DEFENSE TECHNICAL INFORMATION CENTER

Information for the Defense Community

DTIC® has determined on 05/24/2010 that this Technical Document has the Distribution Statement checked below. The current distribution for this document can be found in the DTIC® Technical Report Database.

☒ **DISTRIBUTION STATEMENT A.** Approved for public release; distribution is unlimited.

☐ **© COPYRIGHTED**; U.S. Government or Federal Rights License. All other rights and uses except those permitted by copyright law are reserved by the copyright owner.

☐ **DISTRIBUTION STATEMENT B.** Distribution authorized to U.S. Government agencies only (fill in reason) (date of determination). Other requests for this document shall be referred to (insert controlling DoD office)

☐ **DISTRIBUTION STATEMENT C.** Distribution authorized to U.S. Government Agencies and their contractors (fill in reason) (date of determination). Other requests for this document shall be referred to (insert controlling DoD office)

☐ **DISTRIBUTION STATEMENT D.** Distribution authorized to the Department of Defense and U.S. DoD contractors only (fill in reason) (date of determination). Other requests shall be referred to (insert controlling DoD office).

☐ **DISTRIBUTION STATEMENT E.** Distribution authorized to DoD Components only (fill in reason) (date of determination). Other requests shall be referred to (insert controlling DoD office).

☐ **DISTRIBUTION STATEMENT F.** Further dissemination only as directed by (inserting controlling DoD office) (date of determination) or higher DoD authority.

Distribution Statement F is also used when a document does not contain a distribution statement and no distribution statement can be determined.

☐ **DISTRIBUTION STATEMENT X.** Distribution authorized to U.S. Government Agencies and private individuals or enterprises eligible to obtain export-controlled technical data in accordance with DoDD 5230.25; (date of determination). DoD Controlling Office is (insert controlling DoD office).

Proceedings of an
International Conference

Daytona Beach, Florida USA
November 1-3, 1995



Experimental Analysis and Measurement of Situation Awareness

Edited by

Daniel J. Garland & Mica R. Endsley

Sponsored by

U.S. Federal Aviation Administration
U.S. Nuclear Regulatory Commission
Embry-Riddle Aeronautical University
Texas Tech University

Experimental Analysis and Measurement of Situation Awareness

Experimental Analysis and Measurement of Situation Awareness

Edited by

Daniel J. Garland, Ph.D.
Center for Aviation/Aerospace Research
Embry-Riddle Aeronautical University
Daytona Beach, Florida

Mica R. Endsley, Ph.D., P.E.
Department of Industrial Engineering
Texas Tech University
Lubbock, Texas

20100311130

Embry-Riddle Aeronautical University Press
Daytona Beach, Florida USA

Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness, Daytona Beach, FL, November 1–3, 1995.

Notice

This document is disseminated under the sponsorship of the U.S. Federal Aviation Administration, the U.S. Nuclear Regulatory Commission, Embry-Riddle Aeronautical University, and Texas Tech University in the interest of information exchange. Opinions expressed in this document do not necessarily reflect official policy or approval of the sponsors or the United States Government. The United States Government assumes no liability for the contents or use thereof.

The United States Government does not endorse products or manufacturers. Trade or manufacturer's names appear herein solely because they are considered essential to the object of this document.

Technical Editing by

Mark A. Wise
Center for Aviation/Aerospace Research
Embry-Riddle Aeronautical University
Daytona Beach, Florida

Graphic Designs by

Nancy Rahn-Virden
Embry-Riddle Aeronautical University
Daytona Beach, Florida

For information contact:
Center for Aviation/Aerospace Research
600 South Clyde Morris Blvd.
Daytona Beach, FL 32114-3900

Sponsored by:

- **U.S. Federal Aviation Administration**
- **U.S. Nuclear Regulatory Commission**
- **Embry-Riddle Aeronautical University**
- **Texas Tech University**

Table of Contents

Preface xii

Acknowledgments xiii

Introduction

Situation Awareness Measurement and Analysis:
A Commentary 1
Charles E. Billings

The State of Situation Awareness Measurement: Circa 1995 7
Richard W. Pew

Theoretical Underpinnings of Situation Awareness: A Critical Review 17
Mica R. Endsley

Maintaining Situation Awareness when Stalking Cognition in the Wild 25
John M. Flach

Expert Performance and Situation Awareness 35
Neil Charness

Experimental Analysis and Measurement of Situation Awareness:
A Commentary 43
David Meister

Situation Awareness: A Cognitive Neuroscience Model
Based on Specific Neurobehavioral Mechanisms 49
Robert S. Kennedy and J. Mark Ordy

Performance Based Measurement Techniques

The Tradeoff of Design for Routine and Unexpected Performance:
Implications of Situation Awareness 57
Christopher D. Wickens

Performance Measures and Situational Awareness:
How Strong the Link? 65
John M. Reising

The Role of Scope as a Feature of Situation Awareness Metrics 69
Michael A. Vidulich

Use of Testable Responses for Performance-Based
Measurement of Situation Awareness 75
A. R. Pritchett, R.J. Hansman and E.N. Johnson

Subjective Measurement Techniques

Experiential Measures: Performance-Based
Self Ratings of Situational Awareness 83
R.M. Taylor

Using Observer Ratings to Assess Situational Awareness
in Tactical Air Environments 93
Herbert H. Bell and Wayne L. Waag

SA Measurement: Lessons Learned from Workload 101
Gary B. Reid

Query Techniques

Direct Measurement of Situation Awareness in Simulations of
Dynamic Systems: Validity and Use of SAGAT 107
Mica R. Endsley

SACRI: A Measure of Situation Awareness for
Nuclear Power Plant Control Rooms 115
Stephen G. Collier and Knut Follesø

Measurement and Analysis of Situation Awareness
in Anesthesiology 123
Stephen D. Small

Team Situation Awareness Research: Many Paths to a Destination 129
Carolyn Prince, Eduardo Salas, Clint Bowers and Florian Jentsch

Situation Awareness: Team Measures, Training Methods 135
Carolyn Prince, Eduardo Salas and Renée J. Stout

Physiological Measurement Techniques

Psychophysiological Assessment of SA? 141
Glenn F. Wilson

Role of Volitional Effort in the Application of
Psychophysiological Measures to Situation Awareness 147
Evan A. Byrne

Physiological Measurement Techniques: What the Heart
and Eye Can Tell Us About Aspects of Situational Awareness 155
J. A. Stern, L. Wang, and D. Schroeder

Post-Hoc and Process Assessment Techniques

Post-Hoc Assessment of Situation Assessment
in Aircraft Accident/Incident Investigations 163
Barry Strauch

Air Traffic Controller Awareness of Operational Error Development 171
Mark D. Rodgers, Richard H. Mogford and Leslye S. Mogford

Studying Situation Awareness in the Context of
Decision-Making Incidents 177
Gary Klein

Accuracy Estimation in Situation Awareness Research 183
Richard H. Mogford

Air Traffic Control and Situation Awareness

Factors Characterizing En Rroute Operational Errors:
Do They Tell Us Anything About Situation Awareness? 189
Francis T. Durso, Todd R. Truitt, Carla A. Hackworth,
Daryl Ohrt, Janis, M. Hamic, Jerry M. Crutchfield,
and Carol A. Manning

Situational Awareness at Different Levels of Abstraction: The Distributed Cooperative
Problem-Solving Domain of ATCSCC-Airline Operations 197
C. Elaine McCoy, Judith Orasanu, Philip J. Smith,
Amy VanHorn, Charles Billings, Rebecca Denning,
Michelle Rodvold, and Theresa Gee

Measurement Of Air Traffic Controllers' Situation Awareness
and Performance During Simulator Training 203
Esa M. Rantanen and Joseph S. Butler

A Simulation Study of Air Traffic Controller Situational Awareness 211
Randy L. Sollenberger and Earl S. Stein

Aircraft Systems and Situation Awareness

Construct Validity of Situation Awareness
Measurements Related to Display Design 219
Oscar Olmos, Chia-Chin Liang, and Christopher D. Wickens

FliteScript: A Multimedia Test to Index
Situational Schemata in Pilots 227
Alan F. Stokes and Donna Wilt

Modeling and Measuring Situation Awareness for
Target-Identification Performance 233
Eileen B. Entin, Daniel Serfaty, and Elliot E. Entin

Complex Systems and Situation Awareness

Analyzing Situation Awareness During
Wayfinding in a Driving Simulator 245
Jack Beusmans, Vlada Aginsky, Catherine Harris, and Ronald Rensink

Evaluation of "RLMS" Automotive Rear Lighting 253
David L. Cameron

Comparing Explicit and Implicit Measures of Situation Awareness 259
Leo Gugerty and William Tirre

Cognitive Correlates of Explicit and Implicit
Measures of Situation Awareness 267
Leo Gugerty and William Tirre

Situation Awareness Evaluation for an Operator
Support System in a Nuclear Power Plant 275
Geert Uytterhoeven, Michel De Vlaminck and Denis Javaux

Individual and Team Issues in Situation Awareness

Evaluating Team Situation Awareness through Communication 283
Judith Orasanu

Situation Awareness and Older Workers 289
Cheryl A. Bolstad and Thomas M. Hess

Expertise and Chess: A Pilot Study Comparing
Situation Awareness Methodologies 295
Francis T. Durso, Todd R. Truitt, Carla A. Hackworth,
Jerry M. Crutchfield, Danko Nikolic, Peter M. Moertl,
Daryl Ohrt, and Carol A. Manning

Perspectives on the Appreciation of Team Situational Awareness 305
Iain S MacLeod, Robert M Taylor and Colin L Davies

A Methodology for Analyzing Team Situation
Awareness in Aviation Maintenance 313
Michelle M. Robertson and Mica R. Endsley

Posters

Aircraft Recognition Thresholds and Manual Attitude Control: Individual Performance
Links that Might be Important to Situation Awareness 321
Jeremy M. A. Beer, Robert A. Gallaway, & Fred H. Previc

The Virtual Patient – An Application of Situation Awareness
for System Design and Training in Intensive Care 329
S. Keith Adams, Shane P. Babin and Zeinab A. Sabri

