the driver using a moving-map system. AAVs that did not navigate with a moving-map relied instead on a Precision Lightweight Global Positioning System (GPS) Receiver (PLGR), which simply displays vehicle location as latitude and longitude coordinates on the display of a small handheld device. Drivers tended to miss their course waypoints more often with the PLGR (figure 3) than with the moving-map. Missed waypoints always resulted in steering out of the navigation lane, which in a true operational situation, would leave a vehicle perilously vulnerable to mines and other threats. Furthermore, AAV crews were often unaware of their error and misjudged their location and ensuing vulnerability.

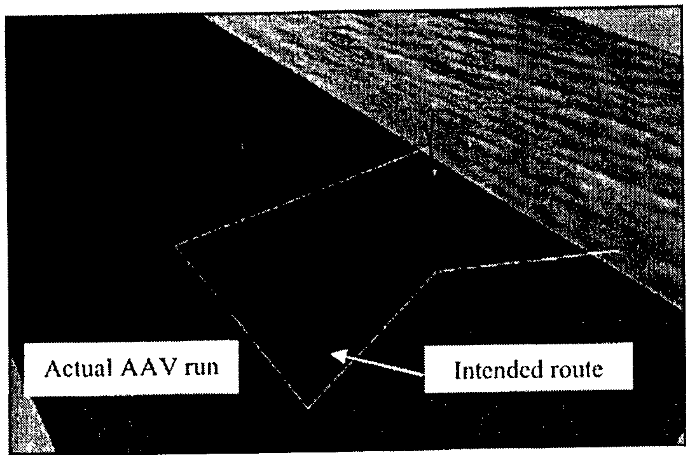


Figure 2. AAV steering patterns used to avoid surface waves.

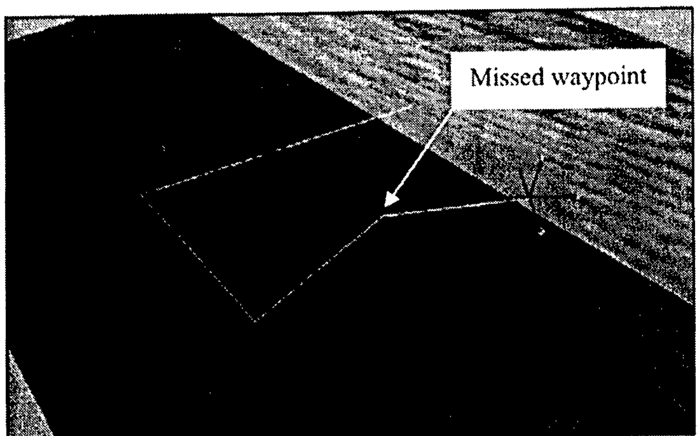


Figure 3. AAV navigation error resulting from a missed waypoint.
Substantial deviation from the designated course leaves the AAV vulnerable to mines.

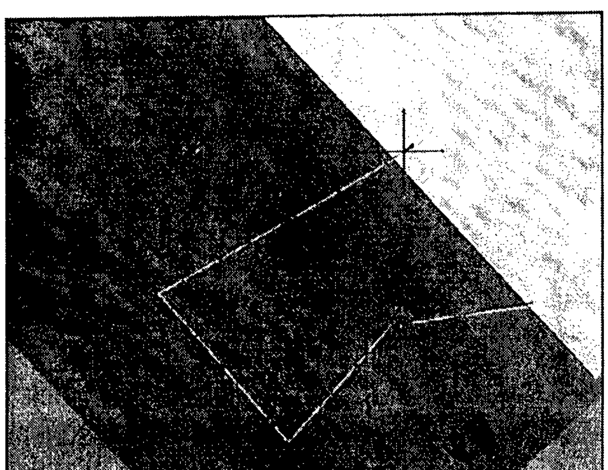


Figure 4. Animation using a plan view perspective.

DISCUSSION

The AAV visualizations depicted in figures 2 and 3 were recently presented at the Oceans '03 Conference (Lohrenz, et al., 2003). Software links were inserted into a PowerPoint presentation to launch the viewer and 3D application at the appropriate time.

The 3D AAV model can be viewed and manipulated separately to convey the physical and visual constraints of the vehicle driver (Fig. 5), or for training and familiarization purposes. For example, the user can rotate the entire vehicle, operate any of its moveable parts (e.g., open the hatches), and even enter the vehicle for viewing from within.

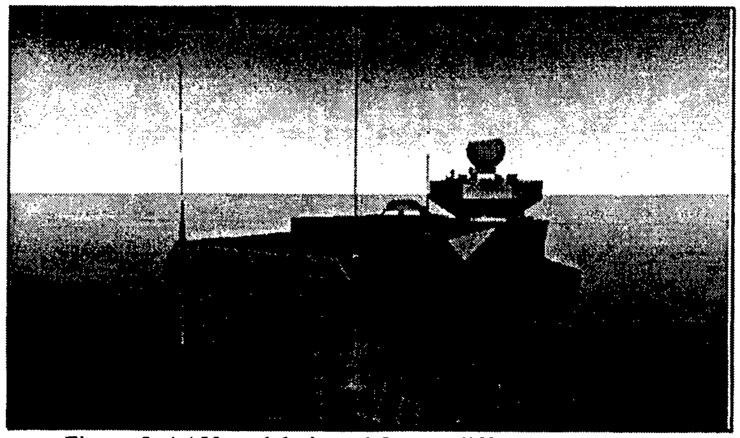


Figure 5. AAV model viewed from a different perspective.

SUMMARY

NRL has developed 3D visualization applications based on VRML and X3D graphics as an alternative means of information presentation. These applications can be displayed with "shareware" viewer software on conventional Internet web-browsers and are equally effectively in oral presentations and in separate on-line viewing via Internet web-browsers. These applications were developed to visualize Marine Corps AAV navigation performances during field demonstrations. They depict steering patterns used to avoid surface waves, how well the drivers negotiate lane turning points, and a vehicle's vulnerability to potential threats when it is steered outside of the cleared navigation lanes.

ACKNOWLEDGEMENTS

The Office of Naval Research (ONR) sponsored this project under program element number 0603782N. We thank Dr. Doug Todoroff (program manager at ONR) and Dr. Dick Root (program manager at NRL) for their support.

REFERENCES

Ames, A.L, Nadeau, D.R, & Moreland, J.L (1997). VRML 2.0 Sourcebook, (2nd ed.). New York: Wiley & Sons, Inc.

Gendron, M.L., Myrick, S.A., Edwards, S.S., & Mang R.A. (2002). Demonstrating a moving-map system with electronic charts for improved lane navigation: testing on AAVs in Gulfport MS. NRL/MR/7440--02-8280. Naval Research Laboratory, Stennis Space Center, MS. Sept. 30.

Lohrenz, M. C., Edwards, S. S., Myrick, S. A., Gendron, M. L., & Trenchard, M. E. (2003). Demonstration of a Moving-Map System for Improved Lane Navigation of Amphibious Vehicles. Proceedings of the Oceans '03 Conference, San Diego, CA

ParallelGraphics Cortona VRML Client v4.1 (Release 90) viewer. Retrieved April10, 2003 from http://www.parallelgraphics.com/

U.S Naval Post Graduate School, Scenario Authoring and Visualization for Advanced Graphical Environments (Savage) Group Amphibious Vehicles: AAV. Retrieved June 12, 2003 from http://web.nps.navy.mil/~brutzman/Savage/AmphibiousVehicles/AAV/_pages/page01.html

U.S Naval Post Graduate School, Scenario Authoring and Visualization for Advanced Graphical Environments (Savage) Group Amphibious Vehicles: Waypoint Interpolator. Retrieved June 12, 2003 from http://web.nps.navy.mil/~brutzman/Savage/Tools/Animation/WaypointInterpolatorExample.wrl

DESIGN AND EVALUATION OF A VISUALIZATION AID FOR STABILITY AND SUPPORT OPERATIONS

Jonathan Pfautz, Kenneth Jones, Greg Zacharias
Charles River Analytics, Inc.

Emilie Roth
Roth Cognitive Engineering

Eva Hudlicka
Psychometrix Associates, Inc.

Ted Fichtl
The Compass Foundation

Bryan Karabaich
Karabaich Associates, Inc.

ABSTRACT

Current and future Joint Task Force stability and support operations (SASO) require intelligence and civic affairs analysis of the attributes of individuals and groups as well as the complex psychosocial and political relationships among these entities. To support analysis in this domain, we have been developing a tool, the Stability and Support Operations Visualization Aid (SASOVA), that combines visualizations (e.g., social network graphs, geo-referenced displays), hyperlinked navigation, and knowledge-based inferencing capabilities to enable analysts to: (1) rapidly profile individuals, groups, and events; (2) assess their inter-relationships; and (3) generate predictions of likely future behavior.

A user evaluation of the SASOVA system was performed using military analysts with extensive SASO experience. Participants utilized the SASOVA system to assess entity characteristics, identify inter-relationships, analyze events, and predict future behavior in a simulated SASO scenario. The results of the evaluation pointed to the value of a multifaceted tool such as SASOVA in increasing speed and accuracy of intelligence analyses. At the same time, the evaluation pointed to the need for additional capabilities to improve observability and traceability of machine agent inferences and assessments, and reduce the potential for fixation effects and premature closure.

Keywords: Decision Aiding, Stability and Support Operations, Intelligence Analysis, Visualization

INTRODUCTION

Current operations in Afghanistan and Iraq point to the increasing need for computerized decision and visualization aids that can support military stability and support operations (SASO) (Cordesman, 2003b; Cordesman, 2003a). The SASO environment is characterized by diverse information requirements, including the need to understand the socio-political climate, the psychosocial characteristics of key individuals, and the causal relationship between groups, constituent individuals, and events. As a result of these multifaceted operational requirements, intelligence analysis in the SASO domain is highly complex and requires careful examination of cognitive demands and critical informational needs to develop effective decision-aids.

We have taken a cyclical approach to developing the Stability and Support Operations Visualization Aid (SASOVA). First, we conducted a cognitive task analysis (CTA) of intelligence analysis in the SASO domain, collecting valuable information on the analyst's decision-support requirements from experienced military personnel. Second, we used this information, with the guidance of two subject matter experts, to develop a computerized visualization aid that integrates a variety of displays and interfaces to support a wide range of intelligence tasks in the SASO domain (e.g., mission planning, execution, re-planning, assessment, review). Third, we conducted a user evaluation of the SASOVA system to identify strengths and weaknesses of our system. Fourth, we are using these results to further drive SASOVA system development and to focus future CTA and evaluation efforts. We summarize these results below because of their potentially broad application to complex decision-making situations beyond the SASO domain.

COGNITIVE ANALYSIS OF THE SASO ENVIRONMENT

We performed a cognitive task analysis to understand the cognitive and collaborative demands that arise in the SASO environment and the kinds of visualization and decision-support elements that could facilitate performance (Potter, Roth, Woods & Elm, 2000). The CTA was based on scenario-guided interviews that were conducted at Ft. Leavenworth, Kansas, with four military personnel with extensive SASO experience, as well as on input and guidance from two Military Intelligence and Psychological Operations experts who served as collaborators on the project.

The scenario-guided interviews were designed to create a concrete SASO context in which the participants could reveal the kinds of information they would seek and factors they would consider in SASO planning tasks. A SASO-specific scenario was developed based on an existing Balkans-like scenario (the TRADOC Kazar scenario). The mission was to enter the country of Kazar to stabilize the local geo-military situation, support the local government in its resumption of sovereign activities, and prepare the region for transition to U.N. control and the next stage in political development. The interviewees were presented specific planning tasks (e.g., developing a plan to find and seize weapons) and were asked to indicate what decisions they would need to make and what information/displays would be useful to support these decisions. The interviewees drew heavily on their own SASO experiences in generating and explaining their decisions.

With respect to decisions/and knowledge requirements -- several major themes arose:

1. In non-traditional operations, established doctrine (both own and enemy) is lacking, placing a premium on rapidly acquiring knowledge of the cultural, group, and individual factors that are likely to influence individual and group behavior.
2. The importance of information gathering and dissemination, and the need to build (bi-directional) communication channels (among U.S. forces, local leaders, non-governmental organizations, other nations participating in operations, and the general populace of host country). They stressed the importance of improved dissemination of information gathered by (and conclusions drawn by) intelligence analysts to the soldiers on the ground that most need it/can best use it.
3. The need to identify emergent patterns suggestive of likely future behavior. They stressed the need to anticipate (and try to dissipate) the next 'flashpoint' or 'hot spot'.
4. The fact that units regularly rotate in and out of positions places a constant need to 'come up to speed' and a premium on methods enabling outgoing units to transfer data to incoming units.

These inputs were used to generate SASO decision-support requirements that served as the basis for design of the SASOVA system. The decision-support requirements included:

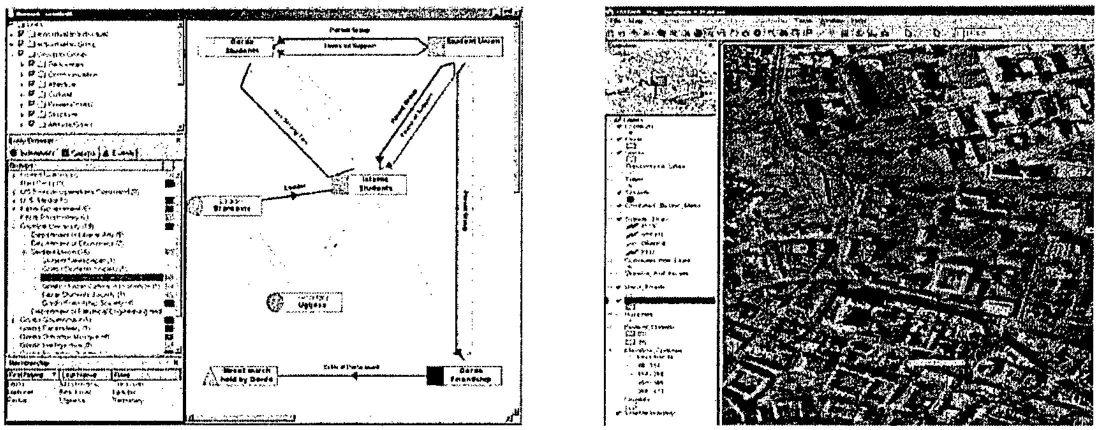
- Support tracking and analyzing the cultural, group, and individual factors affecting individual and group behavior by providing individual, group and event dossiers that collect and integrate information on these entities as well as graphic representations such as social network diagrams that reveal the inter-relationships among entities.
- Provide a repository of intelligence information and 'lessons learned' to enable more effective dissemination of information gathered and conclusions drawn (e.g., by current intelligence; by personnel who held the position previously).
- Provide support for inference and reasoning in data-sparse environments including the ability to infer unknown attribute values from known data; and integrating (social and psychological) theory with known data to maximize individual & group behavior prediction

THE SASOVA SYSTEM

The SASOVA system was developed as an analysis and visualization tool to support Joint Task Force commanders and their intelligence staff in SASO environments. It consists of an integrated suite of capabilities, including: social network displays; dossiers for individuals, groups, and events; geospatial information; history and trend charting; queries and alarms; inferencing tools; hyperlinked navigation among displays; and explicit representations of psychosocial characteristics and relationships. These capabilities are detailed below.

The social network display was designed to facilitate understanding of the relationships among individuals, groups, and events. It allows for the hierarchical exploration of organizations, as well as exploration and selective visualization of a variety of relationships that an individual may have with a group or with other individuals. It explicitly represents source, uncertainty, and inferred information. The social network display also includes an interface for easily creating new entities and links. A screenshot of this display is shown in Figure 1(a).

The geospatial information display includes standard map exploration tools (e.g., pan, zoom), and customizable layers of information. The display is compatible with ESRI's standard geographical information system (GIS) formats to allow for easy importation of data and reduce a user's familiarization time. The interface allows the analyst to draw regions interests and other annotations on the map, and supports 'drill-down'. Information in the map layers can be queried, and map elements can be hyperlinked to the social network and dossier displays. Figure 1(b) shows a screen shot of this display.

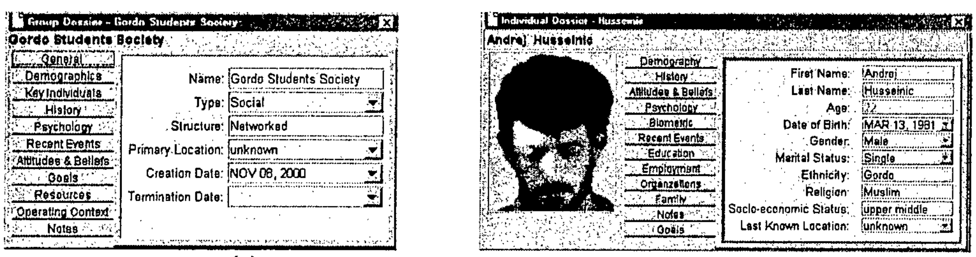


(a) (b)

Figure 1: SASOVA's (a) social network display and (b) geospatial information display

The SASOVA system also includes significant query and alarm capabilities. The user can define specific conditions of interest and query the system's databases, or set an alarm to display an alert when these conditions are met by any of the entities being displayed. The system supports the integration of the results of a query (or the conditions that cause an alarm to be generated) with the existing visualization formats. Alarms and queries can be saved to allow for rapid transfer of case-specific knowledge or general heuristics among analysts.