An Assessment Of Situation Awareness in an Air Combat Simulation:
The Global Implicit Measure Approach 339
Bart J. Brickman, Lawrence J. Hettinger, Merry M. Roe, Dean Stautberg, Michael A. Vidulich, Michael W. Haas and Robert L. Shaw

Displays to Enhance Air Combat Situational Awareness 345
Gerald P. Chubb

Measuring Situational Awareness with the "Ideal Observer" 351
Marc Green, J. Vernon Odom, and J. Terry Yates

Towards a Robust, Quantitative Measure for Presence 359
Jerrold D. Prothero, Donald E. Parker, Thomas A. Furness III, and Maxwell J. Wells

The Effect of Task Automatisation in the Automotive Context:
A Field Study of an Autonomous Intelligent Cruise Control System 369
Nicholas J. Ward, Stephen Fairclough and Mark Humphreys

The Effect of Automotive Head-Up Displays on
Attention to Critical Events in Traffic 375
Nicholas J. Ward, Andrew Parkes and Phil Lindsay

Modeling Situation Awareness: Using the Influence Diagram 383
Joseph Sferrazza and Marc B. Wilson

Give Me Situation Awareness, or Give Me Death 407
Michael Eidelkind, Raymond Moffett, Don Arendt, and Charles McKee

Situation Awareness 415
Francis S. Bennett

Ergodynamics and its Application to the Work
Productivity and Cost Effectiveness of Ergonomic
Projects Implementation 423
Valery F. Venda and Ilona V. Venda

New workstations for High Productivity and Low Risk
of Occupational Injures: Design and Industrial Testing 439
Valery F. Venda and Ilona V. Venda

Preface

On behalf of the U.S. Federal Aviation Administration, the U.S. Nuclear Regulatory Commission, Embry-Riddle Aeronautical University, and Texas Tech University, it is our distinct pleasure to offer these proceedings as an enduring part of the international conference on *Experimental Analysis and Measurement of Situation Awareness*. The conference was held at the Adam's Mark Daytona Beach Resort in Daytona Beach, Florida, November 1-3, 1995 and attracted an assembly of the world's experts in the area of situation awareness. Approximately 200 professionals representing 12 countries attended the conference.

Situation awareness is currently a highly active area of research that has spread from the aviation community to impact on a variety of operational applications. The objective of the conference was to bring together researchers to critically evaluate the state-of-the-art in situation awareness measurement, discuss the conceptual and methodological benefits and inadequacies of different measurement approaches for different purposes and in different settings, and generate constructive recommendations needed for improving situation awareness measurement practices. The conference was the first to explicitly focus on the need for a rigorous examination of measurement techniques being used and proposed for work in this field.

The objectives of the conference were to (a) bring together professionals from a variety of disciplines to critically evaluate and discuss situation awareness research, (b) critically assess the state of situation awareness measurement, (c) discuss the conceptual and methodological benefits and inadequacies of different situation awareness measurement approaches for a variety of purposes in different types of settings, and (d) generate the constructive criticism necessary to push the state of knowledge forward by developing recommendations for improving situation awareness measurement practices.

The format of the conference was developed to encourage active and open participation by all attendees. The conference program included several invited presentations by distinguished researchers, six panel sessions, four paper sessions, a poster session, and tours of ERAU's campus and academic/research laboratories. There was also a reception which was held during the poster session on Wednesday evening (Nov 1) and a dinner, featuring Timothy P. Forté, former Director of the Office of Aviation Safety at the NTSB as the after-dinner speaker for Thursday evening (Nov 2).

These proceedings are based on the information disseminated and generated at the conference. The conference and the following papers are the first to specifically address the topic of situation awareness measurement and analyses, consequently serving a very important role in propelling the state-of-the-art in situation awareness measurement.

Daniel J. Garland
Mica R. Endsley

Acknowledgments

The development and publication of a volume of this nature is not without a tremendous amount of dedication and hard work from a group of quality individuals. The editors would like to acknowledge the work of those individuals and organizations who made the conference and publication of this volume possible.

We must thank our sponsors, without whom the conference could not have taken place. The sponsors for the conference included:

- U.S. Federal Aviation Administration
- U.S. Nuclear Regulatory Commission
- Embry-Riddle Aeronautical University
- Texas Tech University

We extend a fervent thank you to all the conference participants and authors for their enthusiastic contributions and for their boldness in addressing such difficult issues. We must thank an excellent conference staff who worked hard before, during, and after the meeting. The conference was managed with the outstanding and tireless work of Frances L. Cozza, the conference's Program Coordinator. Frances was bolstered by an exceptional staff, particularly Esin O. Kiris, S. Armida Rosiles, Debra G. Jones, and David B. Kaber of Texas Tech University, who were on site at the conference virtually day and night, and Mark A. Wise of Embry-Riddle Aeronautical University who contributed a yeoman's effort in compiling these proceedings. We owe a significant debt to Nancy Rahn-Virden for her outstanding graphics which significantly contributed to the conference program and proceedings. Finally, a heartfelt thank you goes to the graduate and undergraduate students who put in untold hours to meet the omnipresent on-the-spot needs of such a conference.

We are forever grateful for the support of these many individuals and sponsoring organizations in the publication of this volume.

Daniel J. Garland
Mica R. Endsley

Situation Awareness Measurement and Analysis: A Commentary

Charles E. Billings

The Ohio State University

"With what Sart of Sword shall we Swat the Saints?"

Introduction

Many years ago, in my second year of medical school, our very impressive and intimidating Viennese pathology professor addressed us just before the final examination. His lecture, in its entirety, was, "Ladies and Gentlemen: the questions on the examination are the same as last year. Only the answers have changed."

I have been reminded of his words many times during this conference. Most of the important questions raised here are indeed the same as those that confronted us in Orlando in February, 1993 (Gilson, Garland & Koonce, 1994). But my professor was right—some of the answers have changed. In this brief reprise of the Conference, I will attempt to evaluate how they have changed and which ones remain unanswered, or unanswerable, and I shall try to suggest why some of the questions are likely to confront us yet again if we meet again in June of 1998 in as pleasant surroundings as we have enjoyed here.

The Overarching Questions

Just what is situation awareness?

Dr. Meister (1995, this volume) defined situation awareness (SA) as a "lumping concept" rather than a "splitting concept". Either way, SA is an abstraction that exists within *our* minds, describing phenomena that we observe in humans performing work in a rich and usually dynamic environment. Pew (1995, this volume), Endsley (1995, this volume) and Meister all noted its two essential components: a *situation*, or state, of relevant variables in the external world, and the view, or *awareness* within the human operator, of that situation. Dr. Pew made an important distinction, both at Orlando (Pew, 1994) and here between *ideal, attainable* and *actual* situation awareness.

In the ideal case (which I suggest is not attainable because we cannot possess hindsight before the fact) there would be a perfect match between the real situation and the observer's mental model of that situation. The attainable case, which is never ideal, nonetheless represents the best level of awareness in a perfect observer who has assimilated all of the information available about the world state. This level serves as a benchmark against which we can measure actual situation awareness, which is what we fallible mortals actually have at any given moment.

Are such distinctions just academic meandering? Not at all; no more than Flach's (1995) elegantly stated cautions in his editorial in a special issue of *Human Factors*. Dr. Meister began his talk by asking whether the SA model could be applied; "if it cannot be applied, what good is it?" I will propose that if expert operators are found consistently to have actual SA that is inadequate compared to that which is attainable, we have either an information transfer problem due to inadequate processing and integration of information elements (or inadequate representations of those elements), or we have a training problem because we haven't taught operators how to interpret that information, or where to find it, or when they need it. These are design and training issues, and they are tractable, though rarely easy.

A more difficult, but still approachable, design problem arises when attainable situation awareness falls unacceptably short of the ideal. In this case, the information required is not available to the operator in a useful format. The situation may be so indeterminate as to defy description (as may occur during natural disasters or in certain chaotic systems), or it may be uncertain because we do not fully understand, and thus cannot adequately model, the system. More commonly, however, in complex systems, it occurs because information has not been made available, either inadvertently, because the designer didn't think the user would ever need it (Billings, 1991b), or deliberately, to keep the operator from "mucking about" with the system (e.g., Noble, 1983). This is an important cause of automation opacity.

Regardless of the cause, this information deficiency excludes the operator from effective awareness of system state, and thus excludes him or her from effective involvement in system activity. These are design problems, and the SA construct is as good a vehicle as any with which to make designers aware of what the human operator must know to remain in control of the system. As I pointed out at Orlando (Billings, 1994), the human must be *informed of* and *involved in* a system's operation to retain command of that system.

How can we measure, or evaluate, or manipulate, situation awareness?

I have already indicated my bias with respect to this important question. I believe, as do Flach (1995), Andre (1995, this volume) and others among us, that situation awareness is an abstraction that exists in our minds. The depiction or representation of the elements of a situation can be manipulated by the designer. The understanding of those information elements can be manipulated by proper training. The awareness and interpretation of the meaning of that information by the operator can be improved by practice. All of these elements, and their manipulation, can be examined, measured (at least in theory) and understood by knowledgeable observers like ourselves. I might even be able to modify your construct of situation awareness, which of course is what I am trying to do in this paper.

But you cannot measure or quantify an abstraction. You should not use it to explain a human error because of the circularity problem, as Underwood (1957, quoted by Flach, 1995) pointed out. Then what is it good for? Flach said it neatly. "An important contribution of an operational description is to *bound* the problem": to help us focus our research efforts, and to help us abstract our findings with respect to that problem.

Then what *can* we measure or manipulate?

As investigators, we can evaluate and measure the processes involved in acquiring situation awareness, as many of you are doing, both in the laboratory and "in the wild", as Woods (1993) puts it. At Orlando, I called these processes "situation assessment" (Billings, 1994), though I recognize that there are both cognitive and pre-cognitive processes involved and the term "assessment" may seem too narrow.