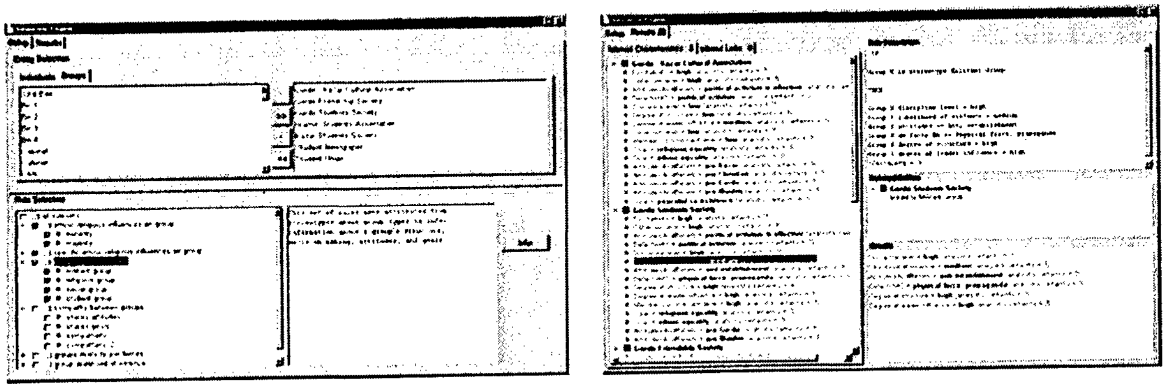
The dossier displays for individual, group, and event attributes were designed to allow the user to seamlessly navigate a large set of hierarchically organized parameters that were identified earlier (Hudlicka, et al. 2002), to edit and/or enter these parameters, and to tie these attributes to specific elements in the GIS and social network displays.



(a) **(b)**

Figure 2: Examples of SASOVA's dossier displays for (a) a group and (b) an individual

Finally, the inferencing capabilities of the SASOVA system provide an interface to a range of profiling tools for individuals and groups, as well as tools for vulnerability assessment and behavior prediction. The inferencing engine supports the use of templates (in the form of specific 'inferencing tasks' that subsume specific knowledge bases), and both expert systems and Bayesian belief network techniques. The inferencing tool supports the creation and editing of rule sets and belief networks within the interface, and presents results hierarchically to support understanding of causal linkages. Figure 3(a) shows the interface for selecting which inferencing task to perform and which entities are of interest. Figure 3(b) shows the display of the results of an inferencing task, with explanatory displays for the rule that fire, the inferred attributes or links, and the associated degree of certainty.



(a) (b)

Figure 3: SASOVA's inference engine's (a) setup interface and (b) results display

EVALUATION OF SASOVA

We conducted an empirical evaluation of the SASOVA system to assess the usability and usefulness of SASOVA as well as to identify opportunities for further improvement. The study employed a work-centered evaluation approach (Roth, Gualtieri, Elm and Potter, 2002; Eggleston, Roth & Scott, 2003) that emphasizes the use of representative scenario tasks that reflect the cognitive and collaborative demands of the domain and collection of both objective performance measures and user qualitative evaluations.

METHOD

Five students at the Naval War College in Newport, Rhode Island participated in the evaluation of a prototype of the SASOVA system. All had military analysis training/experience and included a range of SASO experience (e.g., Somalia, Bosnia, Kosovo). They were presented with a SASO scenario (based on the TRADOC KAZAR scenario) and told to assume they were a newly assigned staff officer. They were asked to use the SASOVA system to respond to a commander's information requests. They were presented a series of questions designed to exercise different features of the SASOVA system (e.g., social network displays, the inferencing tool, the dossiers, query and alerting capabilities). The questions addressed the user's ability to retrieve social/psychological information and draw inferences about individuals and groups (e.g., 'What is Individual X's leadership potential?', 'What is the likelihood of Group Y becoming violent?').

In each case, the participant was asked to utilize the SASOVA system to: (1) generate an answer; (2) explain their answer; (3) indicate their confidence in their answer (using a seven-point scale); and, (4) indicate what other information they would want, if any, to increase their confidence in their answer. We recorded: the user's response, the correctness of that response, the time to respond, and which SASOVA features were used to generate the response. Following the test exercises participant feedback on the SASOVA system and ways it might be improved were elicited via a written feedback questionnaire and a verbal feedback period. Participants were run individually and test sessions lasted approximately three hours.

RESULTS

The test participants' objective performance and their subjective comments (elicited via written questionnaire and verbal debriefing) provided converging evidence that the types of features embodied in the SASOVA prototype would provide useful support. At the same time, they pointed to additional support requirements.

Overall participant performance was good. The average number of correct responses per question was 4.5 (out of a maximum of 5.0) with a range of 2.0 to 5.0, indicating that participants were generally able to answer questions correctly. Questions where performance was less than 100% pointed to opportunities to improve SASOVA features. For example, only 2 of the 5 participants were able to correctly answer the question regarding which individual was 'most well-connected'. This pointed to the need to provide improved features for visualizing 'social-connectedness' and generating social connectedness values.

Interestingly, while mean correct response was high, mean confidence was only moderate (mean of 5.3 on a seven-point scale with a range of 4.5 to 6.0). Participants said that they felt they were giving a 'fast' answer, of the sort they might realistically need to provide in time-critical SASO situations. Their general strategy was to use the inference tool to come up with a response, then use the dossiers and the social network display to crosscheck their answers (or, said that they would do so, given more time).

Their verbal comments during the test scenario helped explain the moderate confidence ratings and pointed to need for improvements to the SASOVA features to increase confidence in the accuracy of quickly generated answers. In particular, participant comments suggested a need for:

- More information to support the numeric source and certainty values provided in the dossiers and social networks (e.g., the exact source of the data – human intelligence, signal intelligence, etc.)
- More information to justify the rules in the inference tool. The participants pointed out that confidence in information in SASOVA was limited by confidence in the previous user who set up the rules. They suggested tagging inferencing tasks (i.e., sets of rules and belief nets) with the author's name and other justification, and linking rules and generalizations to specific events and information that back up the claims (e.g., basis for generalizations such as 'quick to anger').
- An improved ability to follow the reasoning behind the conclusions of the inference tool

Written questionnaire results reinforced the conclusions from the objective performance data. Figure 4 presents the mean usefulness ratings obtained each of the main SASOVA features. Mean usefulness ratings were high for the social network display, (6.4); the dossiers, (6.2); and the querying capability, (6.2). Ratings were more moderate for the alarm feature, (5.5); and the inferencing tool, (5.6).

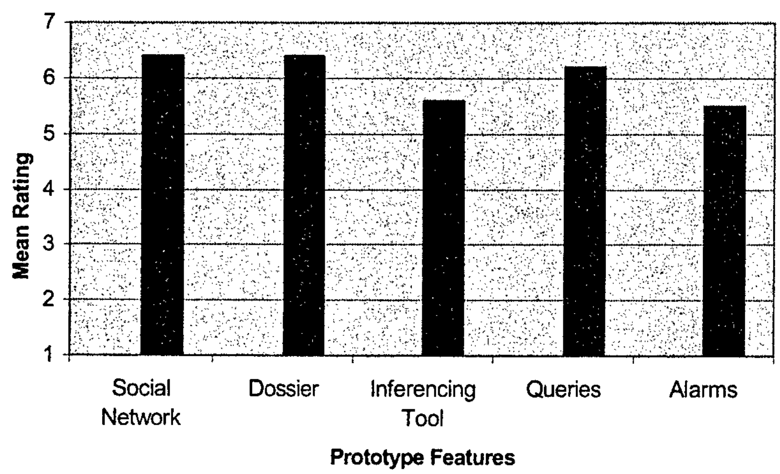


Figure 4: Mean rating of the usefulness of each of the main features of the SASOVA system

To explain the feature ratings, we obtained ratings of cognitive support provided by the SASOVA system and open-ended qualitative assessments. These additional ratings were on a seven-point scale with 1 = 'not at all useful', 4 = 'moderately useful', and 7 = 'extremely useful'. All five participants remarked that the SASOVA system as a whole would provide significant improvement in terms of time and labor over how analyses are done currently. Ratings of the effectiveness of the SASOVA system in supporting different aspects of SASO analysis reinforced this point. SASOVA received a mean effectiveness rating of 6.0 for providing capabilities to explore and connect data, and a mean effectiveness rating of 6.2 for reducing time to perform analysis. Open-ended responses indicated that participants thought that the SASOVA system would be useful for information access and integration tasks for a wide variety of domains beyond SASO operations.

At the same time, participants felt that more information was required to back up the certainty values and the results of the inferencing tool to enable the analyst to evaluate the quality of the information for themselves. This concern was reflected in the ratings of cognitive support. For example, only a moderate rating (mean of 5.0) was obtained for SASOVA's ability to 'broaden set of hypotheses considered' (i.e., prevent fixation, premature closure). Concern with the justification for the rules and certainty values may partly explain why the inferencing tool received only a relatively moderate usefulness rating.

DISCUSSION

The study clearly pointed to the value of a multifaceted tool such as SASOVA in supporting analysis in SASO environments and intelligence analysis more broadly. The accuracy of user's responses on test questions was high, and user feedback indicated that a system such as SASOVA could substantially reduce analysis time. This is especially important for time-critical operations.

At the same time, the results highlighted the importance of supporting analysts in broadening the set of hypotheses considered and preventing premature closure (Patterson, Roth and Woods, 2001). The study pointed to the need for additional capabilities to improve observability and traceability of system inferences and to the need to increase the confidence in the inferences drawn. This includes improving the treatment of source quality and uncertainty; improving the justification for rules and belief nets used in inferencing; and, making it easier to search for, and keep track of, converging and conflicting evidence. Our concerns about preventing premature closure and recommendations for ways to guard against it have general applicability to the design of intelligence analysis support tools.

REFERENCES

Cordesman, A. (2003a). The Instant Lessons of the Iraq War. Center for Strategic & International Studies.

Cordesman, A. (2003b). The Lessons of Afghanistan: Warfighting, Intelligence, Force Transformation, Counterproliferation, and Arms Control. Center for Strategic & International Studies.

Eggleston, R. G., Roth, E. M. and Scott, R. A (2003). A framework for work-centered product evaluation. In Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting. (pp. 503-507). Santa Monica, CA: Human Factors and Ergonomics Society.

Hudlicka, E. Zacharias, G., and Schweitzer, G. (2002). Individual and Group Behavior Determinants: Inventory, Inferencing, and Applications. In Proceedings of the 11th CGF Conference, Orlando, FL, May 2002. (pp.21-28).

Patterson, E. S., Roth, E. M. and Woods, D. D. (2001). Predicting vulnerability in computer-supported inferential analysis under data overload. Cognition, Technology and Work, 3, (pp. 224 – 237).

Potter, S. S., Roth, E. M., Woods, D. D. & Elm, W. (2000). Bootstrapping multiple converging cognitive task analysis techniques for system design. In J. M. Schraagen, S. F. Chipman & V. L. Shalin (Eds.) Cognitive Task Analysis (pp. 317-340). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Roth, E. M., Gualtieri, J. W., Elm, W. C., and Potter, S. S. (2002). Scenario Development for Decision Support System Evaluation. In Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting. (pp. 357-361). Santa Monica, CA: Human Factors and Ergonomics Society.

WHAT PERSONALITY CHARACTERISTICS DO DIGITALLY COMPETENT SOLDIERS HAVE IN COMMON?

Brooke Schaab, J. Douglas Dressel
U.S. Army Research Institute for the Behavioral and Social Sciences

ABSTRACT

Soldiers who elected and qualified for military a occupation that emphasizes digital technology were administered the Self-Directed Search, a measure used to study the overlap between interests and personality. Overall, participants indicated a preference for occupation that were high in the categories of Investigative and Realistic and low in Conventional.

Keywords: Personality; Vocational preferences

INTRODUCTION

Training Soldiers to operate complex digital systems is time consuming and costly. Therefore, understanding characteristics of Soldiers who succeed in this environment has important implications for both selection and training. The purpose of this research is to determine what types of similarities are present among Soldiers who elect and qualify a military occupation that emphasize digital technology.

Measure of Vocational Interests

The Self-Directed Search (SDS) is a self-administered assessment originally designed to provide vocational counseling based on self-reported competencies, abilities, and preferences (Holland, 1985). More recently, this measure has been used to study the overlap between interests and personality.

Holland's typology created in 1973 includes assumptions that people can be categorized into one of six personality types and that people will seek vocations where they can apply their skills and abilities. For example, investigative type personalities seek jobs requiring mathematical and scientific ability because they are inclined to be analytical, curious, and rational.

METHOD

One hundred twenty-seven entry-level Soldiers in training to operate one of the Army's most advanced digital systems participated in this research. Soldiers were administered a paper-and-pencil version of the Self-Directed Search at the beginning of their training. Additionally, they completed a questionnaire where they indicated their preferred high school academics.

RESULTS

Overall, participants indicated a preference for occupations that were high in the categories of Investigative and Realistic and low in Conventional (see Table 1). The *Dictionary of Holland Occupational* Codes defines these occupational classifications as follows:

> *Investigative*-"tend to involve analytical or intellectual activity aimed at problem-solving, trouble-shooting, or the creation and use of knowledge."
> *Realistic*-"tend to involve concrete and practical activities involving machines, tools, or materials"
> *Conventional*- "typically involve working with things, numbers, or machines in an orderly way to meet the regular and predictable needs of an organization or to meet specified standards." (p. 6)

Table 1. Frequency of SDS categories selected

	First Choice	Second Choice	Third Choice	Total
Investigative	39	26	25	90
Realistic	41	19	20	80
Artistic	20	25	26	71
Social	12	29	22	63
Enterprising	10	19	22	51
Conventional	5	8	9	22

As reported above, Investigative and Realistic were the first choice types selected by 63% of these Soldiers. The Chi-square statistic indicates that the number of Soldiers selecting these two categories is significantly different that expected by chance ($X = 27.712$, $\underline{p} < .05$).

Interestingly, there was a significant correlation ($r = .352$, $p < .05$) between Soldiers who chose Investigative as their first choice category and scores on the more difficult items on the end of course test (as determined by subject matter experts ratings and item difficulty calculations).

These Soldiers reported that they enjoyed and had received their best grades in mathematics and technology courses. This supports the vocational categories that were most frequently chosen, Investigative and Realistic. Social Studies and English were the courses that these Soldiers enjoyed the least and where they had received their lowest grades.

CONCLUSIONS

In summary, Soldiers who select and are admitted into an Army occupation that integrates complex digital technology into the job tend to be high on the characteristics of Investigative and Realistic and low on Conventionalism as measured by the Self-Directed Search. There is a tendency for these Soldiers to like and to have received higher grades in Mathematics and Technology courses in the past.

These preliminary findings suggest that it may be possible to use vocational inventories, such as the SDS, to assist Soldiers in selecting their military occupations.

Personality Theory and Human Factors Research

Mark E. Koltko-Rivera
Professional Services Group, Inc.

ABSTRACT:

This paper reflects on issues raised in Schaab's (2004) presentation concerning personality characteristics of the cyber-competent. Schaab's findings raise the possibility that personality traits affect cyber-competence, an insight that is certainly congruent with everyday experience, where personality is seen as affecting human performance in many ways. To apply personality theory to human factors domains, researchers have available to them a variety of theoretical frameworks to study traits (including factorial and circumplex models) and motives (including specific motive and motivational structure theories), for all of which operationalizations are available. There is also a pressing need to develop a set of scales to assess attitudes towards high technology. Human factors researchers should use these theoretical frameworks and operationalizations to study how personality moderates human interaction with the products of high technology (e.g., computers, robots, software agents); this would be the first step in learning how to enhance the cyber-competence of all people.

Keywords: Personality, motivation, human factors, digital competence, cyber-competence, cyber-performance, attitudes toward technology

I have been asked to respond to issues raised by the paper presented by Dr. Brooke Schaab (2004). I concentrate on why and how human factors research should focus on issues addressed by personality theory.

Dr. Schaab administered the Self-Directed Search (SDS; Holland, Powell, & Fritzsche, 1997) to 127 U.S. Army soldiers who had been selected to be trained as Army military analysts; these soldiers were to be trained to work with the Army's most advanced digital systems. The SDS is based on Holland's (1997) model of vocational personalities and work environments; this model posits six vocational personality dimensions, corresponding to six work environment dimensions (Realistic, Investigative, Artistic, Social, Enterprising, and Conventional); the theory proposes that a person with a given personality configuration would perform best in a job with a congruent work environment configuration. In her research, Dr. Schaab found that the overwhelming majority (98%) of the Army analysts-in-training had personality configurations that loaded highly on one or both of the Investigative or Realistic dimensions.

These results are at least compatible with the notion that digital competence (i.e., competence in working within a highly computerized environment) is not equally distributed across personality types; rather, some personality types are simply more digitally competent than others. Such a finding, if replicated, would have profound consequences for human factors theory, research, and practice.

The "Why" of Applying Personality Theory to Human Factors Research

Given the potential consequences, I find it interesting that Dr. Schaab's research was the only report presented at the HPSAA II conference that placed its primary focus upon the influence of personality on a human factors variable. It would appear that human factors research is still guided predominantly by the position of Fitts, who suggested over 40 years ago that personality is of little importance to human factors scientists and practitioners (Fitts, 1963, p. 924)

However dominant this position is in human factors research and practice, it is wildly incongruent with our experience of everyday life in the real world, where we all know that personality affects performance. This is one reason why we assign some kinds of work to some people and not to others. Of course training and experience play a great part in moderating performance, but personality is an important moderator as well.

In the spirit of recognizing this issue, I would suggest that we extend the question raised by the title of Schaab's (2004) paper. Limiting myself to the domain of digital competence (or cyber-competence, as I think it is better designated), I suggest that two fruitful questions for human factors researchers to consider are the following:

- What personality characteristics are typical of more and less cyber-competent people? (I.e., how does personality moderate cyber-competence and cyber-performance?)
- How can we compensate for the personality characteristics of less cyber-competent people?