As investigators, we can evaluate the products resulting from these processes: the state of awareness of a human operator concerning the relevant dimensions of the situation. This has been the focus of a large part of the research performed since our last meeting. Much of the rest of that

research has been devoted to the development and validation of better methods with which to accomplish this difficult task, and this conference has done an excellent job of summarizing the state of our knowledge concerning those methods.

Rather less research, I fear, has been directed at understanding and rigorously describing the situational variables of which we need to be aware. I would echo Dr. Meister's call for a clearer delineation of the independent variables that motivate awareness. We need a better taxonomy of those variables and a clearer understanding of their interdependencies, especially as our human-machine systems become more tightly coupled and complex (Perrow, 1984). We most remember that in many cases we are no longer able to appreciate the true situation without the aid of machines; this is especially true in military aviation. If this is true, however, then those machines must tell us more of what we need to know, and they must do it more effectively and less ambiguously than they have done to date.

What is the real purpose of this activity?

To quote Dr. Meister once more, "the major function of ergonomics is to translate behavioral principles to system design". I believe he also hinted that we're not doing a very good job of it. To do better, we must have a comprehensive and precise understanding of the independent and dependent variables, and more explicit models of how the former influence the latter.

An airline pilot who is also an experienced human factor specialist explains the pilot's dilemma this way, "If you can't see what you've got to know, then you've got to *know* what you've got to know. And if you *don't* know what you've got to know, then you've got to be told!" (Demosthenes, personal communication, 1994). You, as investigators, must figure out what they've got to know, and how to show it to them clearly enough so they have time to do something about it. And you, as ergonomists, have got to model what they will do with that information, to guide designers in developing displays that will tell them and controls that will let them take advantage of the knowledge.

Some Challenges

Over-simplification

I am indebted to Dr. Philip Smith (personal communication, 1995) for suggesting to me that "situation awareness" results from a human operator's perception of, and attention to, world state elements, together with cognitive processes that incorporate a world model to interpret the information. These are real processes and real variables, and their result, an interpretation of the world state, certainly exists. If it exists, it should be describable.

The most serious shortcoming of the situation awareness construct as we have thought about it to date, however, is that it's too neat, too holistic and too seductive. It is too easy to use it, rather than its components, to explain things. We heard here that deficient SA was a causal factor in many airline accidents associated with human error. We must avoid this trap (see Flach, 1995); deficient situation awareness doesn't "cause" anything. Faulty spatial perception, diverted attention, inability to acquire data in the time available, deficient decision-making, perhaps, but not a deficient abstraction!

I don't mean to be over-critical. We're making good progress in a lot of respects. We understand the problem better, though we have a distance to go. We're beginning to understand some ways of getting at it, and we've developed and even validated some very helpful paradigms. We seem to have our feet fairly firmly planted in the real world, which is the only venue in which we're going to solve this real-world problem. This useful conference has pointed out clearly how

much remains to be done, and I have tried to point out above some of the most important questions we must answer.

Predictability and expectations

In her excellent discussion of the theoretical bases of SA, Dr. Endsley (1995, this volume) talked about expectations. She correctly pointed out that they are based on mental models of real systems, and that they also serve as filters. Dr. Meister (1995, this volume) also emphasized that characteristics of the physical system "are almost as important as behavioral variables".

The importance of predictability in a complex system cannot be over-rated (Billings, 1991b). Systems must be predictable to allow a human operator to form mental models of how they work, to develop trust in them and to form expectations that they will continue to work that way. But expectations are a two-edged sword. Cognitive biases and heuristics may lead a human operator to reject hypotheses about world states that are not in accord with expectations. This may lead the operator to ascribe novel symptoms to a familiar cause when in fact they denote a new situation. This predictability double bind imposes a further burden upon designers, who must both provide information to permit operators to build trust in their system, and also information to permit them to retain a degree of skepticism about the situations in which they may find themselves. This is easier said than done; operators will come to rely upon usually reliable systems. But such systems can and do fail, and novel situations do arise and must be dealt with. We must find ways to minimize the impact of this dilemma.

Peripheralization

Dr. Kennedy (1995, this volume) suggested, as I have (Billings, 1991a), that people perceive themselves as becoming more distant from the flying task in highly automated aircraft, and that the requirement for situation awareness has therefore become much higher. I believe that the *requirement* has not changed. Rather, it is the information acquisition and assessment tasks that may have become more difficult because of the plethora of information now available, some of it not very well represented. This again is a system and interface design problem.

I would make one further point in this regard. Kennedy asked whether situation awareness was different from "headwork", or from workload. "Headwork" is the set of processes by which awareness is attained. The headwork required to maintain situation awareness is probably greater in a more complex and autonomous system, and the exercise of "airmanship" (another useful abstract construct allied to SA) may also be more difficult. But airplanes, or other similar systems, should *not* be more difficult to manage than they are to fly. At this time, there can be appreciably more cognitive workload involved in managing an airplane than in flying it, which of course is why pilots "turn off" the automation when they become too heavily loaded (Curry, 1985). We've let the cart get ahead of the horse, and we need to get this back under control.

Outcome measures

In one way or another, many of the research studies discussed here have inferred situation awareness from outcome measures. Just as aircraft accidents should not be attributed to deficient situation awareness, there is a danger in using outcomes as a measure of a psychological construct rather than a measurable psychological variable or set of them. SA is an encompassing, "lumping" concept; there's a lot of stuff in that "black box", and simply saying that the box is faulty is not informative. I believe performance measures can be helpful, provided that we take account of the intervening variables between knowledge and that performance, but that chain of causality is by no means trivial.

Conclusion

The existence of this conference is testimony to the utility of the situation awareness construct. The content of the conference, however, is equal testimony to the complexity of that construct. An old government maxim says, "There's a simple solution to almost every complex problem. It will be neat, plausible—and usually wrong." We will not find a neat, plausible, simple solution to this complex problem and we should not waste time looking for one. When we *do* find the right answers, however, we will know appreciably more about "cognition in the wild" (Woods, 1993), and we will have learned a good deal more about how people function in complex, highly dynamic environments.

With apologies to Dr. Endsley and the rest of our attendees, let me complete this paper as I began it. Let me congratulate all of you for having the "sagatcity[1]" to attend this important conference. On your behalf, let me thank our conference sponsors and the organizers, Dr. Endsley and Dr. Garland, for their sagatcity in mounting it for our benefit. Now it is time for us to return home and to use our own sagatcity as we get on with the tasks before us: to better understand what situation awareness really is, its dimensions, and how to improve it.

References

Billings, C.E. (1991a). Toward a human-centered aircraft automation philosophy. *International Journal of Aviation Psychology* 1(4), 261-270.

Billings, C.E. (1991b). *Human-Centered Aircraft Automation: A Concept and Guidelines.* Moffett Field, CA, NASA-Ames Research Center: NASA Technical Memorandum 103885.

Billings, C.E. (1994). Situation awareness in complex systems: commentary. In: Gilson, R.D., Garland, D.J., & Koonce, J.M., eds. *Situational Awareness in Complex Systems.* Daytona Beach, FL: Embry-Riddle Aeronautical Press.

Curry, R.E. (1985). *The Introduction of New Cockpit Technology: A Human Factors Study.* Moffett Field, CA, NASA-Ames Research Center: NASA Technical Memorandum 86659.

Endsley, M.R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors* 37(1), 65-84.

Flach, J.M. (1995). Situation awareness: proceed with caution. *Human Factors* 37(1), 149-157.

Gilson, R.D., Garland, D.J., & Koonce, J.M., eds. (1994). *Situational Awareness in Complex Systems.* Daytona Beach, FL: Embry-Riddle Aeronautical Press.

Noble, D.F. (1983). *Forces of production: A social history of industrial automation.* New York, NY: Knopf.

Perrow, C. (1984). *Normal Accidents.* New York: Basic Books.

Pew, R. (1994). An introduction to the concept of situation awareness. In: Gilson, R.D., Garland, D.J., & Koonce, J.M., eds. *Situational Awareness in Complex Systems.* Daytona Beach, FL: Embry-Riddle Aeronautical Press.

Woods, D.D. (1993). *Cognitive Systems in Context.* Columbus, OH: The Ohio State University; Cognitive Systems Engineering Laboratory paper.

[1] "Sagatcity" (n.): wisdom arising from appreciation of reality; a high state of situation awareness. From SAGAT, the Situation Awareness Global Assessment Technique (see Endsley, 1995).

The State of Situation Awareness Measurement: Circa 1995

Richard W. Pew[1]

BBN Corp.

Abstract

Achieving situation awareness has become a design criterion supplementing more traditional performance measures. However, measuring SA requires more than an every-day understanding of the term. I build a more formal definition that includes the definition of a "situation," the elements of information and knowledge associated with SA, and the kinds of information resources available in typical applications to achieve it. It is argued that measuring SA implies having a standard against which to compare human performance and such a standard is proposed in terms of an abstract ideal and a practically realizable ideal. A taxonomy of measurement methods is presented and illustrated together with a critique of the potential application of various methods. It is argued that many different classes of methods are potentially appropriate, but each is appropriate to a selected class of measurement requirements.

Introduction

Situation Awareness measurement must be founded on a careful definition of what one means by the term and that understanding must be conditioned on each context in which it is to be measured. I begin this paper with a reiteration of the definitional issues previously presented in Pew (1994). Related material is also presented in Tenney, Adams, Pew, Huggins, and Rogers (1992), and in Deutsch, Pew, Rogers, and Tenney (1994).

Formal Definition of Situation Awareness

In order to adequately define SA we need to understand what we mean by a "situation" and we need to know what it is about situations of which we must be aware. It is also of interest to catalog where that information and knowledge come from. I adopt Table 1 as a working definition of a situation:

---

[1]I wish to thank my colleagues, Yvette J. Tenney, Marilyn Jager Adams, William H. Rogers, and Stephen Deutsch who were my collaborators on some of the ideas reported in this paper. Parts of this work were supported under Contract NAS1-18788 with the NASA Langley Research Center, Dr. Raymond Comstock, Technical Monitor.

Table 1. Definition of a Situation

A situation is a set of environmental conditions and system states with which the participant is interacting that can be characterized uniquely by a set of information, knowledge and response options.

However, the concept of a situation is meaningful, according to this view, only if we can define a discrete and denumerable set of them, that is, that the awareness requirements can be broken discretely into packets, each associated with a set of system states. This ability to partition situations implies that, while the environment is more or less continuously changing with time, only some of the changes are large or severe enough to create a changed situation from the perspective of the crew member. We must be able to identify the boundaries at which we wish to say that a situation has changed. Examples of such changes that are severe enough to redefine the SA might be: a forest fire that has run out of its firebreak, a ship that enters the range of on-coming traffic, a train that encounters a conflicting train on the same track, a power plant that transitions from start-up to full power or an aircraft autopilot disengagement, either expectedly or unexpectedly.

The second part of the definition requires that a "situation" have associated with it the information and knowledge that we are calling awareness. Table 2 shows the elements that need to be included and Table 3 lists the informational sources that the crew member has to draw on to achieve SA.

Table 2. Elements of Awareness, Given the Situation

• Current state of the system (including all the relevant variables).
• Predicted state in the "near" future.
• Information and knowledge required in support of the crew's current activities.
• Activity Phase
• Prioritized list of current goal(s)
  • Currently active goal, subgoal, task
  • Time
• Information and knowledge needed to support anticipated "near" future contexts.

Table 3. Information Resources Contributing to Awareness

• Sensory information from the environment • Visual and auditory displays • Decision aids and decision support systems • Extra- and intra-crew communication • Crew member background knowledge and experience

It is also important to note that, while much of the SA literature focuses on spatial awareness, there are many other aspects of systems and their operations about which awareness is required. Table 3 identifies a set of such concerns. *Spatial awareness* is self-explanatory. *Mission/goal awareness* refers to the need to keep current with respect to the phase of the mission and the currently active goals that are to be satisfied. *System awareness* is especially important in complex

highly-automated systems. The work of Sarter and Woods (1994) identify the critical difficulties associated with understanding and tracking the mode of flight management computers. *Resource awareness* is needed to keep track of the state of currently available resources, including both physical and human resources. One needs to know the current activities of other crew members so that their availability for critical tasks is known. This is different from *Crew awareness* which refers to the need for the team of crew members to share their information and interpretation of current system events. They need to be operating in a common framework of information at all times.

Table 4. Multiple Elements of Awareness that need to be Considered

• Spatial Awareness
• Mission/Goal Awareness
• System Awareness
• Resource Awareness
• Crew Awareness

To have defined situations and the components awareness is still not enough because we must also define the requirements for SA that are presented by a given situation. Otherwise, we have no standard against which to judge how successful a crew member has been in achieving SA. Our thinking about this has led to the consideration of an ideal awareness, and the obtainable ideal, in addition to the actual SA achieved.

The *ideal* is that SA that is defined by experts evaluating the requirements at leisure or even after the fact. It includes both the information and knowledge requirements. The *obtainable ideal* is that subset that is actually available for the crew member to acquire. When defining the obtainable ideal we assume the availability of well-designed information resources and take into account the fact that what is practically available is constrained by expected human cognitive abilities. It does not seem fair to evaluate actual SA achievement by a crew member with respect to a goal that is not practically achievable. Both the ideal and the achievable ideal are assessed independently of crew member performance. However, the *actual* SA must be inferred from measurement or observation. It is the difference between ideal SA and achievable ideal that create a space for evaluating design alternatives contributing to improved SA. The difference between achievable ideal and actual SA that creates the space for evaluating individual differences in the ability to achieve SA and for developing training opportunities.

Having enumerated the elements of a definition of SA I will now move on to the issues of measurement

A Taxonomy of Measurement Methods

The subject of measurement encompasses more than just the selection of performance measures. It includes development of the measurement context, including the systems and scenarios of use. In Endsley (1995a) a taxonomy is presented and measurement methods are reviewed. She found weaknesses in every measurement context except job samples of operationally realistic scenarios. Since her work at the time was based largely on SA in air-to-air combat, this was not inappropriate, however, I take a more eclectic view. I believe there are many relevant

measurement contexts, but each is appropriate to particular purposes. I break down the categories of methods as shown in Table 5.

Table 5. A Taxonomy of SA Measurement Methods

• Direct System Performance Measures
• Direct Experimental Measures
• Verbal Protocols
• Subjective Measures

In Endsley (1995a) it was argued that only scenarios with full face validity were appropriate for the measurement of SA. I believe that any simulation experiment involves compromises from the realism of the real world. The issue is only one of degree - the face validity required of the scenario depends on the purpose of the assessment, as will be illustrated in the discussion to follow. I will consider each category in my taxonomy in turn.

Direct System Performance Measures

I agree with Endsley that there are only limited places where direct system performance measures, *per se*, are appropriate for assessing SA. Those occur in cases where there would be general agreement among an audience of peers that the performance in question was driven solely or largely by SA.

It is more common to invest significant planning resources in the careful design of the scenario to specifically create opportunities for using standard, non-obtrusive performance measures to assess a specific SA issue. Sometimes it involves introducing subtle or counter-intuitive indicators. For example Busquets, Parrish, Williams and Nold (1994), were studying the usefulness of specific navigation displays during aircraft approach and landing on a dual-runway airport. They deliberately introduced a second aircraft apparently intending to land on the other runway. However, just as the pilot-under-study was about to reach final approach, the second aircraft deviated and appeared to be landing on the runway to which the pilot-under-study was committed. They measured the time to take action to avoid the second aircraft. This, by design, was very clearly a measure of the pilot's SA regarding activity in the airspace around him.

A second method of scenario manipulation is to introduce disruptions intended to disorient the crew or operator and from which they must recover. The measure of SA is the recovery time or success of recovery under the assumption that the amount of time required to recover is proportional to how well the crew could use the system displays for recovery from the disorientation. Busquets, et. al., (1994) used this method in their experiment by blanking the displays and, during the period that the displays were blanked, introducing a systematic offset in the aircraft position. The crew had to use the displays to return to their position on the flight path. Both of these methods produced significant differences in performance as a function of the display conditions under study.

A third method that I have advocated, but for which I have no examples, is to introduce anomalous data or instrument readings, readings that create a pattern that could not have been produced by any realistic condition. The measure of performance is the time required to detect the anomaly. Admittedly this method suffers from a difficulty identified by Endsley (1995a), namely that the detection time depends not only on the time to detect the anomaly, but also on the locus and urgency of the crew members attentional focus at the time the anomaly was introduced. The scenario would have to be carefully designed to assure that the primary demand at the time was associated with the pattern that was, indeed anomalous.

Direct Experimental Techniques

The most common approach is to use direct experimental techniques. On my list of such techniques are the use of queries or probes and the use of measures of information seeking. Probes can be introduced during on-going task performance, if the pace of the task is slow or there are many periods of relative inactivity. However, it is more common to suspend the task - to freeze the simulation - and to ask one or more questions about the state of the task or the environment before resuming the action. This method has been formalized by Endsley (1988) as the Situation Awareness Global Assessment Technique (SAGAT) and applied in many situations. It, and variants of it, are now widely used.

I like SAGAT. It makes use of systematic, preplanned procedures. The user is forced to think through in detail before hand exactly what aspects of SA are going to be assessed. It uses computer administration of the set of SA questions during the simulation freeze. This assures control of question administration. I like Endsley's emphasis on three levels of SA, namely, (1) perception of information, (2) integration of data with goals to produce meaning, and (3) projection of the near future and I like her insistence that the questions be context specific. I like the fact that the questions asked at a particular sampling point are drawn at random, although this makes it a very different tool from the kind described above that attempts to assess some specific aspect of SA through performance assessment of some pre-designed scenario feature. Using SAGAT one can obtain interesting and rich data about the aggregate levels of specific classes of SA, but cannot answer specific questions about SA at a particular, perhaps critical, point in a scenario.

Two controversial issues are often introduced when the probe technique is considered. (1) Does the use of the freeze technique disrupt ongoing task performance and thereby place the subject in an unrealistic setting, producing an unrealistic assessment of SA? (2) Does the expectation that probes will be presented change the subjects' behavior? Once the first one is presented, do they anticipate that additional probes will be presented and prepare for them? If so, then the experimenter does not get an accurate assessment of the real level of SA. In Endsley, (1994), (1995a) the first of these issues was addressed in an experiment in which the performance on 25 percent of the trials, on which the subjects received no probes, was compared to the remaining 75 percent on which varying numbers of probes were presented. She found no significant difference in performance among the groups. However, the design of the experiment was such that the same subjects participated in all conditions. They were not told before each trial whether probes should be expected or not. It is not surprising, therefore, that no differences were found.

In this volume Endsley (1996) reports a new experiment in which the presence or absence of probes was manipulated between subjects and still no significant differences were found. This is a much sounder design for examining this question, however, she still presented both groups with pretraining about how to respond to probes and how to behave when the simulation was frozen. The question remains whether this created such an expectation on the part of the subjects that they did not behave differently when probes were or were not presented.

An alternative design that addresses the issue of surprise was used by Wickens and Prevett (1995). In an experiment involving the design of navigation displays, they interrupted the simulation to ask a series of questions, not unlike Endsley's procedure, and found that the first probe had a significantly longer reaction time, but there was no significant difference between the first and subsequent probes in the accuracy of response.

In my opinion, both of these issues are deserving of somewhat less attention than they have been given. Since most measurement is relative, main effects that appear, even though moderated by the experimental techniques can still be useful diagnostic indicators. Yes, you may get a different level of performance on the first and later trials, but both will still accurately reflect the relative differences among treatment conditions. The circumstance that should receive attention is when the interruption or the surprise creates a statistical interaction between two or more of the other variables of interest. For example in an experiment by Olson and Sivak, (1986), the

relationship between driver age and reaction time was impacted by introducing differing expectancies in a driving task.[1]

In addition to queries and probes, I include, in the category of direct measures, assessments of information seeking. Although some investigators define a separate category of physiological measures, I consider the two primary information seeking measures, namely eye-movements and eye blink response to be direct experimental measures. The medium by which they are recorded is not really relevant to their application to SA. It cannot be argued conclusively that simply because an observer's eye is directed toward a specific object or expression, that it is seen. However, the correlation between looking and seeing is likely to be quite high. Vidulich, Stratton, Crabtree and Wilson (1994) in the most comprehensive experiment I have found that is designed to evaluate alternative measures of SA, used eyeblink as an element of a large set of measures that they studied and found a significant effect of display condition on eyeblink duration.