These are not small issues. Within military contexts, the move to network-centric warfare (Galster & Bolia, 2004) suggests that cyber-competence will be important to attain military objectives. Within civilian contexts, all indications suggest that cyber-competence is becoming increasingly important in successfully negotiating both the demands of everyday life and the demands of many work environments. Consequently, an understanding of how personality moderates cyber-competence and cyber-performance is important for enhancing human performance in many contexts. So, how might such research be pursued?

The "How" of Applying Personality Theory to Human Factors Research

Kurt Lewin noted that there is nothing so practical as a good theory. Human factors scientists have several choices when it comes to applying personality theory to the human factors research milieu. Personality theories and variables may be considered as falling into four classes: traits, motives, cognitions, and social context (Winter & Barenbaum, 1999). I will focus here on traits and motives. (My colleagues and I have dealt elsewhere with the issue how the effect of cognitions and social context on human factors variables can be approached, when we describe how theories of worldview and acculturation may be applied to human factors research; Koltko-Rivera, Ganey, Hancock, & Dalton, 2004. Concerning worldview, see also Koltko-Rivera, 2004.)

Trait Approaches to Personality

Trait theories construe personality as a collection or profile of dimensions or traits. These traits are often conceived in bipolar terms (e.g., optimism vs. pessimism). Two major classes of models of traits are *factorial models* and *circumplex models*.

Factorial models consider personality traits to be collected into larger factors. Probably the most popular factorial model currently is the Five Factor model of personality (McCrae & Costa, 1999), which collects dozens of personality traits into five overarching supertraits, which can be recalled by the acronym OCEAN: *O*penness to experience (vs. closedness to new things), *C*onscientiousness (vs. tendency to disorder), *E*xtraversion (vs. introversion), *A*greeableness (versus disagreeableness), and *N*euroticism (vs. mental healthiness). The five-factor approach to personality traits has a long history in personality research, and the five-factor structure seems to be replicable across many cultural contexts (John & Srivastava, 1999). The Revised NEO Personality Inventory (NEO PI-R; Costa & McCrae, 1992; Piedmont, 1998) offers one operationalization of the five-factor theory of personality, and has been used in many research projects. In addition, many instruments are available to assess individual traits or small groups of traits (e.g., Zuckerman & Lubin, 1985). (Of course, there are many instruments to assess psychopathology, which may be considered a superfacet of the Neuroticism factor of personality. For sake of brevity, I will mention only one, which addresses multiple aspects of psychopathology: the Personality Assessment Inventory; Morey, 1991, 2003.)

Circumplex models consider personality traits to be distributed along one or more circular spectra, like a color wheel. On such a circular spectrum, or circumplex, some traits appear close together (e.g., "sarcastic" and "rebellious") while others appear on opposite sides of the circumplex (e.g., "arrogant" and "deferential"). Many circumplex models are possible, depending on the type of traits being studied (e.g., interpersonal traits, psychopathological traits, etc.); a variety of instruments are available to operationalize these constructs (see multiple papers in Plutchik & Conte, 1997).

Motivational Approaches to Personality

Theories of motivation tend towards two types. One we may call the *specific motive* theories, while the other we may consider as *motivational structure* theories.

Specific motive theories focus on specific motives or lists of motives. For example, research has focused on the need for achievement (McClelland, Atkinson, Clark, & Lowell, 1976) and the need for power (McClelland & Burnham, 1976).

Motivational structure theories focus on personality structures that have motivational consequences. For example, Maslow posited a hierarchy of motivations that must be addressed in a specific order, ranging from safety and security through self-actualization and self-transcendence (Maslow, 1969, 1970). The famous developmental sequence derived from psychoanalytic theory is also a motivational theory (defining oral, anal, phallic, and genital needs; Freud, 1940/1969, Chap. 3). An analytical psychology model, Jung's (1921/1971) theory of psychological types, may be construed as a model of motivation: extraverts are motivated to seek stimulation from the external world, introverts from the internal world; sensing types are motivated to seek data for decisions from the sensory world, while intuitive types are motivated to seek data for decisions from the world of intuitions; thinking types then are motivated to make decisions on the basis of linear logic, feeling types on the basis of emotional logic. The Multitheory Personality Assessment Instrument (Koltko-Rivera & Torres, 2004) provides operationalizations for these three models, the Maslovian, Freudian, and Jungian.

Concluding Remarks

When these remarks were shared at the HPSAA II conference, Dr. Christina Frederick-Recascino noted the following:

- The relationship of personality trait and motivation to performance may not be direct, but rather may be mediated by attitudes.

- There is a distinct need to educate human factors professionals in how to apply personality theory to human factors research and practice.

In relation to the first point, it is nothing short of scandalous that, at this late date, we have not developed a general purpose scale regarding attitudes towards higher technology. Anecdotal evidence suggests that there is a great deal of variation in these attitudes; although many people (including, I suspect, most people who inhabit desks near human factors scientists) have a positive and accepting attitude to high technology, many other people regard high technology with suspicion and even fear. Doubtless these attitudes (which may have trait and motivational underpinnings) affect human-computer interaction, and human interaction with any of the products of high technology.

In relation to the second point, this article and others (e.g., Ganey, Koltko-Rivera, Murphy, Hancock, & Dalton, 2004; Koltko-Rivera, Ganey, Hancock, & Dalton, 2004; Koltko-Rivera, Hancock, Ganey, & Dalton, 2004) are an attempt to educate human factors professionals about the need to consider personality theory (as well as theory regarding affect, worldview, and acculturation) in research and practice. This is an area that will only serve to enrich human factors research and practice.

Acknowledgement

The views expressed in this work are those of the author and do not necessarily reflect official Army policy. This research was supported in part by the Office of the Secretary of Defense, Grant DAMD17-02-C-0016 (Principal Investigator: Mark E. Koltko-Rivera), and by the U.S. Department of Defense Multidisciplinary University Research Initiative (MURI) program administered by the Army Research Office, Grant DAAD19-01-1-0621 (Principal Investigator: Peter A. Hancock).

References

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual.* Psychological Assessment Resources: Lutz, FL.

Fitts, P. M. (1963). Engineering psychology. In S. Koch (Ed.), *Psychology: A study of a science: Vol. 5* (pp. 908-933). New York: McGraw-Hill.

Freud, S. (1969). *An outline of psycho-analysis.* New York: Norton. (Original work published 1940)

Galster, S. M., & Bolia, R. S. (2004, March). *Decision quality and mission effectiveness in a simulated command and control environment.* Paper presented at the Second Conference on Human Performance, Situation Awareness and Automation (HPSAA II), Daytona Beach, Florida.

Ganey, H. C. N., Koltko-Rivera, M. E., Murphy, L. M., Hancock, P. A., & Dalton, J. (2004, March). *Affect as an aspect of workload: Why and how to address affect in automation and simulation studies.* Poster presented at the Second Conference on Human Performance, Situation Awareness and Automation (HPSAA II), Daytona Beach, Florida.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Lutz, FL: Psychological Assessment Resources, Inc.

Holland, J. L., Powell, A. B., & Fritzsche, B. A. (1997). *Self-Directed Search (SDS) professional user's guide: 1994 edition.* Lutz, FL: Psychological Assessment Resources, Inc.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102-138). New York: Guilford.

Jung, C. G. (1971). *The collected works of C. G. Jung. Vol. 6: Psychological types* (R. F. C. Hull, Trans.). New York: Princeton University Press. (Original work published 1921)

Koltko-Rivera, M. E. (2004). The psychology of worldviews. *Review of General Psychology, 8,* 3-58.

Koltko-Rivera, M. E., Ganey, H. C. N., Hancock, P. A., & Dalton, J. (2004). *Worldview and acculturation as aspects of workload: Why and how to address these factors in automation and simulation studies.* Paper proposed for conference presentation.

Koltko-Rivera, M. E., Hancock, P. A., Ganey, H. C. N., & Dalton, J. (2004). *An invitation to the dark side: Human factors professionals encounter personality, motivation, affect, worldview, and acculturation psychology.* Manuscript in preparation.

Koltko-Rivera, M. E., & Torres, N. (2004). *The Multitheory Personality Assessment Instrument—Freudian, Jungian, and Maslovian Scales (MPAI-FJM).* Manuscript in preparation.

Maslow, A. H. (1969). The farther reaches of human nature. *Journal of Transpersonal Psychology, 1,* 1-9.

Maslow, A. H. (1970). *Motivation and personality* (2nd ed.). New York: Harper & Row.

McClelland, D. C., Atkinson, J. W., Clark, R. W., & Lowell, E. L. (1976). *The achievement motive* (2nd ed.). New York: Irvington.

McClelland, D. C., & Burnham, D. (1976, March-April). Power is the great motivator. *Harvard Business Review*, pp. 100-111.

McCrae, R. R., & Costa, P. T., Jr. (1999). A five-factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 139-153). New York: Guilford.

Morey, L. C. (1991). *The Personality Assessment Inventory: Professional manual.* Lutz, FL: Psychological Assessment Resources.

Morey, L. C. (2003). *Essentials of PAI assessment.* n.p.: Wiley.

Piedmont, R. L. (1998). *The Revised NEO Personality Inventory: Clinical and research applications.* New York: Kluwer Academic/Plenum.

Plutchik, R., & Conte, H. R. (Eds.). (1997). *Circumplex models of personality and emotions.* Washington, DC: American Psychological Association.

Schaab, B. (2004, March). *What personality characteristics do digitally competent soldiers have in common?* Paper presented at the Second Conference on Human Performance, Situation Awareness and Automation (HPSAA II), Daytona Beach, Florida.

Winter, D. G., & Barenbaum, N. B. (1999). History of modern personality theory and research. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 3-27). New York: Guilford.

Zuckerman, M., & Lubin, B. (1985). *Manual for the MAACL-R: The Multiple Affect Adjective Check List—Revised (MAACL-R).* San Diego, CA: Educational and Industrial Testing Service (EdITS).

INITIAL EXPERIMENTS ON LEADER PRESENCE AND COMMUNICATION MODE ON COMBAT PERFORMANCE

Kip Smith
Linköping Institute of Technology

ABSTRACT

The US Army and NATO forces are in the process of shifting from the traditional in-situ mode of command and control between soldiers and their leaders to a distributed mode of command and control. As part of this shift, a fire-unit's leader may no longer be part of the unit on the battlefield. Rather, the leader may sit at a relatively remote location and use a variety of electronic media to communicate with the team. In the experiments discussed here, we are starting to address the impacts of remote command and control and communication mode in a series of ecologically realistic simulations of a battlefield environment. We have found that participants follow orders more quickly in the leader-present condition. This result suggests that some kind of intervention will be required if soldier performance is to be as efficient in remote command and control as it is in the more traditional, leader-present, mode of control.

Keywords: Remote command and control, Leader presence, Mode of communication, Combat

INTRODUCTION

The practice of having soldiers on the battlefield receive orders from afar through electronic means of communication is known as remote command and control. A reliance on remote command and control is one of the cornerstones of the US Army's plan for modernizing the dismounted infantry. The soldiers who will operate under this plan are (currently) known as the Future Force. With the advent of the Future Force concept, soldiers may no longer take their battle commands from a leader standing within visual range. Instead, the only connection with their commanding officers may be their radios and other portable information devices.

Previous research has shown that varying the physical proximity of an authority figure affects a person's compliance with a command. In the classic study by Milgram (1974), a research participant was far more likely to administer electric shocks to another person at the researcher's command if the researcher was present. If the researcher gave an order to punish an individual from a separate room via telephone, the participant was three times less likely to comply with the command than if the researcher were in the room giving the command. Accordingly, it is reasonable to hypothesize that a change from leaders who are present on the battlefield to leaders who give orders from a distance is likely to have an adverse impact on soldier performance. The study discussed here investigates the effect of leader presence at two levels (present vs. remote) on soldiers' response to commands to move and to shoot. We anticipate that remote command and control will degrade a leader's ability to exercise authority. We expect this degradation in perceived authority will be reflected in slower reaction times and higher levels of psychophysiologic stress when commands are given remotely and when given over a radio than when they are given face-to-face. If this is found to be the case, it will be necessary to design interfaces and training regimes to insure that this degradation of authority can be mitigated.

METHOD

In the set of three experiments presented here, we have modified the Milgram task to make it palatable to institutional review boards and to give it sufficient ecological validity to generalize to a military setting. The technology that enables this simultaneous ethical sanitization and realism is called Paintball.

The first two experiments focused exclusively on behavioral measures and on the effect of leader presence (Pangburn, Freund, Pangburn, & Smith, 2003). Pangburn et al. document the utility of the paintball assault lane as an experimental platform for studying performance under live fire. The third study is in progress. It builds upon the first two to assess the potential for an interaction between leader presence and communication mode. It augments behavioral measures with analyses of two psychophysiologic indicators of stress - heart-rate and heart rate variability.

This section describes elements of the experimental method shared by all three experiments. Each experiment and its results are discussed separately.

Design, Measures, and Task

The simulated combat environment used in this study is a paintball assault lane, Figure 1. Participants advanced through the lane one at a time. The lane consisted of eight protective stations behind which the participant could hide. At the end of the lane was a fortified position where a sniper was positioned. The sniper's task was to shoot the participant moving up the lane. The participant had two tasks. The first was to move from station to station up the lane in response to the command to move. The second was to shoot targets in response to the command to shoot.

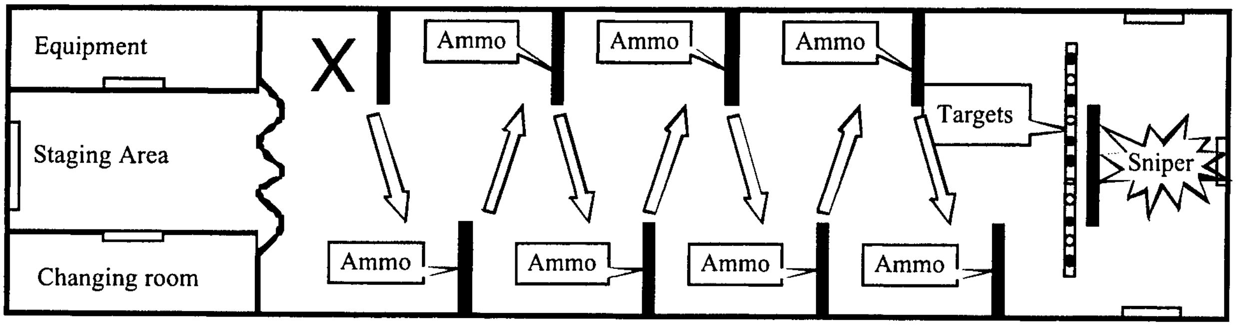


Figure 1 - The layout of the paintball assault lanes.

In all three experiments we manipulated leader presence at two levels (present and remote) in a within-subjects design. In the leader-present condition, the leader was one station behind the participant and communicated by yelling. In the leader-remote condition the only contact between the leader and the participant was by two-way radio.

We used a repeated-measures design with the order of conditions counterbalanced across participants. This design provides the statistical power needed to assess the effect of leader presence and mode of communication on the time it takes participants to respond to commands to move and to shoot.

We measured the participant's response time to the leader's commands to move and to shoot. We predict slower response times in the leader-remote condition but have no a priori hypotheses regarding the effect of communication mode. Statistical analysis used ANOVA to test for sequence effects and within-subjects t tests of the mean differences in response times.

Materials

Participants and the sniper were given one paintball marker (gun), fatigues (overalls), a set of elbow and kneepads, and a paintball face shield. In the leader-remote and present-radio conditions, participants were also given a two-way radio. In the first two experiments, response times to commands to move and shoot were recorded by an observer using a stopwatch. Procedure

Upon arriving at our lane the participants met the leader for the first time. The leader was an army officer wearing a standard battle-dress uniform. The leader briefed participants using the official military Operations Order format and addressed them by their last names. After signing informed consent and liability release forms, participants were told to assemble in a staging area where they could hear the activity in the assault lane while they waited their turns. While waiting, participants were instructed on the safety and use of the paintball markers and read a briefing. All of this was purposefully done to immerse the participant in the experiment and to heighten the sense of realism and their anxiety.

Participants were sent down the lane individually. Whenever the participants took aim at the targets or moved between stations, they exposed themselves to the sniper's fire. Participants were instructed to attempt to shoot enemy targets without hitting friendly targets. No measures were made of firing accuracy, however, because our hypothesis concerns the participant's reaction time to commands given by the leader. The shooting task was created only to give focus to the participant's activity and to give the experiment the feel of a combat environment. (Post-experimental conversations suggest that shooting accuracy was strongly correlated with hunting experience.)

A small container with five paintballs was placed at each of the eight stations. The 40 paintballs in the eight containers were the participant's only ammunition. The participant started at one end of the lane, shown by the

X in Figure 1. At this station and all subsequent stations, the leader gave the participant the command "Fire" when it appeared to be safe to do so. The time elapsed from the issue of the command to the first shot fired is the first dependent measure.

When the participants ran out of paintballs, they reported "Out of ammo" to the leader, who then gave the command "Move" when it appeared to be safe to do so. The participants had to move across the lane to the next station and immediately pick up its container of five paintballs. The time elapsed between the issue of the command to move and the time the participant's hand first touched the new supply of ammunition was the second dependent measure. When the participants finished loading, they reported "Loaded" to the leader who then started the cycle over again with the command "Fire."

The study was intended to generate some anxiety so that the measures would more readily generalize to the battlefield. The major sources of stress were the fear of being shot and actually being hit by paintballs. The pain associated with being struck by a paintball is slight but real. Protective gear minimized the risk of injury.