Verbal Protocols

By a verbal protocol I mean information recorded from the observer or crew, either during, immediately after, or in the course of a video replay of the exercise of a scenario. The subject may be asked to "think aloud" or to explain the information relied on. This is a technique that is most useful early in development of an evaluation. It may help to solidify SA concepts that need to be measured more systematically. Yes, it is disruptive when the observer is asked to report in real time during execution of the scenario, but the information gained may be worth enduring the disruption.

Subjective Measures

It is so difficult to obtain quantitative objective measures of SA that many investigators rely on subjective measures instead. They may be self-assessments, expert judgments, peer ratings, supervisor or instructor ratings.

The most thoughtful and systematic development of a self-administered test of SA to date is reported in Taylor (1990), and is referred to as the Situation Awareness Rating Technique (SART). It consists of three sub-scales which are combined in an equation to produce an overall estimate of the subject's SA. The technique is described in more detail elsewhere in this volume. It includes, as one of the subscales, the demand on attentional resources. Including attention demand in the integrated equation defining SA confounds SA with workload. In my opinion, workload and SA should be viewed as different, but, of course, they may be correlated in practice.

In reflecting on the relationship between SA and workload, I am reminded of the somewhat analogous relationship between speed and accuracy of performance. While we often speak of the speed-accuracy trade-off, this is only one of the ways that speed and accuracy can be related. First, across a set of individuals, we can expect a high positive correlation between speed and accuracy. Those subjects who tend to be fast, also tend to be accurate. Similarly, thinking of it as an individual difference variable, subjects with good intrinsic ability to maintain SA would be expected to achieve it with less workload. Second, if we train an individual to improve performance, we expect her/his speed and accuracy to improve more or less together. There is a high positive correlation between speed and accuracy within an individual across a period of practice. The analogous statement with respect to SA and workload is that training can be expected to increase SA for an equivalent or lower level of workload. Finally, analogous to the speed accuracy trade-off, if we challenge an individual to improve his/her SA at a particular point in practice, we would expect that the workload associated with achieving that improvement to be

---

[1] I am indebted to Neil Charness for supplying this example.

increased. To include aspects of attentional demand in the formula for the assessment of SA makes it difficult for these relationships to emerge.

The experiment of Vidulich, et., al. (1990), evaluated SA in the context of a military air-to-ground attack target detection task and had three display conditions. The condition predicted to provide the most SA used an integrated navigation and targeting display and made the targets a distinctive color, as might be provided by an automatic target recognizer. The condition predicted to be next best, from the point of view of SA, used the same display, but made the targets a camouflaged color that was more difficult to detect. The third condition, predicted to be worst, presented separate displays for navigation and targeting having information that had to be integrated cognitively in order to assess the relative positions of the various targets. This experiment further illustrates the difficulties of confounding workload and SA. The first two conditions, I would argue, differ in the amount of workload required for detection. The second and third conditions differ in the cognitive effort associated with assimilating and integrating the information from one or two displays. These manipulations do not seem to me to be varying the same dimensions of SA, the first simply requiring more workload to achieve the detection on which SA is based, while the second genuinely manipulates the difficulty of achieving an integrated SA picture.

By definition self-ratings can only reflect self-awareness. The operators do not necessarily have a perspective on what they *should be* aware of. The "bad news" is that it usually reflects little more than a confidence rating or preference on the part of the operator or crew member. The "good news" is that sometimes that is exactly what is of interest to the investigator.

The best example of supervisor/peer rating is reported in the U.S. Air Force SAINT study (McMillan, 1994). The investigators were interested in assessing SA as an individual difference variable in the combat skill of F-15 fighter aircraft pilots. They developed a rating scale of SA, that was actually inclusive enough to be called a combat skill rating rather that strictly an SA rating. They obtained supervisor (squadron commanders) and peer ratings for 200 line combat pilots. Interestingly, the peer ratings (that is pilots rating each other) correlated with their supervisors ratings 0.80.

In a separate study a representative subset of 40 of these pilots flew a demanding combat simulation and two instructor pilots, who observed these simulation exercises, produced a set of independent SA ratings. The instructors did not know about the peer or supervisor ratings. When the investigators then correlated the combined supervisor/peer ratings with the instructor ratings, they found that the supervisor/peer ratings accounted for 31% of the variance in the instructor's ratings.

Model-Referenced Performance Measurement

I want to mention one final measurement technique that, to my knowledge, has not yet been tried. I call it Model-Referenced Performance Measurement. It is applicable to human performance measurement in general, but could be used, I believe, to collect very interesting SA-relevant data. The technique is illustrated in Figure 1. To employ the technique we create a simulation and run a human crew member through one or more simulation scenarios, just we would do in any other SA experiment. Then, in parallel with the running simulation, we connect a duplicate simulation (the system model in the figure), to a human performance model. The same scenario inputs are received by both systems. Periodically the system model must be updated from the live simulation to make sure that they do not get out of synchronization.

To the extent that the human performance model is an accurate representation of the human operator, then to that extent the variables being recorded and used in the model are reflective of the behavior of the human subject. However, in the case of the model we have access to variables that we cannot directly assess in the human operator. For example human operators must keep in mind and monitor several activities at once. We might model this as a dynamic priority stack of items to be checked and actions to be taken when certain variables are out of a tolerable range. While a subject might have difficulty reporting on what was on his "priority stack" at any particular time, in

the model we have direct access to that stack. The number of items on the stack and their relative position in priority might be an interesting index of SA.

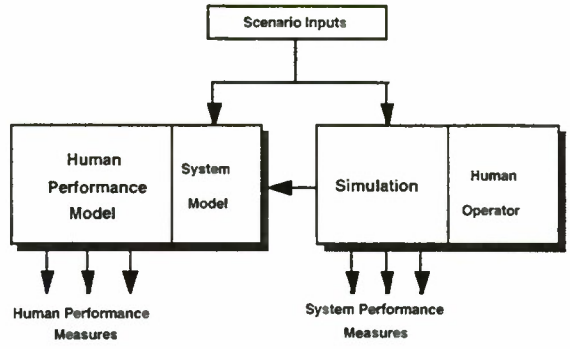


Figure 1. Block Diagram illustrating Model-Referenced Performance Measurement

Summary

As I have illustrated, there are many SA measurement techniques that are now beginning to be investigated in practical situations to evaluate their relevance and usefulness. We are beginning to understand what measures are good for what purposes. To search for the universal measure is to search for the Holy Grail. Rather, each context in which an investigator desires to assess SA must make a judicious choice of the measurement context together with the appropriate choice of measures. The choices should be made with full understanding of the classes of situations that are to be measured and some thought about what will index transitions from one situation to another. The investigator must also figure out ahead of time, perhaps in pilot studies, just what the SA requirements of the situation(s) are so that they can assess achieved SA in relation to those requirements.

Reference

Busquets, A.M., Parrish, R.V., Williams, S.P., and Nold, D.E. (1994). Comparison of pilots' acceptance and spatial awareness when using EFIS vs pictorial display formats for complex, curved landing approaches. In Gilson, R.D., Garland, D.J., and Koonce, J.M., (Eds.). *Situational Awareness in Complex Systems*. Daytona Beach, Fl: Embry-Riddle Aeronautical University Press.

Deutsch, S.E., Pew, R.W., Rogers, W.H., and Tenney, Y.J. (1994). Toward a Methodology for Defining Situation Awareness Requirements - A Progress Report. National Aeronautics and Space Administration, BBN Report No. 7983, April.

Endsley, M.R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (Vol. 1). Santa Monica, CA: Human Factors Society.

Endsley, M.R. (1994). Situation awareness in dynamic human decision making: Measurement. in Gilson, R.D., Garland, D.J., and Koonce, J.M., (Eds.). *Situational Awareness in Complex Systems*. Daytona Beach, Fl: Embry-Riddle Aeronautical University Press.

Endsley, M.R. (1995a). Measurement of situation awareness in dynamic systems." *Human Factors*, *37*, *1*, pp. 65-84.

Endsley, M.R. (1996). Theoretical underpinnings of situation awareness: A critical review. In Garland, D.J., and Endlsey, M.R., (Eds.). *Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness.*

McMillan, G.R., (1994) "Report of the Armstrong Laboratory Situation Awareness Integration (SAINT) team." In Vidulich, M., (Ed.)*Situation Awareness: Papers and Annotated Bibliography*, Crew Systems Directorate, Human Engineering Division, Wright-Patterson AFB, OH. pp. 37-47.

Olson, P.L., and Sivak, M. (1986). Perception-response time to unexpected roadway hazards. *Human Factors*, *28*, *1*, pp. 91-96.

Pew, R. W. (1994). An introduction to the concept of situation awareness. In Gilson, R.D., Garland, D.J., and Koonce, J.M., (Eds.). *Situational Awareness in Complex Systems.* Daytona Beach, Fl: Embry-Riddle Aeronautical University Press.

Sarter, N.B., and Woods, D.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*, pp. 1-28.

Taylor, R.M. (1990). Situation awareness rating techniques (SART): The development of a tool for aircrew systems design. In *Situational Awareness in Aerospace Operations* (Chapter 3). France: Neuilly-sur-Seine, NATO-AGARD-CP-478.

Tenney, Y.J., Adams, M.J., Pew, R.W., Huggins, A.W.F., and Rogers, W.H. (1992). *A principled approach to the measurement of situation awareness in commercial aviation .* NASA Contractor Report 4451.

Vidulich, M.A., Stratton, M., Crabtree, M., and Wilson, G. (1994). Performance-based and physiological measures of situational awareness. *Aviation, Space, and Environmental Medicine*, *65*, (5 Supplement), pp. A7-A12.