EXPERIMENT 1

Location and Participants

For the first experiment, the US Army provided access to the 25 ft x 200 ft building at Range 52, Fort Riley, Kansas, home of the US Army 1st Infantry and an active training center for artillery. We set up our paintball assault lane in this building.

Twenty volunteers from Kansas State University (18 men, 2 women; median age 19, range 18 to 28) participated in the first experiment. Attendance was limited because a one-way trip to Fort Riley took 45 minutes and required passing through a security gate and a variety of active firing range complexes. All told, the experiment took at least four hours of the participants' time.

Results

Figure 2a is a graph of response times to the command to move. The open symbols show the means and standard errors of response times for the group of participants who first ran the lane in the remote-leader condition. This group responded more quickly in the second trial when the leader was present in the lane. The closed circles, for the group who first ran in the leader-present condition, show that participants responded more quickly in the first trial, again when the leader was present in the lane. A two-factor ANOVA was conducted to assess sequence and group effects. Neither group (remote-first, present-first), $F(1,36) = .221$, $p > .64$, nor sequence (first trial, second trial), $F(1,36) = .064$, $p > .80$, were significant. This result allows us to merge data across groups and to conduct a within-subjects t-test for the effect of leader presence. The test, $t(19) = 2.958$, $p < .004$, indicates that, as expected, leader presence made a significant difference in the participants' response times to commands to move. Cohen's d as adjusted for the lower variability inherent in a repeated-measures design at an alpha of .05 is approximately .94, indicating ample statistical power with 20 participants.

Figure 2b is the corresponding graph of the response times to the command to fire. The pattern of results is the same: both groups of participants responded more quickly when the leader was present in the lane. A two-factor ANOVA found that neither group (remote-first, present-first), $F(1,36) = .120$, $p > .73$, nor sequence (first trial, second trial), $F(1,36) = .155$, $p > .69$, were significant. The within-subjects t-test for the effect of leader presence, $t(19) = 2.317$, $p < .016$, indicates that leader presence made a significant difference in the participants' response times to commands to fire. The adjusted Cohen's d at an alpha of .05 is approximately .73. Again, the experiment had ample statistical power with 20 participants.

EXPERIMENT 2

Location and Participants

To test the generality of the indoor result from Fort Riley, we moved the second experiment to an outdoors venue on-campus. The setup was exactly the same as the lane in Figure 1 with one exception. The lane was set up in a small field rather than in a building. The change in setting made the leader-remote condition less remote. The first experiment was conducted indoors which allowed the remote leader to be completely out of sight. In the second experiment, the remote leader hid behind a tree approximately 50 meters behind the lane. Thus the leader was in fact visible if the participant chose to turn around and look. Twenty-two students, three women and 19 men, participated in the second experiment. The median age was 19 with a range of 18 to 26.

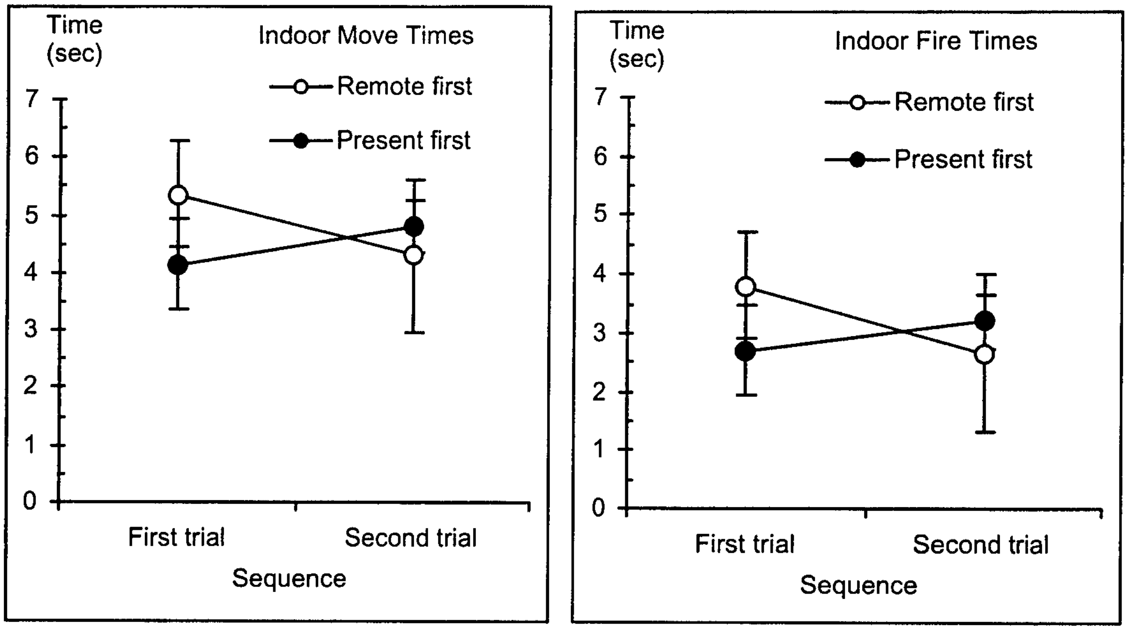


Figure 2 Data from Experiment 1 which was conducted inside a military building. A) Response times to the command to move. B) Response times to the command to fire. Responses are always faster in the leader-present condition.

Results

The graphs of Figure 3 show the response times to the commands to move and to shoot. The open symbols show the means and standard errors of response times for the remote-first group. This group responded more quickly to both commands in the second trial when the leader was present in the lane. The closed circles, for the present-first group, show that participants responded more quickly in the first trial, again when the leader was present in the lane. The ANOVA on sequence and group effects for the command to move show that group was significant, $F(1,40) = 3.779$, $p < .058$. The remote-first group moved significantly more quickly in the second trial when the leader was present in the lane. The test for the effect of sequence, $F(1,40) = 1.17$, $p > .18$, shows no effect of sequence on move time. The ANOVA for fire time indicates that neither group nor sequence were significant, $F(1,40) = .284$, $p > .59$ and $F(1,40) = .181$, $p > .67$, respectively.

We merged the data across groups to conduct a within-subjects t-tests for the effect of leader presence. The test for both move times, $t(21) = 2.798$, $p < .005$, and fire times, $t(21) = 2.211$, $p < .019$, indicates that, as expected, leader presence made a significant difference in the participants' response times to commands to move and shoot. The adjusted Cohen's d at an alpha of .05 is approximately .55 for commands to move and .70 for commands to fire. The experiment had ample statistical power with 22 participants.

Given the similarity of the two experiments' results, it appears the subtle difference in the degree of remoteness of the leader across the two experiments did not have a significant impact on response times. The similarity also allows us to aggregate the data. The test on the composite move data is significant $t(41) = 4.122$, $p < .0001$. The test on the composite shoot data is also significant $t(41) = 3.218$, $p < .0013$. The aggregate power is very high.

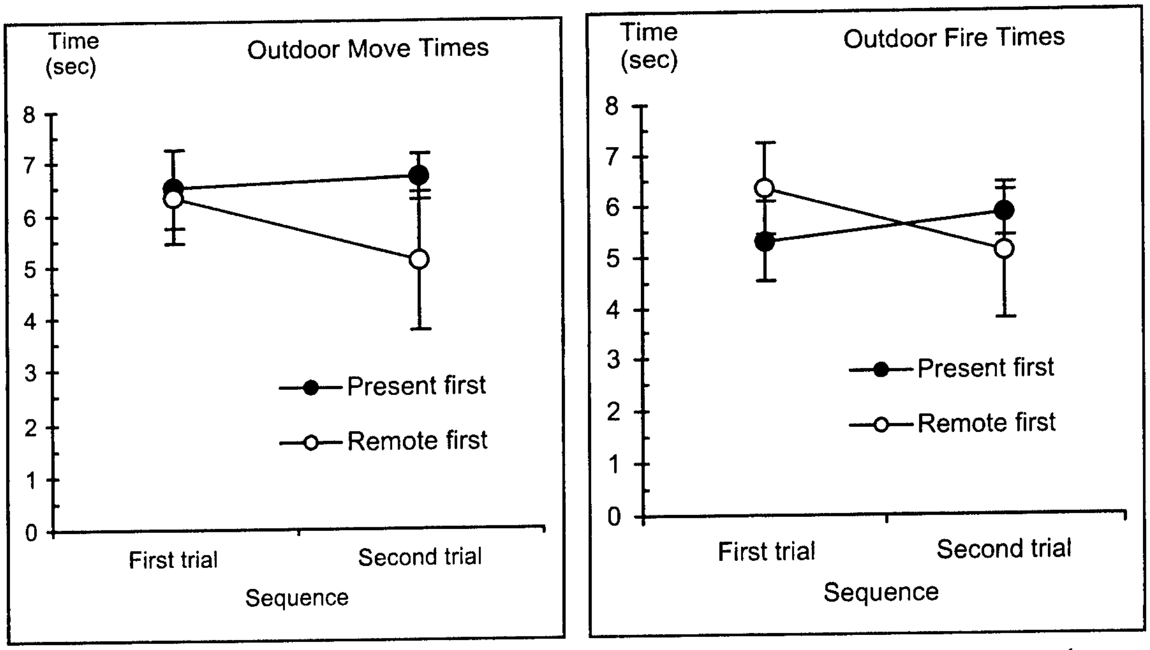


Figure 3 Data from Experiment 2 which was conducted outdoors. A) Response times to the command to move. B) Response times to the command to fire. Responses are always faster in the leader-present condition.

EXPERIMENT 3

The third experiment is in progress at a commercial, indoor paintball arena in Tidan, near Skövde, Sweden. Most of the participants are students at the Skövde Högskolan (college). There is little reason to expect we will find a significant difference between Swedish youth and American youth when asked to follow commands to move and to shoot. The change in setting does, however, present the opportunity to address cross-cultural phenomena (Sutton, 2003). We are currently planning experiments to assess the effects of mixing leader and fire-team nationality.

The third experiment has added an intermediate condition (leader-radio-present) to disambiguate the effects of leader presence and mode of communication. In the intermediate condition, the leader is on the lane one station behind the participant communicating by radio. If leader presence is the major source of variability observed in the first two experiments, then performance in the leader-radio-present condition will be approximately the same as it is in the leader-present condition. In contrast, if the effect is due to radio communication, then performance in the leader-radio-present condition will be like that in the remote-leader condition.

Biometric telemetry is being used to improve measurement of response time. Goniometers (strain gauges) are attached to the participants' and the leader's trigger fingers. Moving the finger stretches the gauge which changes the resistance that is telemetered to a portable computer (Biopac Systems MP150 system, with 2 TEL 100 C-RF remote monitoring modules). The leader bends his finger when he issues a command. Shooting and picking up new ammunition produce distinctive signals. Response times are calculated from the difference in the times of signals in the leader's and the participant's telemetered goniometer data. The telemetry system enables continuous electrocardiographic monitoring of the leader and selected participants. The resulting time series of interbeat intervals are the raw data for studying the correlation between experimental conditions and heart rate and heart-rate variability, two psychophysiologic measures of stress (Backs & Boucsein, 2000).

Data collection will be completed in February 2004 and the results reported at the conference.

DISCUSSION

These results from Experiments 1 and 2 support our hypotheses. Participants were faster to react to the leader's commands when the leader was present than when the leader was remote. This result suggests that some kind of intervention will be required if soldier performance is to be as efficient in remote command and control as it is in the more traditional, leader-present, mode of control. Two classes of intervention come to mind. The first is training.

Does current military training overcome the inherent disadvantage posed by a leader's absence? We hope to address this question by conducting similar experiments with conscripts from the military garrisons in Skövde.

The second intervention is the development of technology that enables 'virtual leaders' to take to the field with their fire teams. The requirements for a virtual leader are not physical or holographic presence but psychological presence. We plan to test alternative designs for information telemetry and display that offset the decrements in performance that accompany remote command and control.

ACKNOWLEDGEMENTS

The views expressed in this work are those of the author and do not necessarily reflect official Army policy. This work was supported by the DOD Multidisciplinary University Research Initiative (MURI) program administered by the Army Research Office under grant DAAD 19-01-1-0621. Dr. Elmar Schmeisser is the project monitor. The paintball paradigm and the assault lane were designed by Lt. Keith Pangburn.

REFERENCES

Milgram, S. (1974). Obedience to Authority. New York: Harper & Row.

Pangburn, K., Freund, S., Pangburn, H., & Smith, K. (2003). Beyond Milgram: An experimental study of leader presence. Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society, 364-368.

Sutton, J. L. (2003). Validation of cultural awareness training concept. Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society.

POTPOURRI

RECOGNITION AND RESPONSE TIME FOR NON-VERBAL CROSS-CULTURAL COMMUNICATION

Sherri A. Rehfeld and Florian G. Jentsch
University of Central Florida

ABSTRACT

Gestures vary widely around the world in regard to their type and meaning. This research project sought a way to display various gestures to benefit travelers in learning the gesture and it's meaning. Therefore, hand gestures with similar and different meanings across cultures were tested to optimize learning and transfer of learning to novel stimuli across displays. The four displays include (a) a text description of the gesture and meaning, (b) the same descriptive text augmented by a full body image of the gesture, (c) the text with a stereotypically dressed, full body image, and (d) the text with a detailed image of the hand gesture. Results showed that gestures with same meanings across cultures produced higher accuracy and shorter response times. In addition, participants responded faster in the transfer of knowledge condition. Finally, the addition of an image to illustrate the gesture decreased response time considerably over simple textual description, with no significant differences between the conditions with images. Limitations and future follow-up studies are discussed.

Keywords: cross-cultural communication, gestures, emblem gestures, iconic gestures

INTRODUCTION

In April of 2003, a U.S. military convoy was filmed traveling past Iraqi citizens during Operation Iraqi Freedom. Many of the citizens were waving and cheering. The atmosphere was one of hesitant jubilation. The military personnel showed the Western gesture of victory, the index and middle fingers in the shape of a "V" and the remaining fingers and thumb tucked together with the palm facing outward. This specific gesture, however, does not carry meaning in Iraq. Further, one of the Iraqi citizens made the same gesture with the palm facing inward. In Iraq, this gesture is vulgar and represents an insult. Neither of the two cultures may have recognized or realized the meaning disparity between the gestures. Whereas people who travel to different countries may be excused for using improper language since the accent of the traveler lets people know that they are not familiar with the customs of the country, use of non-verbal communication does not provide such an excuse since no accent is realized.

Gestures Defined

Although wide variations exist in gesture definitions in the literature, Kendon's continuum is a suitable and thorough breakdown of non-verbal communication (Kendon, 1988; McNeill, 1993). The original continuum presented by Kendon (1988) suggests the growth of gestures from simple gesticulation, to emblem, pantomime, and, lastly, to sign language. *Gesticulation* is the effortless movement of the hands to accentuate and assist speaking. This can vary from simple hand movements during a conversation to planned illustrations during a speech. Conversely, *emblem* gestures are small movements of the hand that convey a meaningful thought or expression such as the American "OK" gesture (index finger and thumb form a circle and the remaining fingers are pointed straight up). *Pantomime* is the deliberate movement of the entire body with exaggerated facial expressions to tell a story sans spoken language. Finally, *sign language* is the use of motion for the replacement of verbal speech altogether, most often for people who cannot speak, hear, or both. The idea behind the Kendon's continuum was the progression of rudimentary movements to polished motion that completely replaces speech.

The current project changed the order of the continuum slightly to reflect the influence of culture on the evolution of gestures (Figure 1). Gesticulation may be more common and exaggerated in some countries (e.g., Italy), but it is generally used worldwide without meaning attached to the motion. Similarly, pantomiming surpasses cultural boundaries as one can determine the story line no matter the dialect of the performer or the audience. Emblem gestures, however, have specific meaning attached to the motion or signal depending on the country and culture of the person giving and receiving the gesture. Likewise, sign language is culture specific; for example, there is the American Sign Language, Australian Sign Language (AUSLAN), and Italian Sign Language. Each differentiation along this modified continuum increases the cultural influence on the gesture motion.

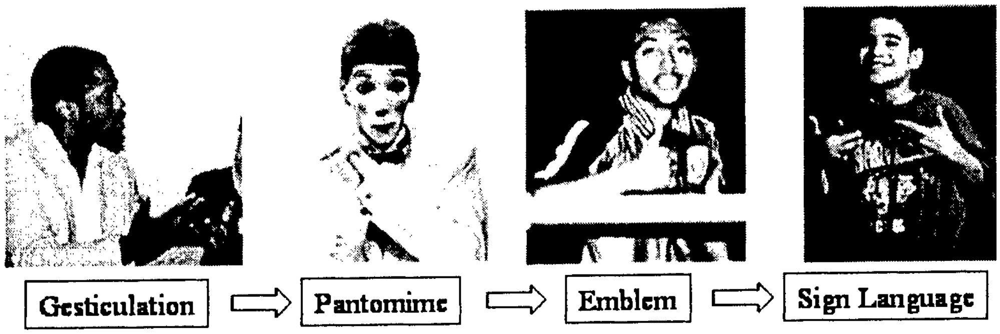


Figure 1. Kendon's continuum of gestures, modified to show cultural influence.

Gesture Research Study

As military personnel and civilians travel around the world for extended periods of time, smooth interaction with a new and different culture is important for the traveler to communicate effectively. Although focus on the ability to communicate in a different language is paramount, non-verbal communication is just as important. A simple gesture given in a novel environment may (a) mean nothing at all, (b) invite an unwanted response (such as a sexual connotation), or (c) be unintentionally offensive and vulgar.

Currently, cultural awareness training for the military, while detailed, may be lacking in the area of non-verbal communication. Similarly, a sample of the commercially available literature shows very little to no information available to enlighten travelers of the influence that their own gestures may have on a different population or the gestures that they may encounter. To address this issue, the current study combined the relevant research with commercially available information to build a database of worldwide gestures (Axtell, 1988; Bauml & Bauml, 1975; "Cultural Gestures," 2003; Kavanagh, 2000; Morris, Collett, Marsh, & O'Shaughnessy, 1979).

The aim of this project was to determine the best way to present the relevant information of the gesture database in a way to best enhance human performance through learning, memory, and speed of response. It is essential to note that some gestures carry the same meaning across cultures while other gestures have vastly different connotations. Likewise, the ability to learn the information is negated if the person is unable to transfer this knowledge to the country or culture visited, which will be novel in nature. Therefore, the current study used 11 countries and 16 gestures (eight with the same meaning across countries and eight with different meanings) and tested the ability to learn the information and apply it to novel stimuli. Specifically, we studied different ways to present cross-cultural gesture information. Four formats were tested: (a) a description of the gesture and its meaning with only text, (b) the same descriptive text augmented by a plain, full body image of the gesture, (c) the text with a stereotypically dressed, contextually relevant, full body image, and (d) the text with a detailed image of the hand gesture.

Given that there are fewer meanings to learn and remember, it was hypothesized that the gestures with similar meanings would produce shorter response times and higher accuracy. Since testing of the gestures would be repeated with the novel stimuli, the transfer of learning was expected to generate shorter and more accurate responses. Finally, of the four displays, the text only display was anticipated to have the longest, and least accurate responses, followed by the plain image, the contextually relevant image, and the detailed image was predicted to produce the shortest, most accurate responses.

METHOD

Participants

Forty-four undergraduate students from a large Southeastern university participated in exchange for research credit. Five participants were removed from analyses due to a change in the computer program, and one participant was determined to be an outlier (more than three standard deviations below the mean; most likely due to disinterest, as all answers were very fast negatives). These cases were not included in the analyses, resulting in 38 datasets used in the analyses. There were 10 participants each in the Text-only and Detailed image with text conditions and nine participants each in the Plain image with text and Contextual image with text conditions.

Apparatus

An IBM-compatible computer accommodated the Visual Basic program that presented the gesture stimuli and recorded response times (RT) and errors.

Design and Measures

A four level (display type; Text only, Plain image with text, Contextual image with text, Detailed image with text), between subjects design was employed. The Text only display presented the name of the gesture, the country where the gesture is used, and the meaning of the gesture in that country in Time New Roman 16 point font. This identical text was presented in every condition. The Plain image with text included a full body, gender neutral, and expression neutral figure created in Poser 5.0. The Contextual image with text included a full body, expression neutral figure in stereotypical dress of the country for that gesture. Finally, the Detailed image with text showed a close-up view of just the gesture itself. View time was measured for the presentation of the gesture information as well as for the multiple-choice questions. The reading rate throughout the training session was used as a covariate during the multiple-choice questions to mitigate varied reading rates among participants. In addition, accuracy of the questions was also recorded as the proportion of correct responses.

Procedure and Task

After receiving instructions from the experimenter, the participants began the computer program. To advance each slide throughout the program (viewing the gestures and answering multiple-choice questions), the participant clicked on the NEXT button with the computer mouse and 32 slides followed in which each of the 16 gestures would be represented by two countries, eight have the same meaning for both countries and eight have different meanings between the countries (11 total countries used). Therefore, each gesture was presented two times, sequentially, with two countries and meanings per gesture provided on two different slides. The countries were then reviewed to remind the participant of the countries to which the gestures pertained.

Between the training and transfer tests, a distractor task was implemented consisting of 5 min worth of long division and multiplication problems to prevent rehearsal of the gesture meanings. Following this task, 32 multiple-choice questions were presented. The questions stated the country name, gesture description in the format presented for that condition (Text only, Plain image with text, Contextual image with text, or Detailed image with text), and gave four choices for what the gesture means in the country stated. All of the distracter items in the multiple choice options were taken from other gesture meanings so that all meanings had been viewed previously. An additional set of multiple-choice questions (32) were presented wherein the format was consistent across all conditions and contained a contextual image with different colored clothing to test transfer of learning. At the end of the computer program, the participants were thanked and given research credit in accordance with university policy.

RESULTS

The dependent measures of interest were the response times (the amount of time, in milliseconds, to respond to the multiple-choice questions) and the accuracy of each response (the percentage of correct responses). The between group factor was the type of display with four levels (Text only, Plain image with text, Contextual image with text, Detailed image with text). The within group factors were the gesture meaning across countries (same, different) and testing (learning, transfer of learning).

All analyses were conducted using SPSS, 11.5. Preliminary analyses of the data were performed to assess the underlying assumptions of normality and homogeneity of variance. No serious violations of the assumptions were noted. Unless stated otherwise, the alpha level used in the analyses was conservative and set at 0.017 to account for alpha inflation with the number of tests used (six tests with original 0.10 alpha level from stated hypotheses).

Meaning

As expected, the gestures that had the same meaning ($M = 6724$ ms, $SD = 2063$ ms) showed a shorter response time to the multiple-choice questions than the gestures with different meanings ($M = 7560$ ms, $SD = 2119$ ms) across countries, $t(37) = -4.726$, $p < .0005$. The gestures with the same meaning also produced a higher proportion of accurate responses ($M = .934$, $SD = .079$) than different meanings ($M = .887$, $SD = .088$) across countries, $t(37) = 5.373, p < .0005$.

Testing

The second set of multiple-choice questions represented the transfer of learning to novel stimuli and showed a shorter response time ($M = 6594$ ms, $SD = 1765$ ms) than the first set of multiple-choice questions ($M = 7846$ ms, $SD = 2722$ ms), $t(36) = 4.193, p < .0005$. However, there were no significant differences in accuracy between the learning and the transfer of learning questions, $t(37) = 0.358, p = .722$.

Condition

Two one-way, four level ANOVAs tested response time (with the covariate of reading time during training) and accuracy among the conditions (Text only, Plain image with text, Contextual image with text, Detailed image with text). As shown in Figure 2, the text alone condition had longer response times than any of the other four conditions, $F(3, 34) = 11.481$, $p < .0005$, $\eta^2 = .503$. However, there was no significant difference between the conditions, $p = .735$, in regard to accuracy.

DISCUSSION

The results of the analyses supported a number of the hypotheses: First, gestures that have the same meaning across countries are more effectively learned (as expressed by more accurate responses and shorter response times) than gestures that have different meanings across cultures. This was expected since gestures with a universality of meaning across countries should be easier to learn than when specific gesture meanings must associated with each particular country. Second, the multiple-choice questions testing the transfer of learning (the second set of questions) had shorter response times than the initial learning test (the first set of multiple-choice questions). As explained by the hypotheses, this may be due to practice since the second set of questions, while presented in a different order, are the same as the first set of questions.

The initial hypotheses also stated that the text information alone should result in the highest errors and longest response time. While the Text only condition showed the longest response time, the accuracy results were not significantly different between any of the four conditions. Finally, the Detailed image with text condition was not significantly different in terms of response time or accuracy than the remaining conditions.

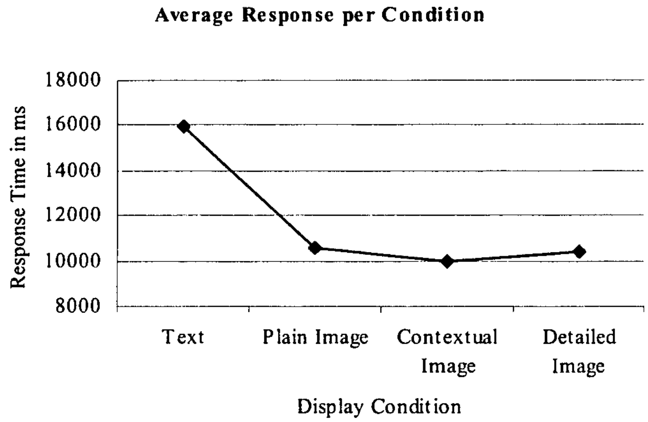


Figure 2. Average response time across display conditions.

One explanation for the data is that given the consistently high (all above 90%) accuracy results of the four conditions, there is a possible ceiling effect within the data. This ceiling effect does not allow for differences between the conditions to be distinguished. Given this likelihood, a follow-up study is planned to more accurately reflect the application of the gesture knowledge in a bona fide, real-life situation. This will change the task from one of recognition (by choosing from multiple options for the answer) to a one of recall (having to remember the information and write it down). For example, in a situation in which a gesture is seen, the person must remember the connotation of the gesture as well as its meaning.

Implications

This was a first study to test the learning of gestures across cultures, specifically as a function of the presentation of the gesture information during learning. The results were encouraging, as (a) hypothesized differences in learning between gestures with same and different meanings showed up consistently, and (b) participants were able to learn hand gestures quite effectively, even when their meanings differed across countries. The consistently high accuracy in the responses negated any effects of the format-manipulation on the one hand, but on the other hand also suggested that learning gestures is a comparatively natural task. Further investigations, however, are needed to determine whether memory for gestures is equally good when recall, rather than recognition, of gestures and their meanings are required.

REFERENCES

Axtell, R.E. (1988). Gestures-the DO's and TABOOs of body language around the world. New York: John Wiley & Sons.

Bauml & Bauml (1975). A dictionary of worldwide gestures. Lanham, MD: Scarecrow Press.

Kavanagh, J. (2000). Worldwide gestures: A traveller's introduction to common international gestures and symbols. Blaine, WA: Waterford Press.

Kendon, A. (1988). How gestures can become like words. In F. Poyatos (Ed.), Cross-cultural perspectives in nonverbal communication, pp 131-141. Toronto, Canada: Hogrefe.

McNeill, D. (1993). The circle from gesture to sign. In M. Marschark and M. D. Clark (Eds.) Psychological Perspectives on Deafness, pp 153-183. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Morris, D., Collett, P., Marsh, P., & O'Shaughnessy, M. (1979). Gestures, their origins and distributions. London: Jonathan Cape.

No author (2003). Cultural Gestures. Retrieved January 29, 2003, from http://www.webofculture/worldsmart/gestures.asp.

CAN REGULATION WORK? UNDERSTANDING THE CONSEQUENCESOF FORCED COMPLIANCE IN HUMAN FACTORS REGULATION

Christina M. Frederick-Recascino, Ph.D.
Embry-Riddle Aeronautical University

ABSTRACT

Human factors-based regulations are currently in place in a variety of areas within the aviation environment; however some of these regulations have been met with resistance and non-compliance. An example of this of this type of resistance can be found in crew resource management (CRM) training. Although mandated by the FAA, Helmreich & Wilhelm (1991) report that a subset of pilots continue to reject CRM and its applications in the cockpit. Due to situations such as these, understanding the dynamics of non-compliance is important for researchers and practitioners within the HF field. Although non-compliance has been studied in a general sense and has been linked to lack of expertise and cost issues, few researchers have examined individual barriers related to non-compliance with HF regulations. The proposed paper has three purposes. First, the paper will address why regulations may not facilitate use and acceptance of human factors programs in the aviation environment. This discussion will emphasize psychological states that arise as a result of forced compliance. A second purpose of the current paper is to present a specific framework for studying and understanding a specific psychological barrier, that of motivation, in implementation of human factors policies. The final purpose of the paper is to provide suggestions for overcoming negative psychological states and motivational barriers in human factors' implementation, even in those cases when structured regulations are deemed necessary.

Keywords: Motivation, compliance, self-determination

INTRODUCTION

Issues related to compliance and coercion have been of interest to social psychologists since the late 1940's. Milgram's now famous shock experiments indicated that 65% of people asked to comply to a experimenter's demands to deliver shock to another individual actually did comply (Milgram, 1963). It may be surprising at first that so many participants went along with Milgram's request, however equally interesting were the large minority of participants who did not comply fully. Why is it that some people easily accept forced regulation and compliance, while others fight such pressures every step of the way?

One explanation involves individual differences in the reaction to requests for compliance. A variable that has been shown to be a powerful correlate of non-compliance is psychological reactance (Brehm & Brehm, 1981). Reactance is a personally experienced negative and emotional reaction to a request for compliance or obedience. Adults are more likely experience this phenomenon when confronted with controls, rules and regulations they perceive as externally decided and/or arbitrary. If issues of reactance are not addressed and diffused as they occur, non-compliance, entrenchment of position, hostility and even aggressive action often occur.

Related to the issues of compliance and reactance are the concepts of conformity and conversion. An individual experiencing reactance may indeed behaviorally conform to a regulation, however he/she may continue to privately object to, or not accept, what he/she is being asked to do. In contrast, over time some individuals move from reactance to conformity to conversion. Conversion occurs when one not only behaviorally conforms, but also privately accepts the requests for compliance as being legitimate and valuable. In the aviation domain, regulations and requests for compliance are developed in order to enhance the safety and efficiency of the system. Ideally, one would wish to move recalcitrant employees away from reactance and toward conversion.

Understanding Motivational Factors as Barriers to Compliance

It is important to understand the psychological principles underlying compliance and non-compliance. However, it is equally valuable to have a framework from which situations can be analyzed for their likelihood to create non-compliance, and for their effect on the perceptions of the individuals operating within them. It is proposed that Self-Determination Theory (SDT:Deci & Ryan, 1985, 1991) provides a viable framework for understanding compliance issues. SDT is a motivational theory that distinguishes between two different types of motivation: extrinsic and intrinsic. Each type of individual motivation is derived from situational factors and the interpretation and experience of those factors. Compliance can exist with either type of motivation, however as motivation becomes more internalized, compliance is likely to become more self-determined and consistent.

Levels of Internalization. As mentioned above, addressed within Self-Determination Theory is the distinction between intrinsic motivators and extrinsic motivators. Furthermore, extrinsic motivation is differentiated into levels reflecting various degrees of internalization. At the lowest level of these levels is external regulation. At an external level, behavior is completely determined by external sources. Coercion and forced compliance with no behavioral options can be examples of external regulation. At this level of motivation, feelings of pressure and control are salient. Psychological reactance, and resistance to authority are very real issues.

At the next level of regulation, behavior moves from being entirely externally controlled to being internalized at an introjected level. Introjection occurs when individuals act in order to gain social approval or alleviate feelings of guilt. As such, it is likely that compliance will occur sporadically and may be based upon the value of the authority figure to the individual being asked for compliance and the social repercussions of non-compliance. As behavior continues, individuals can further internalize their action and attain a level of identified internalization. Identification occurs when individuals relate to, or identify with, the goals and purposes of their actions. Internalization of behavior often occurs as a developmental progression as individuals become more familiar with a particular domain or as they mature cognitively (Ryan & LaGuardia, 2000). At an identified level, compliance would be perceived as a self-determined choice made in service to the domain goals desired by the individual. Feelings of achievement, satisfaction and purpose are often experienced in association with identified regulation.

In contrast to the levels of extrinsic motivation, intrinsic motivation reflects true self-determination of behavior, driven by personal interest and challenge. Affective states associated with intrinsic motivation include: satisfaction, feelings of competence, self-esteem, exhilaration, happiness and interest. Behavior has been shown to be most persistent in a state of intrinsic motivation. Compliance at an intrinsic level can only be obtained choicefully by the individual, in a situation that provides optimal challenge.

Based on this conceptualization of motivation, it is easy to see that regulations tend to place motivational action within the externally controlled level of extrinsic motivation. It is, however in the best interest of those implementing regulations to try to facilitate rapid internalization of behavior within those individuals adhering to regulations, in order to enhance compliance. It is in the best interest of those in management to focus on moving individual to a more internalized level of motivation. Reeve (2001) has indicated that for tasks that are deemed important, but are not experienced as intrinsically motivating, identified motivation should be the goal toward which one should strive. Past literature has indicated across a variety of domains, including education (Deci & Ryan, 1985), sports (Frederick, 2001) and work (Deci, Connell & Ryan, 1989), that greater internalization of behavior is associated with better performance and higher levels of behavioral adherence. By analyzing the regulatory environment and the motivational state it creates in the individual within that environment, it is likely we can predict associated levels of compliance and then intervene to increase compliance.

Before concluding this section, a word needs to be said in support of a motivational analysis of aviation work environments. Motivation is not often associated with human factors issues, however support is growing for use of a motivational perspective. Paries & Amalberti (2000) present a safety paradigm for aviation that emphasizes an underlying philosophy of "freedom". A freedom-based paradigm is a motivational one that stresses personal choice, challenging and meaningful training and a system's perspective for understanding safety errors. Further support can be gleaned from Maurino, Reason, Johnston & Lee's (2001) analysis of the causes of 24 CFIT aviation accidents. According to their results, 3 of 24 accidents involved organizational deficits in motivating employees. It is believed that a focus on using motivational techniques to understand and then facilitate compliance may be valuable in meeting Maurino et al.'s goal of moving to the zone of maximum resistance to safety errors and remaining in that zone.

Solutions for Non-Compliance

Based on the motivational theory just discussed there are a variety of ways in which organizations can develop techniques designed to facilitate internalization of regulations.

Some of these solutions include: the correct use of rewards and feedback, peer-modeling behaviors, changes in cognitive strategies, and structural changes in airline organizations.