Wickens, C.D., and Prevett, T.T. (1995). Exploring the dimensions of egocentricity in aircraft navigation displays. *Journal of Experimental Psychology: Applied*, *1*, 2, pp. 110-135.

Theoretical Underpinnings of Situation Awareness: A Critical Review

Mica R. Endsley

Texas Tech University

Introduction

The enhancement of operator situation awareness (SA) has become a major design goal for those developing operator interfaces, automation concepts and training programs in a wide variety of fields, including aircraft, air traffic control, power plants, and advanced manufacturing systems. To evaluate the degree to which new technologies or design concepts actually improve (or degrade) operator SA, it is necessary to systematically evaluate them based on a measure of SA, thus providing a determination of which ideas have merit and which have unforeseen negative consequences.

Explicit measurement during design testing determines the degree to which design objectives have been met. Performance parameters must be carefully specified and supporting data collected. In addition, for many systems, operator situation awareness and workload need to be directly measured during design testing. High level performance measures (as collected during the limited conditions of simulation testing) are often not sufficiently granular or diagnostic of differences in system designs. Thus, while one system design concept may be superior to another in providing the operator with needed information in a format that is easier to assimilate with operator needs, the benefits of this may go unnoticed during the limited conditions of simulation testing or due to extra effort on the part of operators to compensate for a design concept's deficiencies. If situation awareness is measured directly, it should be possible to select concepts that promote situation awareness, and thus increase the probability that operators will make effective decisions and avoid poor ones. Problems with situation awareness, frequently brought on by data overload, non-integrated data, automation, complex systems that are poorly understood, excess attention demands, and many other factors, can be detected early in the design process and corrective changes made to improve the design.

In addition to evaluating design concepts, a measure of SA may also be useful for (a) evaluating the impact of training techniques on SA, (b) conducting studies to empirically examine factors that may effect SA, such as individual abilities and skills, or the effectiveness of different processes and strategies for acquiring SA, and (c) investigating the nature of the SA construct itself.

To adequately address these goals, however, the veracity of available SA measures needs to be established. Ultimately, validity and reliability must be established for any SA measurement technique that is used. It is necessary to establish that a metric (a) indeed measures the construct it claims to measure and is not a reflection of other processes, (b) provides the required insight in the form of sensitivity and diagnosticity, and (c) does not substantially alter the construct in the process, providing biased data and altered behavior. In addition, it can be useful to establish the existence of a relationship between the measure and other constructs as would be predicted by theory. To this end, the available theoretical foundation for the concept of situation awareness will be reviewed and the implications of this information for SA measurement discussed.

Theories of SA

Situation awareness can be described as a person's state of knowledge or mental model of the situation around them. Many definitions of SA have been developed, some very closely tied to the aircraft domain and some more general (see Dominguez (1994) or Fracker (1988) for a review). A general, applicable definition describes SA as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" (Endsley, 1988). Thus, it includes more than perceiving or attending to information, but also the integration of multiple pieces of information and a determination of their relevance to the person's goals, and the ability to forecast future situation dynamics, thus providing for timely and effective decision making.

Several researchers have put forth theoretical formulations for depicting the role of numerous cognitive processes and constructs in SA (Adams, Tenney, & Pew, 1995; Endsley, 1988, 1994, 1995b; Fracker, 1988; Smith & Hancock, 1994; Taylor, 1990; Taylor & Selcon, 1994; Tenney, Adams, Pew, Huggins, & Rogers, 1992). There are many commonalties in these efforts pointing to key mechanisms that are important for SA and which have a direct bearing on the appropriateness of proposed measures of SA.

Endsley (1988; 1994; 1995b) proposed a framework model based on information processing theory (Wickens, 1992). Key features of the model will be summarized here. The model shows situation awareness as a stage separate from decision making and performance. Situation awareness is depicted as the operator's internal model of the state of the environment. Based on that representation, operators can decide what to do about the situation and carry out any necessary actions. Situation awareness therefore is the main precursor to decision making, however, many other factors can come into play in turning good situation awareness into successful performance.

Several factors will impact the accuracy and completeness of situation awareness that individual operators derive from their environment. First, humans are limited by working memory and attention. The way in which attention is employed in a complex environment with multiple competing cues is essential in determining which aspects of the situation will be processed to form situation awareness. Once taken in, information must be integrated with other information, compared to goal states and projected into the future - all heavily demanding on working memory.

Long-term memory stores in the form of mental models or schema are hypothesized to play a major role in dealing with these limitations. With experience operators develop internal models of the system they operate and the environments they operate in. These models serve to help direct limited attention in efficient ways, provide a means of integrating information without loading working memory, and provide a mechanism for generating projection of future system states. Associated with these models may be schema of prototypical system states. Critical cues in the environment may be matched to such schema to indicate prototypical situations which provide instant situation classification and comprehension. Scripts of the proper actions to take may be attached to these situation prototypes, simplifying decision making as well. Schemata of prototypical situations are incorporated in this process and in many instances may also be associated with scripts to produce single-step retrieval of actions from memory, thus providing for very rapid decision making such as has been noted by Klein (1989). The use of mental models in achieving SA is considered to be dependent on the ability of the individual to pattern match between critical cues in the environment and elements in the mental model.

Goals are also important for situation awareness. Essentially human information processing in operating complex systems is seen as alternating between data driven (bottom-up) and goal driven (top-down) processing. In goal driven processing, attention is directed across the environment in accordance with active goals. The operator actively seeks information needed for goal attainment and the goals simultaneously act as a filter in interpreting the information that is perceived. In data driven processing, perceived environmental cues may indicate new goals that need to be active. Dynamic switching between these two processing modes is important for successful performance in many environments.

In addition, preconceptions or expectations influence the formation of situation awareness. People may have certain expectations about what they expect to see or hear in a particular environment. This may be due to mental models, prior experiences, instructions or other communications. These expectations will influence how attention is deployed and the actual perception of information taken in. That is, there is a tendency for people to see what they expect to see (or hear).

Finally, automaticity is another mechanism developed with experience that can influence situation awareness. With experience, the pattern-recognition/action-selection sequence can become highly routinized and developed to a level of automaticity (Logan, 1988). This provides good performance with a very low level of attention demand in certain well-understood environments. In this sense, automaticity can positively impact situation awareness by reducing demands on limited attention resources. Situation awareness can also be negatively impacted by automaticity due to a reduction in responsiveness to novel stimuli. Information that is outside the routinized sequence may not be attended to. Thus, situation awareness may suffer when that information is important.

Fracker (1988) similarly points to the importance of both working memory and schemata in long-term memory for SA. He points out that while schemata may be very useful for facilitating situation assessment by providing a reduction in working memory demands, they can also lead to significant problems with biasing in the selection and interpretation of information that may create errors in situation awareness.

Adams, et. al. (1995) stress the importance of the inter-relationship between one's state of knowledge, or SA, and the processes used to achieve that knowledge. Framed in terms of Neisser's (1976) model of perception and cognition, they make the point that one's current knowledge effects the process of acquiring and interpreting new knowledge in an ongoing cycle. This agrees with Sarter and Woods (1991) statement that SA is "the accessibility of a comprehensive and coherent situation representation which is continuously being updated in accordance with the results of recurrent situation assessments". Smith and Hancock (1994) further support this proposition by stating that "SA is up-to-the minute comprehension of task relevant information that enables appropriate decision making under stress. As cognition-in-action (Lave, 1988), SA fashions behavior in anticipation of the task-specific consequences of alternative actions." (p. 59) In defining SA as "adaptive, externally directed consciousness", they take the view that SA is purposeful behavior that is directed toward achieving a goal in a specific task environment. They furthermore point-out that SA is therefore dependent on a normative definition of task performance and goals that are appropriate in the specific environment.

The relationship between SA and workload has also been theorized to be important. Taylor (1990) includes a consideration of supply and demand of resources as central to situation awareness. Adams, et. al. (1995) also discuss the task management problem, involving prioritizing, updating task status, and servicing tasks in a queue, as central to SA. Endsley (1993), however, shows that for a large range of the spectrum, SA and workload can vary independently, diverging on the basis of numerous factors. Only when workload demands exceed maximum human capacity is SA necessarily at risk. SA problems may also occur under low workload (due to vigilance problems) or when workload is in some moderate region.

Implications for Situation Awareness Measurement

Several implications can be drawn from these viewpoints for developing measures of situation awareness (Endsley, in press).

Processes vs. States

First, situation awareness as defined here is a state of knowledge about a dynamic environment. This is different than the processes used to achieve that knowledge. Different individuals may use different processes (information acquisition methods) to arrive at the same state of knowledge, or may arrive at different states of knowledge based on the same processes due to differences in comprehension and projection of acquired information or different mental models, schema, etc... Measures that tap into situation assessment processes, therefore, may provide information of interest in understanding how people acquire information, however, they will only provide partial and indirect information regarding a person's level of situation awareness.

Situation awareness, decision, and performance disconnect

Secondly, just as there may be a disconnect between the processes used and the resultant situation awareness, there may also be a disconnect between situation awareness and the decisions made. With high levels of expertise in well understood environments, there may be a direct situation awareness-decision link, whereby understanding what the situation is leads directly to selection of an appropriate action from memory. This is not always the case, however. Individuals can still make poor decisions with good situation awareness. They may have inadequate strategies or tactics guiding their decision processes. They may be limited in decision choices due to organizational or technical constraints. They may lack the experience or training to have good, well-developed plans of actions for the situation. Individual personality factors (such as impulsiveness, indecisiveness or riskiness) may also make some individuals prone to poor decisions. A recent study of human error in aircraft accidents found that 26.6% involved incidents where there was poor decision making even though the aircrew appeared to have adequate situation awareness for the decision (Endsley, 1995a).

Furthermore, the link between human decision making and overall performance is also indirect in many environments. A desired action may be mis-executed due to physical error, other workload, inadequate training or system problems. The system's capabilities may limit overall performance. In some environments, such as the tactical aircraft domain, the action of external agents (e.g. enemy aircraft) may also create poor performance outcomes from essentially good decisions (and vice-versa).