Use of Rewards. Using rewards and feedback to motivate individual compliance to regulations is widely used, however this technique is difficult to use correctly from a motivational standpoint. There are many problems associated with the use of rewards including the fact that once a reward system is implemented, it needs to be maintained. One cannot gain compliance through rewards and then cease the reward structure. An example of a program developed via the use of rewards is the behavioral safety program used by the U.S. Department of Defense to regulate the nuclear industry (Waters & Duncan, 2000). A positive feedback and reward system has been successful in lowering incidents, however this program can never be reduced or ended, because once reinforcement is ceased, employees will abandon their safety focus. These systems tend to regulate behavior at an external level of motivation and although outward conformity may be gained, internal compliance is not. If a feedback and reward system is used to facilitate compliance to regulations it must be entered into carefully and rewards should not be continuous, expected or too low to guarantee compliance. Providing appropriate situation-centered, behavior contingent and honest feedback can be used to help motivate and engage employee behavior. It is important that the feedback is provided in an informational manner and not in controlling way in order to reduce psychological reactance. In order to facilitate the correct use of feedback, the organization must adopt a learning and developmental perspective in which feedback is appreciated and not used as a punisher.

Involvement in the Process. One key technique that has been used in order to facilitate internalization of motivation and commitment to behavioral options has been inclusion of those individuals affected by regulations in the decision-making process. Having a forum in which one can ask questions, express opinions and even work within a team to help modify and improve regulations is likely to facilitate adherence to those regulations. From a motivational viewpoint, this type of participation creates a situation in which the employee operates at an identified level of motivation, focusing on regulation adherence because it meshes with their own beliefs, plans and goals.

Peer Modeling. When actual participatory management cannot occur, using successful peers to teach and model desired behaviors may be an option. This process needs to take place in an environment centered around cooperation. Cooperative learning facilitated by peer mentors is an excellent way to develop organization-wide recognition of the value of regulatory adherence. From a motivational perspective, this type of intervention can begin the process of internalization of behavior and would likely appeal to younger employees who look for mentoring and social approval.

Cognitive Change. Additional training can also be provided, which teaches individuals affected by regulations to be aware of their own cognitive reactions and illogical thought processes, so that they may be able to self-monitor and decrease undesirable psychological reactions such as reactance. Once an individual is trained to be aware of his/her cognitions, he/she can learn to gain some control over his/her responses. This type of training process is usually referred to as cognitive restructuring.

Organizational Change. Often motivational changes influencing behavior can result from changes in the structure of the environment. Changes which have proven valuable in creating higher levels of internalized motivation toward regulations include: consistency in organizational attitudes toward regulatory behaviors, facilitating employee input in training for regulatory behaviors, viewing each employee as being on a developmental trajectory within the organization and knowing that internalization often occurs naturally over time. The more the organization believes in and promotes the importance of regulatory policies, as benefiting employees and consumers, the more likely the organization will be creating the foundation upon which identified motivation can be built.

Analysis of CRM, Motivation and Compliance Issues

CRM training is one area in which non-compliance to regulations has been extensively documented. It is believed that psychological reactions to CRM regulations, a lack of self-determination, and industry-wide inconsistency in CRM training have contributed to this situation (Maurino, 1999). In the U.S., although CRM is required, each airline has a different training program adapted over time. Some airlines provide personality testing, others focus more on crew coordination. Some provide detailed review of past accidents and little else. In addition, the skill and knowledge of CRM facilitators also varies widely. A skilled facilitator can challenge students in a learning-based environment. However, an unskilled facilitator often creates an environment of boredom, disrespect and reactance.

The result of these conditions is an environment in which 10% of pilots are openly anti-conformist in their attitudes toward CRM (Helmreich & Wilhelm, 1991). What has never been estimated is the percentage of pilots and crewmembers that manifest outward conformity to CRM training, while still being inwardly non-compliant. For all functional purposes, both of these groups are problematic due to their low level of motivation and personal investment in CRM.

It is important that an industry-wide dialogue be established to create consistency in the goals of CRM training and its importance to the industry as a whole. Honest belief in the importance of CRM for the industry and provision of information supporting this position will help alleviate feelings of reactance and external pressures. This type of clear structure can provide a foundation for the developmental process of motivational internalization to occur.

Actual CRM sessions should be run by trained facilitators who are able to provide challenge and learning to all levels of expertise. If this is not possible within a single training, then domain-specific student experts could be utilized within the CRM training. Another possibility for creating a challenging environment that can foster intrinsic motivation is to break individuals into groups based on expertise and knowledge levels. Thus, a more specialized training can be provided to all students. Students with very high expertise levels could be groomed to move into CRM facilitator slots, providing peer role models.

Any of the suggestions just made could and should be tested in a systematic fashion in both laboratory and real-life settings.

DISCUSSION

This paper presents a framework for understanding why regulations in the aviation environment do not always achieve their desired ends. However, safety concerns do require that regulations are created and enforced. With this in mind, the paper presents a conceptual framework that can help explain non-compliance, as well as a set of strategies that could be used to increase overall compliance rates for a variety of regulated behaviors in the aviation domain.

REFERENCES

Brehm, S. & Brehm, J. (1981). Psychological reactance: A theory of freedom and control. New York: Academic Press.

Deci, E.L., & Ryan, R.M. (1991). A motivational approach to self: Integration in personality. In R. Dienstbier (Ed.), Nebraska symposium on motivation 1990: Perspectives on motivation, (pp. 237-288). Lincoln: University of Nebraska Press.

Deci, E.L., & Ryan, R.M. (1985). Intrinsic motivation and self-determination in human behavior. New York: Plenum.

Deci, E.L., Connell, J.P., & Ryan, R.M. (1989). Self-determination in a work organization. Journal of Applied Psychology, 74, 580-590.

Frederick-Recascino, C.M. (2003). Self determination theory and participation motivation research in the sport and exercise domain. In E. L Deci, & R. M. Ryan (eds.), Handbook of self-determination research. Rochester, NY: University of Rochester Press.

Helmreich, R. & Wilhelm, J. (1991). Outcomes of crew resource management training. International Journal of Aviation Psychology, 1, 287-300.

Kelman, H. C. (1958). Compliance, identification and internalization: Three processes of attitude change. Journal of Conflict Resolution, 2, 51-60.

Maurino, D. (1999). Crew resource management: A time for reflection. In D. Garland, J. Wise & D. Hopkins (Eds.), Handbook of aviation human factors. New Jersey: LEA.

Maurino, D. E., Reason, J., Johnston, N., Lee, R. B. (2001). Beyond aviation human factors: Safety in high technology systems. Burlington, Vermont: Ashgate Publishing.

Milgram, S. (1963). Behavioral study of obedience. Journal of Abnormal and Social Psychology, 67, 371-378.

Paries, J. & Amalberti, R. (2000). Aviation safety paradigms and training implications. In N. Sarter and R. Amalberti (Eds.) Cognitive engineering in the aviation domain (pp.253-286). New Jersey: LEA.

Reeve, J. (2001). Understanding motivation and emotion. Orlando, FL: Harcourt College Publishers.

Ryan, R.M. & La Guardia, J. (2000). What is being optimized: Self-determination Theory and basic psychological needs. In S. Qualls and N. Abeles (Eds.), Psychology and the aging revolution. Washington: American Psychological Association.

Thibault, J. W. & Kelley, H. H. (1959). The social psychology of groups. New York: Wiley.

Waters, R. & Duncan, M. (2000). Behavioral safety programs in the department of energy. Unpublished manuscript.

THE EFFECT OF OLFACTION ON IMMERSION INTO VIRTUAL ENVIRONMENTS

Lauriann Jones, Clint A. Bowers, Don Washburn, Art Cortes and Ravi Vijaya Satya
University of Central Florida &
Institute of Simulation & Training

ABSTRACT

This study was part of Research in Augmented & Virtual Environment Systems (RAVES), a cross-disciplinary project researching multi-modal virtual environments. The purpose of this research was to test the impact of olfaction on a human operator's sense of immersion into a virtual environment. The application of which could enhance military training environments to optimize performance in the field. The study was a 2 x 3 x 2 mixed factorial design with gender (i.e., male, female), condition (i.e., control/no scent, experimental/concordant scents, discordant scent), and time (before vs. after) as the independent variables. Scores from an augmented immersion questionnaire served as the dependent variable. The experimental group did not differ significantly from the control or discordant groups in any analyses but the conditions differed significantly on their ratings of the augmented virtual environment and genders differed significantly in their experience in the augmented virtual environment, but not by condition.

Keywords: Augmented reality; Virtual environments; Virtual reality; Immersion; Olfaction

INTRODUCTION

The RAVES objective is to gain a deeper understanding of the development and utilization of virtual environments through research with unique applications of existing technologies and the development of new technologies to optimize human cognitive processing. Simulation training (the use of computer simulations of environments and/or situations to train individuals or groups) has proven to successfully utilize dual modalities, such as visual and verbal, limiting cognitive overload and aiding human cognitive processing (Bowers & Jentsch, 2001; Wickens & Hollands, 2000). Possibly, the addition of simulated olfactory environments would enhance the training experience by increasing immersion. Or possibly, the use of an olfactory component may be used to convey messages when one's visual or auditory modalities are already being utilized, reducing interference.

Olfaction, "the sense of smell or the act of smelling", appears on the surface to maintain separation from visual/spatial or verbal/auditory modalities (Reber, 1995). Olfactory/odor memory is considered to have reliable qualities, commonly known as "Proustian characteristics" which include resistance to interference, uniqueness, and independence from other modalities, (Annett, 1996, Danthiir, Roberts, Pallier, and Stankov, 2001; Herz & Engen, 1996). Larsson (1997) stated that, "verbal/semantic factors play a negligible role in olfactory memory".

Olfaction has proven to play a significant role in human learning and memory. The addition of an olfactory component has been found to reduce stress, increase information processing, enhance memory performance (e.g., enhanced problem-solving, reduced response times and errors, increased recall, recognition, and retention), and enhance productivity, physical performance (e.g., running speed, hand grip strength, number of push-ups), and odor identification (Cain, de Wijk, Lulejian, Schiet, & See, 1998; Degel, Piper & Koester, 2001; Herz, 2000; Kole, Snel & Lorist, 1998; Lesschaeve & Issanchou, 1996; Livermore & Lainge, 1996; Rabin, 1988; Raudenbush, Corley, & Eppich, 2001; Parker, Ngu, & Cassaday, 2001; Schab, 1991; Wickens & Hollands, 2000; White & Treisman, 1997; Wood & Eddy, 1996). If olfaction works separate from other modalities, than the addition of an olfactory component may uniquely augment the cognitive processes of human operators experiencing the least optimal stress levels (low or high) for optimal performance without additional cognitive overload (Chu & Downes, 2001; Kole et al., 1998; Parker et al., 2001; Raudenbush et al., 2001; Schab, 1991).

The purpose of this research was to test the impact of olfaction (i.e., the sense of smell or act of smelling) on a human operator's sense of immersion into a virtual environment (i.e., augmented reality). The application of which could enhance military training environments to optimize performance in the field. Future applications could extend benefits of tapping the olfactory modality to human performance on tasks that already involve dual-modalities. Such as, when the human operator's visual and auditory modalities are overloaded.

METHOD

Design

The study was a 2 x 3 x 2 mixed factorial design with gender (i.e., male, female), condition (i.e., control/no scent, experimental/concordant scents, discordant scent), and time (i.e., before vs. after) as the independent variables and scores from an augmented immersion questionnaire as the dependent variable.

Participants

Participants were 30 volunteer college students from the southeast, U.S. (ages 17 – 27 yrs.) Ten participants (5 males, 5 females) were randomly assigned to each condition.

Materials & Procedure

Each participant was given a consent form, a pre-manipulation check (to identify any odors present in the room), a demographics form with embedded pretest items, a map, computer controls sheet, and fitted with a headset (Plantronics) with a hidden olfactory dispersion system (ScentAir Technologies). Participants played a computer game (i.e., IGI-2 Covert Strike) on a large (approx. 5'x5') panoramic screen for 5 minutes, where depending on the condition, the participant experienced no scents throughout the game (i.e., control), "ocean mist" by the ocean and "musty" scent in the fort (i.e., experimental/concordant scents), or "maple syrup" (i.e., discordant scent) throughout all environments. After completion of the virtual environment task (i.e., computer game), the participants answered an augmented immersion questionnaire (for rating their experience, environment, immersion, etc.) followed by a post-manipulation check to identify any odors left in the room.

RESULTS

The addition of an olfactory component did not significantly enhance immersion into a simulated environment (i.e., the experimental group did not differ significantly from the control or discordant groups in any analyses). Repeated measures ANOVAs were run on Condition x Gender x Time (pre/post items) and there were no significant findings. Pre and post tests revealed an experimental group with unusually high ratings for their previous experiences (Graph 1a, 1b), environments (Graph 2a, 2b), and reality.

A multivariate ANOVA was run for Condition x Gender on the augmented immersion questionnaire. The conditions/groups differed significantly on their ratings of the augmented virtual environment, $F(2,24) = 3.43$, $p = .049$. Tukey HSD revealed that the Control group had significantly higher ratings of the augmented virtual environment than the Discordant group, $p = .04$ (see Graph 3). Genders differed significantly in their experience in the augmented virtual environment, but not by condition, $F(1,24) = 6.13$, $p = .02$. Males had significantly higher ratings of their experience in the augmented virtual environment than did females (see Graph 4). There were no significant interactions.

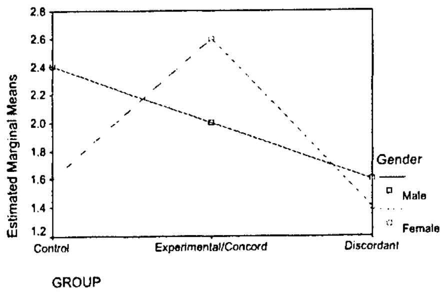
DISCUSSION

It appears that in the attempt to create an immersed environment (e.g., panoramic screen, very realistic graphics and sound) an overall "wow" effect may be created from which the addition of an olfactory component went unnoticed or ignored. It is recommended, future studies utilize a within-subject repeated measures design where subjective differences in conditions may be better differentiated. Additionally, the development of automated systems to run the rather complex experiment would be preferable to reduce experimenter error.

Research into the possible benefits of olfaction to multi-modality, immersion, and augmented reality systems for the optimization of human information processing, is an important and difficult line of research for which technology is only beginning to breach. It is our hope that the results of this study help guide future research in pursuit of such goals.

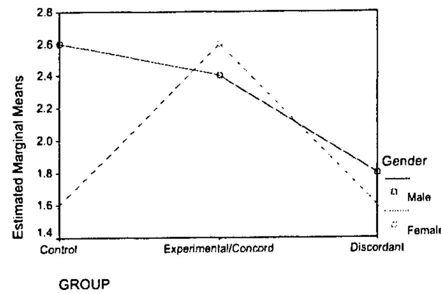
Graph 1a: Means of Pre-Test

Augmented Environment Experience



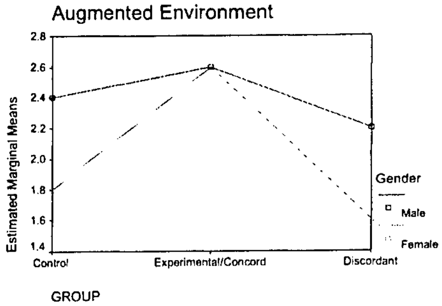
Graph 1b: Means of Post-Test

Augnmented Environment Experience



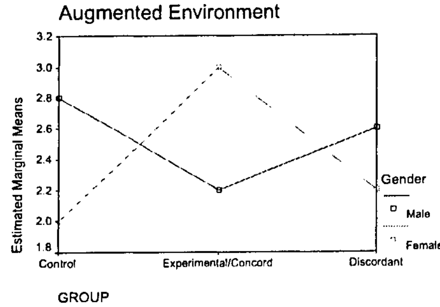
Graph 2a: Means of Environment

Augmented Environment

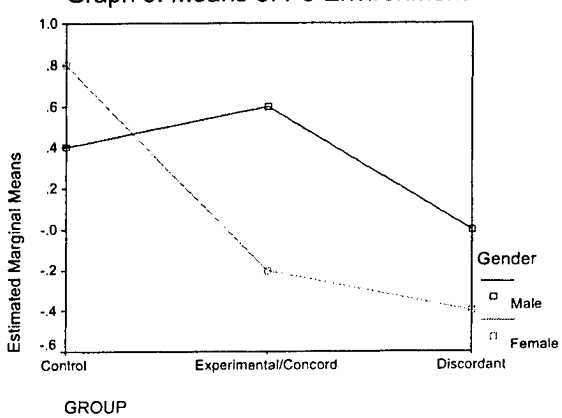


Graph 2b: Means of Environment

Augmented Environment

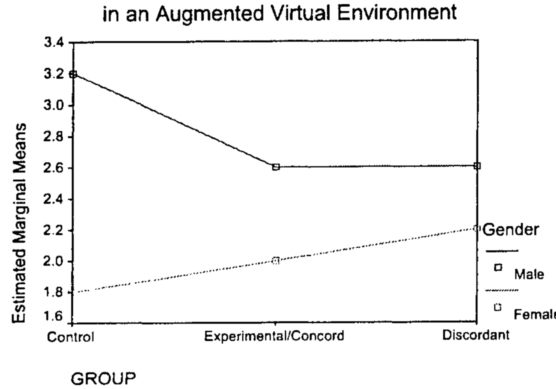


Graph 3: Means of P5 Environment



Graph 4: Means of P3 -Experience

in an Augmented Virtual Environment



REFERENCES

Annett, J. M. (1996). Olfactory memory: A case study in cognitive psychology. Journal of Psychology, 130(3), 309-319.

Bowers, C. A., & Jentsch, F. (2001). Use of commercial off-the-shelf, simulations for team research. In C. A.

Bowers & E. Salas (Eds.), Advances in human performance and cognitive engineering research (pp. 293-317).

Cain, W. S., de Wijk, R., Lulejian, C., Schiet, F., & See, L.-C. (1998). Odor identification: Perceptual and semantic dimensions. Chemical Senses, 23(3), 309-326.

Chu, S., & Downes, J. J. (2000). Odour-evoked autobiographical memories: Psychological investigations of Proustian phenomena. Chemical Senses, 25(1), 111-116.

Danthiir, V., Roberts, R. D., Pallier, G., & Stankov, L. (2001). What the nose knows: Olfaction and cognitive abilities. Intelligence, 29(4), 337-361.

Degel, J., Piper, D., & Koester, E. P. (2001). Implicit learning and implicit memory for odors: The influence of odor identification and retention time. Chemical Senses, 26(3), 267-280.

Herz, R. S. (2000). Verbal coding in olfactory versus nonolfactory cognition. Memory and Cognition, 28(6), 957-964.

Herz, R. S., & Engen, T. (1996). Odor memory: Review and analysis. Psychonomic Bulletin & Review, 3(3), 300-313.

Kole, A., Snel, J., & Lorist, M. M. (1998). Caffeine, morning-evening type and coffee odor: Attention, memory search and visual even related potentials. In J. Snel & M. M. Lorist (Eds.), Nicotine, caffeine and social drinking: Behaviour and brain function (pp. 201-214). Amsterdam: Harwood Academic.

Larsson, M. (1997). Semantic factors in episodic recognition of common odors in early and late adulthood: A review. Chemical Senses, 22(6), 623-633.

Lesschaeve, I., & Issanchou, S. (1996). Effects of panel experience on olfactory memory performance: Influence of stimuli familiarity and labeling ability of subjects. Chemical Senses, 21(6), 699-710.

Livermore, A., & Laing, D. G. (1996). Influence of training and experience on the perception of multicomponent odor mixtures. Journal of Experimental Psychology: Human Perception and Performance, 22(2), 267-277.

Parker, A., Ngu, H., & Cassaday, H. J. (2001). Odour and Proustian memory: Reduction of context-dependent forgetting and multiple forms of memory. Applied Cognitive Psychology, 15(2), 159-171.

Rabin, M. D. (1988). Experience facilitates olfactory quality discrimination. Perception and Psychophysics, 44(6), 532-540.

Raudenbush, B., Corley, N., & Eppich, W. (2001). Enhancing athletic performance through the administration of peppermint odor. Journal of Sport & Exercise Psychology, 23(2), 156-160.

Reber, A. S. (1995). Dictionary of Psychology (2nd ed.). London, England: Penguin Books.

Schab, F. R. (1991). Odor memory: Taking stock. Psychological Bulletin, 109(2), 242-251.

White, T. L., & Treisman, M. (1997). A comparison of the encoding of content and order in olfactory memory and in memory for visually presented verbal materials. The British Journal of Psychology, 88, 459-472.

Wickens, C. D., & Hollands, J. G. (2000). N. Roberts & B. Webber (Eds.), Engineering psychology and human performance (3rd ed.). Upper Saddle River, New Jersey: Prentice Hall.

Wood, J. T., & Eddy, R. M. (1996). Filtrates & residues: Olfactory tritation. Journal of Chemical Education, 73(3), 257-258.

HUMAN FACTORS OF INTELLIGENCE ANALYSIS

Malcolm J. Cook, Carol S. Angus, and Corinne S.G. Adams
Centre for Usability Test and Evaluation,
University of Abertay Dundee,

ABSTRACT

Kuhns (2003) has identified intelligence failures as one of the highly developed areas of academic study of intelligence. Other reviews of intelligence have supported the view that failures are associated with potentially consistent social and psychological factors as their contributors (Herman, 2002). It is proposed that there are significant ways to improve the use of intelligence analysis in achieving significantly improved results with limited data that is multi-source, multi-attribute, and possesses dubious validation criteria. This paper discusses a more detailed analysis of why experts with knowledge of the critical issues fail to deliver the correct analysis of all-source intelligence material. A current study on intelligence processes is described in terms of the suitability of the methodological approach used.

Keywords: Intelligence, Errors, Socio-Cognitive Processes, Situation Awareness

INTELLIGENCE

Kuhns (2003) has identified intelligence failures as one of the most developed areas of academic study of intelligence and Herman (2002) has suggested that consistent factors contribute to the occurrence of the failures. Intelligence failures can be analysed in a manner similar to accidents with a sequence of contributory causes leading up to significant events (Reason, 1990; 1997). Reason has proposed that any error or failure in system operation is normally not a result of a single cause but rather it is a consequence of a concatenation of errors that result in operational failure. To develop this approach and apply it to intelligence one needs to consider the stages in the intelligence process. Intelligence processes are normally segmented into collection, analysis and dissemination see figure 1 and 2 below outlining the intelligence process (Berkowitz and Goodman, 2000) with collection and analysis identified as problematic areas that contribute to intelligence failure (Herman, 2002; Kuhns, 2003). The emphasis for many agencies is naturally on superior collection (Combs, 2000) because there is a belief that this would diminish uncertainty associated with decision-making but it is argued that analysis is often weak. In the final analysis it is very unlikely that critical elements of the intelligence picture would be captured and as a consequence intelligence will always rely upon an incomplete, uncertain and confused image of the operational environment. The investigative guesswork of actual operations is well captured in Baer's (2002) book that describes his pursuit of terrorists in the Middle East. While Baer was a in the Directorate of Operations and not the Directorate of Intelligence his insights as a field officer suggest that the *image* or assessment of the intelligence problems are rarely complete. In addition, Baer indicates a very important role for HUMINT as a special source and one of the most effective in corroboration.

Currently, intelligence analysis does not make use of effective information technology (Berkowitz and Goodman, 2000) and the system interface to the knowledge is weak in supporting searching. This is surprising as the information technology revolution has been identified as a potential revolution in military affairs (O'Hanlon, 2000; Hall, 2003) and it would be not unreasonable to expect that the same might be the case for intelligence operations. Indeed, some authors have specifically identified the information age as a unique opportunity for re-thinking the manner in which intelligence operations are conducted (Berkowitz and Goodman, 2000). The visibility of the intelligence failures has in recent years become something that has been a matter for Congressional Intelligence Committees in the U.S.A. because of the failures in intelligence predictions prior to the events of September 11th 2001 (Johnson, 1996; Posner, 2003). The problems with intelligence (Benjamin and Simon, 2002; Powers, 2002) were already a matter for subject debate before the release of US Governmental evidence and Congressional judgements. The failure of intelligence services to grasp what was a fairly clear footprint, if somewhat diverse (see Gunuratna, 2002), for Al-Quaeda was identified in more popular reviews of intelligence function (Farren, 2003). The tactical surprise of the Al-Quaeda attacks can be set along side other attacks like that on Israeli athletes at the Olympic Games 1972 and the Aum Shinri Kyo gas attacks on the Tokyo underground (Murakami, 1997; Henderson, 2001), even though the scale of the assault by Al-Quaeda was far greater. With more information available in the public domain it has been made clear that a significant body of information existed and further data collection would only have corroborated the potential method of attack, place of attack and time of attack (see Fouda and Fielding, 2003) indicating a post-collection failure in analysis or dissemination. The links between individuals and Al-Quaeda were obvious, as indicated by associates that were caught and imprisoned (Moussaoui, 2002) and Fouda and Fielding's (2003) account. These failures in insight strongly support the view that there was a failure to exploit intelligence in an information age knowledge management system that suggests that the proposals for more effective processes designed to exploit information technology (Berkowitz and Goodman, 2000) have largely been ignored. The body of evidence on the attackers was sufficient to introduce measures that would have mitigated and pre-empted the attacks, even though the organisation was not attacked. The arrogance with which the Al Quaeda forces were viewed may be a contributory factor in the intelligence analysis. Arrogant or dismissive assessments of enemy forces have contributed to military operational failures in the past and they are still a frequent occurrence even though the technology of intelligence has changed (Regan, 2000; Keegan, 2003). The success of the attacks on the African Embassies should have been a viewed as a prelude to the attacks mainland U.S.A.. In the same manner the recent attacks on Spanish targets in March 2004, are a further indication of terrorist intent and capability.

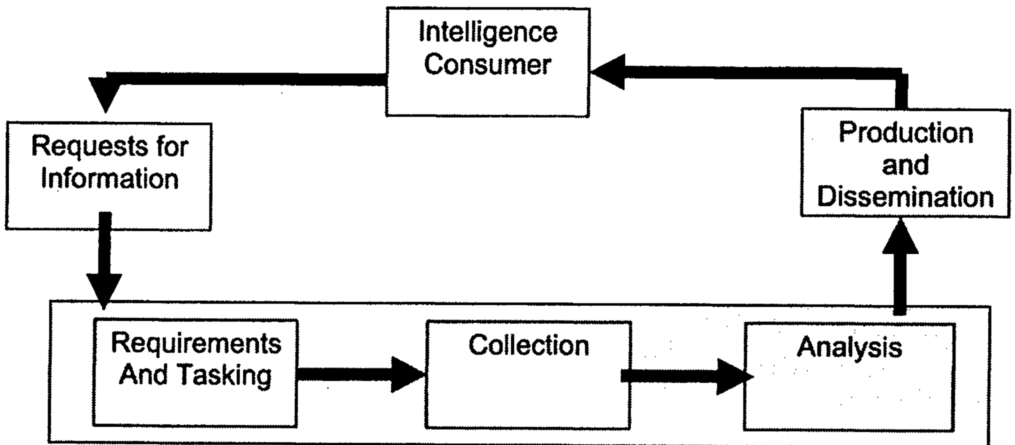


Figure 1: Intelligence cycle after Berkowitz and Goodman (2000)

Intelligence failures are not new and the frequent comparison between the events of 9-11 to Pearl Harbour has some basis in fact. It has been suggested that 9-11 was only a tactical surprise. It was recognised that cooperation between organisations and within organisations was weak in fusing this intelligence that was reminiscent of the failures prior to Pearl Harbour (McNeilly, 2001). Even if the information was made available in a single organisation it is likely that the thematic linkages between the individual items of information could not have been successfully exploited as a consequence of procedural, technological and organisational limitations (Benjamin and Simon, 2002). In an era of global terrorism it is necessary to overcome these difficulties. The financial and economic impact of 9-11 has been global and strategic with the airline industry the most visible casualty so that the surprise attacks on 9-11 should not be dismissed. Intelligence failures at Pearl Harbour resulted from critical areas of information capture that were neither exploited nor circulated to effectively exploit the critical information. There are many psychological issues involved in effective exploitation of intelligence that are critical in developing projection situation awareness based on uncertain, contradictory and incomplete information sources. The management of uncertainty in intelligence is a key issue in the continuing war on terrorism.

The intelligence services require a sophisticated group of knowledge workers able to collate, analyse and interpret complex patterns of information to make predictions about the future course of events (Hulnick, 1999). The intelligence services need to transfer their knowledge to other groups and this multi-agency collaboration is used to create policy and justify actions (Hulnick, 1999). Thus, there is a need to store information in a manner that a specialised community can use it but in a way in which it can easily be transformed into a format that is easily assimilated by other agencies, where cooperation is required. Herman (2002) notes the vast majority of intelligence failures are associated with various types of human factors issues in which the role played by the individuals within the intelligence community, with regard to failure, is critical.

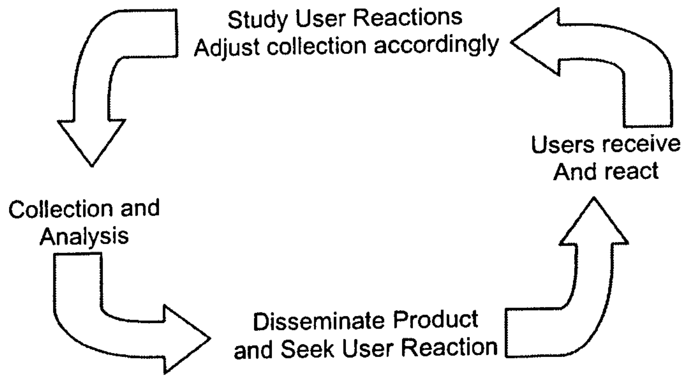


Figure 2: A simplified outline of the intelligence process

Psychological models have been used previously in evaluating the risk of bias in intelligence preparation (Cremeans, 1971; Heuer, 1978) but organisational, technological and economic factors have radically re-shaped intelligence services and processes in the period of time following these initial investigations during the cold war. Herman (2002) uses dated models of human psychological process to explain the mistakes observed in intelligence and it is not clear if the same types of error will propagate into future intelligence operations dominated by information technology and organisational change. A more detailed analysis by appropriately qualified human factors and domain experts could provide valuable insights to enhance the transitional process because of the wide range of social and cognitive issues associated with the use of information technology as a knowledge mediating system. Currently participant observation and ethnographic studies are taking place to determine what processes shape the intelligence process and how it might be improved. Initial reports suggest that a combination of social and cognitive issues might critically determine intelligence performance in a manner that is broadly similar to military command intelligence functions (Macklin, Cook, Angus, Adams, Cook, and Cooper, 2002). It would be useful to develop and validate a socio-cognitive model of intelligence functions using a combination of observational and empirical research based on quantitative and qualitative measures.

It is generally recognised that many information search technologies currently operate poorly because the user is not able to apply their conceptual understanding of the domain of interest via the interface. Thus, the current knowledge warehouses may not structure or collect knowledge in a manner that meets the needs of intelligence functions (Odom, 2003). In combination with potential information overload this will result in inefficient use of critical information. Thus, the aim is to develop a knowledge structure that enables a novel type of interface, which is intended to support conceptual appreciation of the information held as knowledge. In particular, it is proposed that a narrative structure be used to organise information into a coherent package of intelligence. Intelligence functions are used in a wide variety of governmental, commercial and institutional environments but each user group has a diverse range of operational uses. The requirements analysis proposed would aim to consider intelligence specifically applied to terrorism because of the diverse range of sources used to derive the intelligence picture and the uncertainties associated with information sources, content and interpretation.

The events of September 11th 2001 created significant concerns about the work of intelligence agencies and their ability to effectively process available information to accurately predict intent and actions of terrorist organisations (Betts, 2002; Pettiford and Harding, 2003). Information is not equivalent to knowledge and this was clearly illustrated by the events of September 11th. The production of knowledge in specific areas requires knowledge and meta-knowledge to infer what is a realistic interpretation of the information available. Knowledge is crucially important in intelligence. As Shulsky and Schmitt (2002) note intelligence refers to the creation of knowledge, by an organisation and through an activity, with knowledge creation at the core of that process. Knowledge creation in intelligence is divided into three parts, collection, analysis and dissemination. It is generally recognised that failures occur in intelligence analysis (Berkowitz and Goodman, 2000; Carter, 2001; Herman, 2001a; Herman, 2002; Odom, 2003) and there are many reasons to suspect that this may reflect cognitive limitations of operators, social factors shaping the handling of data and technological limitations in supporting the process. Currently the empirical evidence in the area is scant because of the severe controls over access to the operational environment. The process of managing intelligence information has been revolutionised by the sheer volume of information that can be collected and submitted for analysis from secret and open source media (Shulsky and Schmitt, 2002; Treverton, 2001; Berkowitz, 2003). Electronic management of information has in turn revolutionised the dissemination of information (Sharfman, 1996) making the propagation of inappropriate interpretations more problematic and potentially resulting in conservative estimates. Herman (2002) has identified a number of issues with direct bearing on intelligence which in turn relate to psychological and social aspects of information sharing and usage. In intelligence a delicate balance must be struck between revealing information in aiding the process of collection and guarding intelligence to protect the sources of information. If one accepts that the ebb and flow of information may vary in speed and quality the level of shared situational awareness amongst the potential users will vary. Allowing for retention of information at one time and rapid sharing of information at other critical times a new format of information storage must be created. The danger in using technology alone to solve the problem is the ability to create large warehouses of information that are inaccessible, unintelligible and unusable. Two issues should be considered with regard to an intelligence warehouse. First the ease of using the methods for encoding and retrieving information to develop intelligence briefs needs considered. It has been suggested that the development of intelligence briefings is a major performance indicator in the community and a significant factors in career progression. It might be assumed that this would produce higher quality output but is more likely that this will polarise inputs into conservative estimates producing no surprises or exaggerated estimates that will never be qualified by experience. The evidence from history suggests that both types of failure have occurred in the past. Second, the appropriateness of the knowledge structure, implicit in an interface to an intelligence warehouse, will be considered with regard to the conceptual requirements of intelligence. Previous work with high-level decision makers in command and control teams (Macklin et al., 2002) suggests that it may be possible to construct more effective interfaces by using a conceptual structure derived from critical incident debriefing of practitioners (Macklin et al., 2002). Critical incident debriefing has been used successfully in human factors research to acquire knowledge structure information for use in system design (Klein, 2000b).

One candidate knowledge structure for effective storage and retrieval is a narrative or storyboard format that inter-relates level 1 SA (perception of events), with level 2 SA (comprehension or interpretations of events), and level 3 SA (prediction of future events). The codification of information in terms of these levels of situational awareness and in terms of a narrative format (with temporal and spatial codes) allows agent-based representation of searches and inquiries to be executed on behalf of human operators on a continual basis, by other human and computer-supported agents. Thus, a new format for information storage and retrieval could simultaneously improve encoding of information, subsequent retrieval, re-use of information by other agencies and integration of all-source intelligence material into a single integrated framework. These improvements in intelligence functions have been considered by a number of authors (Berkowitz and Goodman, 2000; Treverton, 2001) as a result of the open-source availability of information and the information revolution. The events of September 11th made clear that intelligence lapses needed further investigation to understand the mechanisms and processes that had failed to capture and use the relevant information that was available after the events (Herman, 2001b).