The relation between situation awareness and performance, therefore, can be viewed as a probabilistic link (Endsley, 1990, 1995b). Good situation awareness should increase the probability of good decisions and good performance, but does not guarantee it. Conversely, poor situation awareness increases the probability of poor performance, however, in many cases does not create a serious error. For instance, being disoriented in an aircraft is more likely to lead to an accident when flying at low altitude than when flying at high altitude. Lack of situation awareness about one's opponent in a fighter aircraft may not be a problem if the opponent also lacks situation awareness. In relation to situation awareness measurement, these issues indicate that behavior and performance measures may be only indirect indices of operator situation awareness.

Attention

The way in which a person deploys his or her attention in acquiring and processing information has a fundamental impact on situation awareness. Particularly in complex environments where multiple sources of information compete for attention, which information people attend to has a substantial influence on their situation awareness. Design changes that influence attention distribution (intentionally or inadvertently) therefore can have a big impact on situation awareness. Similarly, measurement techniques that artificially influence attention distribution should be avoided, as they may well change the construct that is being measured in the process.

Memory

Direct measures of situation awareness tap into a person's knowledge of the state of the dynamic environment. This information may be resident in working memory for a short period of time or in long-term memory to some degree and under certain circumstances. A significant issue for measures which attempt to tap into memory is to what degree people can report on mental processes to make this information accessible.

Automaticity may influence memory recall. With automaticity there is very little awareness of the processes used. A careful review of literature regarding automatic processing, however, reveals that while people may not be able to accurately report processes used in decision making, they are usually aware of the situation itself at the time (Endsley, 1995b). A low level of attention may make this information difficult to obtain from memory after the fact, however.

Time also affects the ability of people to report information from memory. With time there is a rapid decay of information in working memory, thus only long-term memory access may be available. Nisbett and Wilson (1977) demonstrate that recall of mental processes after the fact tends to be over-generalized, over-summarized, and over-rationalized, and thus may not be an accurate view of the actual situation awareness possessed in a dynamic sense. Real-time, immediate access of information from memory can also be difficult, however, as this process may influence ongoing performance and decision processes and situation awareness itself. Direct access of a person's memory stores can be problematic, therefore, and indicates that careful strategies for obtaining this information must be employed.

Workload

Situation awareness and workload, although inter-related in certain circumstances, are essentially independent constructs in many ways. This can be conceived of as a two-dimensional continuum between four possible extremes (Endsley, 1993).

1. Low SA and low workload - This would be a situation in which a person has little idea of what is going on and is not actively working to find out. Vigilance conditions, inattentiveness or low motivation may produce this state.

2. Low SA and high workload - If the volume of information and demand of tasks is too great, a loss of SA can easily result due to the operator's ability to attend to only a subset of required information. Attentional narrowing and disruption of scan patterns have been cited as leading to this condition. An inability to put together separate pieces of information may also cause this state.

3. High SA and high workload - This may occur where the person is working hard, but is successfully achieving an accurate and complete picture of the situation. The ability to maintain SA under conditions of high workload is one of our greatest design challenges.

4. High SA and low workload - Ideally, if the required information can be presented in a manner which is easy to process, high SA can be achieved under conditions of low workload. This is ultimately our biggest design goal.

Thus, SA and workload may dissociate in numerous ways, depending on characteristics of the system design, tasks, and the individual operator. As people can make tradeoffs between the level of effort extended and how much they feel they need to know, it is important that both SA and workload be measured independently in the evaluation of a design concept. A particular design may improve (or diminish) SA, yet workload may remain stable. That is, operators may be putting forth the same amount of effort, and getting more (or fewer) rewards in terms of the SA

achieved. With other designs, it may be that operators are able to maintain the same level of SA, yet may have to work much harder. In order to get a complete understanding of the effects of a particular design concept, therefore, both situation awareness and workload need to measured during design testing.

Conclusions

Numerous approaches to measuring situation awareness have been proposed. Ultimately, each class of measures may have certain advantages and disadvantages in terms of the degree to which the measure provides an index of situation awareness. In addition, the objectives of the researcher and the constraints of the testing situation will also have a considerable impact on the appropriateness of a given measure of SA. Certain classes of measures may be highly suitable for qualitative investigations of SA processes, for instance, yet be inadequate for design testing, and vice versa. Regardless, it is vital that the veracity of any measure used for measuring SA be established, so that informed research and design testing can take place.

References

Adams, M. J., Tenney, Y. J., & Pew, R. W. (1995). Situation awareness and the cognitive management of complex systems. *Human Factors, 37*(1).

Dominguez, C. (1994). Can SA be defined? In M. Vidulich, C. Dominguez, E. Vogel, & G. McMillan (Eds.), *Situation awareness: Papers and annotated bibliography (AL/CF-TR-1994-0085)* (pp. 5-15). Wright-Patterson AFB, OH: Armstrong Laboratory.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 97-101). Santa Monica, CA: Human Factors Society.

Endsley, M. R. (1990). Predictive utility of an objective measure of situation awareness. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 41-45). Santa Monica, CA: Human Factors Society.

Endsley, M. R. (1993). Situation awareness and workload: Flip sides of the same coin. In R. S. Jensen & D. Neumeister (Eds.), *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 906-911). Columbus, OH: Department of Aviation, The Ohio State University.

Endsley, M. R. (1994). Situation awareness in dynamic human decision making: Theory. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 27-58). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Endsley, M. R. (1995a). A taxonomy of situation awareness errors. In R. Fuller, N. Johnston, & N. McDonald (Eds.), *Human Factors in Aviation Operations* (pp. 287-292). Aldershot, England: Avebury Aviation, Ashgate Publishing Ltd.

Endsley, M. R. (1995b). Toward a theory of situation awareness. *Human Factors, 37*(1), 32-64.

Endsley, M. R. (in press). Situation Awareness Measurement in Test and Evaluation. In T. O'Brien & S. Charlton (Eds.), *Human Factors Testing & Evaluation*. Hillsdale, NJ: Lawrence Erlbaum.

Fracker, M. L. (1988). A theory of situation assessment: Implications for measuring situation awareness. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 102-106). Santa Monica, Ca: Human Factors Society.

Klein, G. A. (1989). Recognition-primed decisions. In W. B. Rouse (Ed.) *Advances in man-machine systems research* (5) (pp. 47-92). Greenwich, Conn: JAI Press, Inc.

Lave, J. (1988). *Cognition in practice.* Cambridge, UK: Cambridge University Press.

Logan, G. D. (1988). Automaticity, resources and memory: Theoretical controversies and practical implications. *Human Factors, 30*(5), 583-598.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology.* San Francisco: W. H. Freeman.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology,* 1(1), 45-57.

Smith, K., & Hancock, P. A. (1994). Situation awareness is adaptive, externally-directed consciousness. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational awareness in complex systems* (pp. 59-68). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

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 - 3/17). Neuilly Sur Seine, France: NATO - AGARD.

Taylor, R. M., & Selcon, S. J. (1994). Situation in mind: Theory, application and measurement of situational awareness. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational awareness in complex settings* (pp. 69-78). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Tenney, Y. T., Adams, M. J., Pew, R. W., Huggins, A. W. F., & Rogers, W. H. (1992). *A principled approach to the measurement of situation awareness in commercial aviation* (NASA Contractor Report 4451). NASA Langely Research Center.

Wickens, C. D. (1992). *Engineering Psychology and Human Performance* (2nd ed.). New York: Harper Collins.

Maintaining Situation Awareness when Stalking Cognition in the Wild

John M. Flach

Wright State University

The term "situation awareness" originated with tactical fighter pilots as they attempted to articulate the difficulties of managing the complex information processing demands of air combat. The term has more recently been embraced by the human factors community to define a domain of research whose goal is to study cognition as it occurs in complex, dynamic work environments. Military and civil aviation are but two examples of complex, dynamic work environments. Other examples include modern medicine and chemical process control. It might also be argued that *science* is a complex, dynamic work environment (although the time constants are much longer than even process control). If this is true, then it might be useful to recursively apply the question of situation awareness to the science of situation awareness. What traps in the scientific enterprise might lead to loss of situation awareness? What constitutes good situation awareness on the part of the scientist?

Circular Reasoning

Flach (1995), reviewed Underwood's (1957) discussion of scientific reasoning, to highlight one potential trap for the scientist --- the trap of circular reasoning. This occurs when a construct (such as situation awareness) that originates as a description of a phenomenon (such as a pilot's awareness of the threats and opportunities in the environment) becomes an explanation for the very same phenomenon. So, scientists might lose situation awareness, and find themselves explaining a pilot's failure to respond appropriately to a threat or opportunity as being "caused" by the lack of situation awareness.

Einstein and Infeld (1938) have identified such naive scientific reasoning as "substance theories." Thus, for example, thermodynamic phenomenon were, at one time, attributed to a substance called "caloric." Temperature differences were attributed to differential amounts of this substance which somehow traveled from one object to another to account for thermodynamic equilibrium. A scientist, who loses situation awareness, might lose sight of the fact that situation awareness originally described a dynamic cognitive coupling between an actor and a situation. When this occurs, SA becomes both the measure of different degrees of coupling (a dependent variable) and the causal explanation for those differences (an independent variable). It becomes the caloric "substance" that explains differences among individuals and across situations. It should be noted that the calorie remains an important measure for thermodynamic processes. So the empirical observations of caloric differences remain important even though the early theoretic explanations are now considered to be naive.

Piecemeal Reasoning

Newell (1966) once criticized cognitive psychology for playing twenty questions with nature and losing. The basis for this criticism was the observation that the phenomenon of cognition was being parsed in terms of simple dichotomies (e.g., parallel or serial processing). Newell was skeptical that it would be possible to integrate across the various niches defined by these dichotomies to build a complete description of cognition. Newell's concems reflect a potential trap of reductionism in which explanations are sought in terms of fundamental elements --- at some point, reduction to smaller elements will destroy the essence of the phenomenon of interest. It is not a simple question of reductionism or not, but rather a question of a stopping rule. At what point does further division result in breaking --- where the pieces can no longer be assembled in a way that leads to further understanding?