FUTURE IMPROVEMENTS IN INTELLIGENCE

Human factors approaches to the development of computer supportive technology, in decision-aiding and information analysis have developed rapidly over the last fifteen years. There is now a need for more sophisticated performance measures for evaluation of the technology and theoretical models to help conceptualise design problems. One aim of this research is to identify human factors models suitable for application in the field of intelligence gathering and knowledge creation. One of the key models applied to individual cognition is the model of situational awareness (Endsley, 2000). Situation awareness consists of three components, level 1 (perception of events), level 2 (comprehension of the meaning of events), and level 3 (projection of future events based on current understanding). This model can be applied to descriptions of the technology, systems and processes for intelligence to determine if the emphasis in current intelligence is weighted towards supporting level 1 Situation Awareness (SA), the perception of events. Current analyses of intelligence functions suggest that intelligence information collection is adequate but the analysis of information is not. This observation is in direct contrast to situation awareness errors in real-time systems management, where the failures are usually related to missing significant events. If one accepts that the cognitive weighting of current systems inadequately supports the development of level 2 or 3 situational awareness it is easy to interpret the shortcomings with regard to recent terrorist incidents. Retrospective analysis of the events leading up to September 11th indicate that significant clues existed from a number of sources that identified an airborne threat to a limited number of U.S. mainland targets (Hawthorne, 2002; Posner, 2003). There is an obvious hindsight bias in the interpreted significance of the cues but it seems likely that this type of operation will be a regular feature of terrorist actions in the future that needs to be guarded against. Thus, intelligence processes, technology and systems should be designed to make better use of this type of construct to develop insight based on uncertain data.

One approach taken from the applied psychology literature relates to the manner in which decision-making processes occur, where it is suggested that decision-making is more correctly described as a pattern recognition process where environmental cues are associated with schematic knowledge of previous events. This process of recognition-primed decision-making (Klein, 1993b) (also termed naturalistic decision making by Gary Klein (1993a)) has been used to aid the designers of new information management systems in real-time control systems. It is likely that the same models of decision-making, given their reliance on knowledge (explicit and implicit) and on expertise are applicable to the intelligence community operators. While many knowledge workers do not consider themselves decision-makers their role as filters of information and intelligent observers of events has strong similarities to the properties of decision-makers in command and control. The information management process is essentially a socio-technical filtering operation whereby the information deluge is narrowed and shaped into a manageable stream of relevant data. This process of narrowing is subject to type 1 and type 2 errors of marking as relevant irrelevant information or discarding irrelevant information that is actually relevant. In addition, intelligence operations must manage attempted decoys, deceptions and bluffs.

Human factors research has identified useful methodologies for the development of new technology called cognitive task analysis or cognitive work analysis (see Chipman, Schraagen and Shalin. 2000; Vicente, 1999; Hollnagel, 2003). While not true equivalents both methodologies have been successful in gaining insight into complex socio-cognitive technologies where individual and group psychology factors influence performance. Cognitive task analysis is well described by Chipman, Schraagen and Shalin (2000) who suggest that it is an extension of traditional task analytic techniques to include information about knowledge, thought processes, and goal structures that underlie observable task performance. Thus, it is clearly applicable in an area such as intelligence operations, which involves the use of knowledge and critical thinking to create the intelligence product. Cognitive work analysis attempts to understand the nature of the operational domain by attempting to identify the semantics of the relevant domain (Vicente, 1999). In simple terms work only makes sense within a context and abstract representations of work can create misleading indications for system developers and process management. It has been argued that work analysis is an important method for developing computer-based systems that effectively supports human work within a complex socio-technical system. Again the emphasis with these modern approaches is not description but explanatory appreciation of what work is done, the demands on the human operator and how they are best supported. Recent reviews of intelligence have already identified the significance of the analysis process and of the information revolution in intelligence there is clearly a need to appreciate the nature of the work with an appropriate methodology, such as Cognitive Work Analysis (CWA) or Cognitive Task Analysis (CTA). Similar concerns are found in Wieck's (2001) work on making sense in complex socio-technical organisations because sense-making emphasises both the social and cognitive elements of the cooperative enterprise. The significance of social context, personal identity, salient cues, ongoing projects, plausibility, and enactment, can be easily identified in intelligence communities. Indeed, there is no reason to expect intelligence operations to be sterile because the human and organisational factors will cause the process to deviate from optimal function. Historically it has been found that governments can influence the craft, individuals can undermine the process with malicious intent or as a way of influencing their career progression and theories of enemy intent can be upheld in the face of incontrovertible and antagonistic evidence. Any analysis of intelligence can only explain a proportion of the data if it does not address the multi-facetted web of influence on the process.

To understand the human factors issues in intelligence it is necessary to outline the steps whereby information makes sense and information is dismissed from the system. Most models of human cognition propose three major types of memory, a very short-term sensory memory that gives us access to all the environmental information, a much more limited short-term or working memory in which information is processed and a long-term memory that retains all the products of experience. The capacity, speed and organisation of each type of memory are different and this shapes the way in which information is processed. Working memory is relatively small and the main danger is information overload where the amount of information exceeds the capacity of the memory. Working memory is critical because effective processing of information results in transfer of processed information to long-term memory and the development of experience (Carlson, 1997). Long-term memory is much slower to access and a major problem is retrieval, where information is available but inaccessible. Long-term memory does not have capacity problems but humans can mislay information, failing to retrieve information. Access to long-term memory can change in expert individuals but only when the information accessed is repeatedly and exhaustively used, under these conditions expertise is highly limited and situation specific (Ericsson and Delaney, 1999; Proctor and Dutta, 1995). This is why the long-term analysis of the Soviet threat was much easier to manage than the highly volatile terrorist threats in recent times. It is clear that even after short periods of training intelligence analysts will change their methods for processing information and the type of structure they impose on the knowledge. However, their real information processing sophistication may be the meta-knowledge about which sources, which type of information and what types of corroborative evidence which is ae likely to be significant in specific analyses.

Having considered briefly the ways in which the different elements of memory inter-relate one might consider why a human analyst is considered more appropriate than machine intelligence. First, reason is the sparse nature of the information in intelligence that requires conjectural developments using experience beyond the scope of current inferential logic driven by machine intelligence. Second, the presence of misleading information in the database designed to draw attention away from or mask the intent of the group under scrutiny. Third, the consideration of intangible and qualitative qualifications of the sources, methods and coverage of the information collected. The accomplished intelligence analyst needs to use implicit knowledge of the information, often described as gut instinct, to qualify the judgements made. This is strength and weakness of intelligence preparation by human analysts because feelings of uncertainty associated with complexity of the information can be confused with the interpretation of analysis, to produce an uncertain or qualified interpretation. Psychologists examining information processing strategies have suggested that affect is an integral part of how we manage the world and it impacts judgements and reasoning (Bower and Forgas, 2000; Forgas, 2000). Accepting that this is the case technology should be designed to help the user explore their uncertainties and to protect against errors of judgement driven by decision-related anxiety. However, the need for certainty, to sanction actions, and the uncertain nature of the judgements in intelligence represents a conflict that is intrinsic to the process and would not be eliminated completely by the use of technology. Thus, the solution requires training, technology and processes to prevent erroneous judgement. What makes the area of intelligence somewhat unique is the focus largely on the support of interpretative analysis on information to generate knowledge or comprehension without some form of direct or immediate feedback from the real world. In effect the plausibility or accuracy of the model proposed is unknown at least until further events occur and further evidence is accrued, as such it resembles science in only finding supporting evidence that is relatively accurate and not absolute evidence that is unquestionable. Intelligence analysis is an open system and as such it is important to develop metrics which assess both the process and the product of intelligence activity, as the value of the latter may never be totally without doubt.

The focus of any research program on intelligence should be geared towards the practical implementation of an improved intelligence process by socio-cognitive improvements in information sharing techniques. An appropriate research program would enable an appreciation of culture and its impact in intelligence circles, as it has been suggested that this may be destructive and undermine the exploitation of new technology (Berkowitz and Goodman, 2000). Some attempt should be made to understand the organisational culture as a factor influencing work-related activities and for this reason the type of interpretative analysis used by Wieck (2001) and the work analysis approach (Vicente, 1999) should be used. Some consideration of the more detailed issues in collaborative and coordinated working mediated by computer (see Olson, Malone and Smith, 2001) have been examined in the computer science literature but many of the studies conducted have failed to look at mature organisations with subject matter experts, typical of intelligence services.

In conclusion, the time has come for the revolution in information technology to be developed to meet the requirement of the intelligence services more adequately than currently is the case. A simple technological fix will not improve the analysis process because there is currently a knowledge gap with regards to the actual process. A superficial and subject-matter led analysis has not taken the process far and the absence of a human factors approach to analysing and aiding the intelligence process will mean that future attempts at improvement are more likely to fail. In recognising that intelligence is knowledge craft but accepting that knowledge is not impartial, and the processes creating it are influenced by a myriad of causes, one accepts the central role of the human operator. Machines do not think and currently do not discern intent it is the human operator that must do this. As intelligence operations against terrorism is the discernment of intent then human issues are the key to any future improvements.

REFERENCES

Benjamin, D. and Simon, S. (2002) A failure of intelligence? In R.B. Silvers and B. Epstein (Eds.) *Striking terror: America's new war.* New York, USA: New York Review Books.

Berkowitz, B. (2003) *The new face of war: How war will be fought in the 21st century.* London: Free Press.

Berkowitz, B.D. and Goodman, A.E. (2000) *Best truth: Intelligence in the information age.* London: Yale University.

Betts, R.K. (2002) Fixing intelligence. *In The War on Terror.* New York: Foreign Affairs.

Bower, G.H., and Forgas, J.P. (2000) Affect, memory and social cognition. In E. Eich, J.F. Kihlstrom, G.H. Bower, J.P. Forgas, and P.M. Niedenthal (Eds.) *Cognition and Emotion.* Oxford: Oxford University Press.

Boyatzis, R.E. (2000) Transforming qualitative information. London: Sage Publishing.

Carlson, R.A. (1997) Experienced cognition. London: Lawrence Erlbaum Associates.

Carter, A.B. (2001) Keeping the Edge: Managing defense for the future. Carter, A.B., and White, J.P. (Eds.) *Keeping the Edge: Managing defense for the future.* Oxford: MIT Press.

Combs, C.C. (2000) *Terrorism in the twenty-first century.* Upper Saddle River, N.J.: Prentice Hall.

Cremeans, C.D. (1971) Basic psychology for intelligence analysts. Studies in Intelligence 15(1).
Reproduced in H. Bradford Westerfield (Ed.) *Inside the CIA's private world* (1995).

Davenport, T.H. and Prusak, L. (2000) Working knowledge. Boston, USA: Harvard Business School Press.

Davies, N. (2003) *Terrorism: Inside a world phenomenon.* London: Virgin Books.

Endsley, M.R. (2000) Theoretical underpinnings of situational awareness a critical review. In M.R. Endsley and D.J. Garland - *Situation awareness analysis and measurement.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Ericsson, K.A., and Delaney, P.F. (1999) Long-term working memory as an alternative to capacity models of workling memory in everyday skilled performance. In A. Miyake and P. Shah (Eds.) *Models of working memory.* Cambridge: Cambridge University Press.

Farren, M. (2003) CIA: *The secrets of the "company".* London: Chrysalis Press.

Forgas, J.P. (2000) Affect and information processing strategies. In J.P. Forgas (Ed.) *Feeling and thinking: The role of affect in social cognition.* Cambridge: Cambridge University Press.

Fouda, Y, and Fielding, N. (2003) *Masterminds of terror.* London: Mainstream Publishing.

Gunuratna, R. (2002) *Inside Al Qaeda: Global Network of terror.* London: Hurst and Company.

Hall, W.M. (2003) *Stray voltage: war in the information age.* Annapolis, Maryland, U.S.A.: Naval Institute Press.

Hawthorn, S. (2002) Knowledge related to purpose: Data-mining to detect terrorism. (unpublished article).

Heath, C., and Luff, P. (2000) *Technology in Action.* Cambridge: Cambridge University Press.

Henderson, H. (2001) *Global terrorism: The complete reference guide.* New York: Checkmark Books.

Herman, M. (2001a) Keeping the edge in Intelligence. Carter, A.B., and White, J.P. (Eds.) *Keeping the Edge: Managing defense for the future.* Oxford: MIT Press.

Herman, M. (2001b) *Intelligence services in the information age.* London: Frank Cass Publishers.

Herman, M. (2002) *Intelligence power in peace and war.* Cambridge: Cambridge University Press.

Hollnagel, E. (Ed.) (2003) *Handbook of cognitive task design.* London: Lawrence Erlbaum Associates.

Huer, R.J. (1978) Cognitive biases: Problems in hindsight analysis. Studies in Intelligence 22(2) 21-28. Reproduced in H. Bradford Westerfield (Ed.) *Inside the CIA's private world* (1995).

Johnson, L.K. (1996) *U.S. intelligence in a hostile world: Secret agencies.* London: Yale University Press.

Keegan, J. (2003) *Intelligence at war: Knowledge of the enemy from Napoleon to Al-Qaeda.* London: Hitchinson Press.

Klein, G. (1993a) Naturalistic decision making. Wright Patterson Air Force Base, Ohio: Crew Systems Ergonomics Research Information Analysis Centre (CSERIAC).

Klein, G. (1993b) A recognition-primed decision (RPD) model of rapid decision making. In G.A. Klein, J. Orasanu, R. Calderwood, and C.E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 138-147). Norwood, NJ: Ablex.

Klein, G. (1997a) An overview of naturalistic decision making applications. In C.E. Zsambok and G. Klein, *Naturalistic Decision Making* (pp. 49-59). Mahwah, N.J.: Lawrence Erlbaum Associates.

Klein, G. (1997b) The current status of naturalistic decision making framework. In R. Flinn, E. Salas, M. Strub, and L. Martin (Eds.) *Decision making under stress: Emerging themes and applications.* Aldershot: Ashgate Publications.

Klein, G. (2000a) Cognitive Task Analysis of Teams. J.M., Schraagen, S.F., Chipman, and V.L., Shalin, (Eds.) *Cognitive Task Analysis (p. 417-430).* London: LEA.

Klein, G. (2000b) Analysis of situation awareness from critical incident reports. In M.R. Endsley and D.J. Garland - *Situation awareness analysis and measurement.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kuhns, W.J. (2003) Intelligence failures: Forecasting and the lessons of epistemology, In R.K. Betts and T.G. Mahnken *Paradoxes of Strategic Intelligence.* London: Frank Cass Publishers Ltd.

Luff, P., Heath, C., and Greatbatch, D. (1994) Work, interaction and technology: The naturalistic analysis of human conduct and requirements analysis. In M. Jirokta and J. Goguen (Eds.) *Requirements Engineering: Social and Technical Issues.* London: Academic Press.

Macklin, C. M., Cook, M. J., Angus, C. S., Adams, C. S. G., Cook, S. and Cooper, R. (2002) Qualitative analysis of visualisation requirements for improved campaign assessment and decision making in command and control. In C. Johnson (Ed) 21st *European annual conference on Human decision making and control: GIST Technical Report G2002-1* (pp 169 - 177). University of Glasgow.

McNielly, M. (2001) *Sun Tzu and the Art of Modern Warfare.* Oxford, UK: Oxford University Press.

Moussaoui, A.S. (2002) *Zacarias Moussaoui: The making of a terrorist.* London: Serpent's Tail Publishing.

Murakami, H. (1997) *Underground: The Tokyo gas attack and the Japanese psyche.* London: Harvill Press.

Odom, W.E. (2003) *Fixing intelligence: For a more secure America.* London: Yale University Press.

O'Hanlon, M. (2000) *Technological change and the future of warfare.* Washington D.C., USA: Brookings Institution Press.

Olson, G.M., Malone, T.W., and Smith, J.B. (Eds.) (2001) *Coordination theory and collaboration technology.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Pettiford, L. and Harding, D. (2003) *Terrorism: The new world war.* Slough, UK: Arcturus.

Posner, G. (2003) *Why America slept: The failure to prevent 9/11.* New York: Random House.

Powers, T. (2002) The trouble with the CIA, In R.B. Silvers and B. Epstein *Striking Terror.* New York: New York Review Books.

Proctor, R.W. and Dutta, A. (1995) Skill acquisition and human performance. London: Sage Publications.

Reason, J. (1990) *Human error.* Cambridge: Cambridge University Press.

Reason, J. (1997) *Managing risks of organizational accidents.* Aldershot: Ashgate Publications.

Regan, G. (2000) *Great military blunders.* Basingstoke: Channel 4 Books (Imprint of Macmillan).

Robson, C. (2002) *Real world research, 2nd Ed.* Oxford: Blackwells.

Schraagen, J.M., Chipman, S.F., and Shalin, V.L. (2000) Introduction to cognitive task analysis. In J.M. Schraagen, S.F. Chipman, and V.L. Shalin, (Eds) *Cognitive task analysis.* London: Lawrence Erlbaum Associates.

Sharfman, P. (1996) Intelligence analysis in an age of electronic dissemination. In D.A. Charters, S. Farson, and G.P. Hastedt (Eds.) *Intelligence analysis and assessment.* London: Frank Cass Publishers.

Shulsky, A.N., and Schmitt, G.J. (2002) *Silent warfare: Understanding the world of intelligence.* London: Brassey's Inc.

Treverton, G.F. (2001) *Reshaping national intelligence in the information age.* Cambridge: Cambridge University Press.

Vicente, K.J. (1999) *Cognitive work analysis.* London: Lawrence Erlbaum Associates.

Wieck, K.E. (2001) *Making sense of the organisation.* Oxford: Blackwell.