Einstein and Infeld (1938) have identified the next level of sophistication in scientific reasoning, beyond substance theories, as mechanistic theories. The Newtonian program in physics is an example of a mechanistic theory. The problem of mechanistic reasoning is that dichotomies are always present (e.g., mass and energy, matter and force). In the Newtonian program there are the particles or matter and then there is space and time. Space and time are independent from each other and independent from the matter which they encompass. Thus, space and time have an absolute existence independent of matter. At the same time, they play fundamental roles as causal independent variables that determine the behavior (e.g., motion) of the particles. Ray and Delprato (1989) identify the reduction of biological activity to physiochemical causal chains as typical of mechanistic explanations. Ray and Delprato identify mechanistic thinking with "the era of the world machine, mechanism, materialism, causal determinism, and reductionism" (p. 82). The problem with this view is that the "cause" (in terms of energy, or space-time, or physiochemical activity) is some how independent and distinct from the phenomenon explained. It acts like a homunculus just beyond the curtains pulling the puppet's strings.

Philosophy and psychology have struggled with the dichotomy of mind and body. There has been one ontology for mental events and the constraints that govern mind and a second ontology for physical events and the constraints that govern the body. For example, a major criticism of the computer metaphor has been that the constraints that govern the body have been defined as being outside the problem of cognition (e.g., Dreyfus, 1992). The PDP or neural computing movement has tried to acknowledge the physical constraints at a micro-level (i.e., in term of biological constraints) and the ecological movement (e.g., Gibson, 1979) has attempted to recognize the role of physical constraints at a macro-level (i.e., in terms of functional constraints). For human factors this duality shows up in terms of actor and environment. As with the Newtonian view of space and time, the environment tends to be seen as a "container" that exists independent of the actor (i.e., substance). Thus, we have different languages and constructs for the user and the environment.

The very term, situation awareness, seems to demand a theoretical framework that spans the mental (awareness) and the physical (situation) constraints that govern performance. Yet, it seems that a lot of energy is being wasted breaking the phenomenon into even smaller pieces. Debates over whether situation awareness is distinct from workload, or naturalistic decision making and efforts to distinguish SA as product from SA as process are symptomatic of low SA. The phenomenon of SA demands that we move up in abstraction from the information processing model. Just as theories of naturalistic decision making (e.g., Klein, et al., 1994) recognize that decision making in dynamic work environments is intimately coupled with perception and action, theories of SA must recognize that the cognitive coupling between human and dynamic work environment spans the full range of information processing stages. To reduce SA to a distinct stage of information processing, would be a major blunder. Theories of SA must recognize that cognition is both situated within (e.g., Suchman, 1987) and distributed over (e.g., Hutchins, 1995) environments. We must not forget Simon's (197) parable of the ant in which he illustrated the importance of the environment for determining behavioral trajectories We must build theories

that encompass both the reality of the ant and the reality of the beach. As I have argued before (Flach, 1994; Flach & Warren, 1995), a theory of a disembodied mind will never be a theory of "what matters!"

Heisenberg Uncertainty

Heisenberg's uncertainty principle states that we cannot know both the position and momentum of a subatomic particle. The more we know about one, the less we know about the other. However, the implications of this principle are much broader (e.g., see Zukav, 1979). It challenges the myth of the scientist as a passive observer. It reminds us that every act of measurement introduces variance. For physical sciences, the variance due to measurement is often so small, relative to the phenomenon of interest, that it can be ignored. However, at the level of quantum mechanics, the variance of the phenomenon and of the measurement intervention are at comparable scales. Cognitive psychology faces the same problem as quantum mechanics. Our measurements create variance that is at a comparable scale to the phenomena that we study. Thus, the situationally aware cognitive scientist must be alert to the possibilities of demand characteristics and reactivity.

In general, cognitive psychology has been cognizant of the impact of demand characteristics and reactivity associated with self reports. Because self reports have been important for measuring SA (e.g., SART - Taylor, 1995; and SAGAT - Endsley, 1995), there has been much concern about this measurement problem. However, we tend to forget that demand characteristics and reactivity are not specific to self reports. Every experimental intervention produces its own variance. In fact, Howard (1994) has shown that for many situations the construct validity coefficients for self-reports were superior to the validity coefficients of other measurement approaches. Howard (1994) writes:

> for whatever reasons, social scientists frequently recite the litany of 'known problems of self-reports' but rarely do we focus upon the 'known problems of (for example) behavioral measures' . . . . Researchers' suspicions of self-reports (as well as their unjustified confidence in many non-self-report measures) are apparent in criterion validation studies wherein the validity of self-reports is estimated by their correlation with some non-self-report measures of the constructs of interest. Have you ever seen the opposite case --- where a self-report is used as a criterion measure to validate some behavioral index? (p. 400)

The moral is that every act of measurement in cognitive science (observational, verbal reports, behavioral, physiological) interacts with the phenomenon of interest and that the resulting variance is at a scale of magnitude that is comparable to the phenomena of interest. So, the only way we can differentiate the variance associated with the phenomenon from variance associated with the measurement act is to use a set of converging operations that employ a range of different measures. It is important to note that we use the term measure to refer not only to the dependent variable, but to the whole experimental context.

An important implication of situation awareness is that the situational constraints are an important source of variance. If those constraints are not present in the measurement context, then important aspects of the phenomenon will be lost and the measurements will not be representative of the phenomenon. There is a certain irony, related to the Heisenberg problem in cognition. On the one hand there is a wealth of research in social psychology that illustrates that subtle differences in the expectations of the experimenter can have a significant impact on observations in artificial, laboratory contexts (e.g., Rosenthal, 1966). On the other hand, observations by Rasmussen (1994) and Klein (1995) on their interactions with experts in their domains of expertise show that extraordinary efforts to "bias" the experts to behave in ways consistent with the

experimenter's expectations had little impact. In the work domain, the constraints of the situation tend to overwhelm any variance contributed by the expectations of the experimenter. It seems that the further actors are removed from their domains of expertise, the more susceptible they will be to demand characteristics of the experimental situation. The irony is, that the rationale for sterile, controlled laboratory environments is to reduce unwanted sources of variance. With regard to situation awareness, and perhaps cognition in general, the effect may actually be to create a situation that maximizes the potential impact of demand characteristics. Thus, in pristine laboratory settings the experimenters may be seeing little more than their own reflections. This leads naturally to the issue of confirmation bias.

Confirmation Bias

Science is essentially an inductive process. That is, science tries to infer general theories or laws from particular observations. With induction, one contradiction is sufficient to eliminate a potential theory, but no amount of confirmation is sufficient to prove a theory. Yet, humans show a strong bias toward seeking confirmation (e.g., Wason, 1960; Klayman & Ha, 1987). Once a belief or theory is formulated the search and interpretation of information tends to be biased so as to maintain the theory. In more general terms, Norman (1986) has referred to this bias as "cognitive hysteresis." That is, a tendency for humans to hold on to a belief beyond the point where the evidence would warrant it. Functional fixedness (Dunker, 1945), conservatism (Edwards, 1968), attribution (Hastie & Kumar, 1979), and anchoring (Tversky & Kahneman, 1974) are all examples of this cognitive hysteresis.

Cindy Dominguez and I were recently reminded of the confirmation bias by a comment that was made by a surgeon Cindy interviewed as part of an ongoing study of laparoscopic surgery. The surgeon commented, "A good surgeon believes what he sees. A poor surgeon sees what he believes." I think that cognitive science and human factors has often fallen into the trap of seeing what we believe. For example, in studying workload, there was a strong tendency to see the world in terms of the additive factors logic of the Sternberg Task (Sternberg, 1966). It appears as if some believed that understanding a complex task, like landing an airplane, was simply a matter of determining whether the landing task resulted in a slope or intercept effect when paired with the Sternberg Task. In fact, I have heard the remarkable claim that the Sternberg task was a "dip stick into the mind." Clearly, this is naive science. Classical AI has focused almost exclusively on logical puzzles (e.g., cryptarithmatic, theorem proving, tower of Hanoi, etc.) that fit the computer metaphor; as opposed to problems that human do well (e.g., walking, driving cars, putting out fires, etc.). It also appears to me, that theories such as the multiple resource model of workload (e.g., Wickens, 1984) are little more than reifications of the logic of Analysis of Variance in which the patterns of interaction or additivity are given the status of causal explanation. Not only our theories, but our analyses tools can sometimes become blinders that unnecessarily restrict our awareness of the situation. It is not that the observations made in these contexts are not valid or insightful, but they are at best small slices of the total picture. The question, that Newell wrestled with is --- will these slices add up to a complete picture at the end of the day? Is the study of SA just another slice of the total picture or do the problems of SA require a fundamental reorientation of how we attack cognition in the wild?

Of course, the surgeon's comments were a gross oversimplification. There are no "good" surgeons (in the sense of being totally free of biases due to expectations and beliefs). Everyone's perceptions are influenced by their beliefs. Careful surgeons/scientists, however, are aware of the bias and proceed with appropriate caution --- checking and double checking to verify their perceptions. It is really a question of the appropriate balance of theory and empiricism. Watkins (1990) recently questioned whether there was an appropriate balance of theory and empiricism in the study of memory. He wondered whether the theoretical construct of the memory trace was

obfuscating progress in the study of memory. He suggested that the mediational constructs driving the empirical studies were not generating the kind of data needed for incremental progress in understanding memory. He illustrates the problem with the following parable:

> A few years ago, the Psychology Department at Rice University was in need of furniture, and for a time I kept a vigil for pieces discarded by other departments. One day I chanced upon such a piece. It was handsome and well constructed, but at the same time it was complex --- a sort of table but with a two-level top and a rather odd shape. Clearly, it had been made to meet a particular need, one that presumably no longer existed. Notwithstanding all the time and expertise that had gone into its making, the singular nature of this item of furniture rendered it of no use even to those whose needs were great, and it was indeed thrown away. In the same way, research designed to address some person's individual theory is unlikely to be of any use once that person allows the theory to wither and die. (p. 332)

It is interesting to read Wickens' (1992) comparison of SA and workload in light of Watkins' parable:

> In the same way that the seeds of workload research were planted and nourished in the late 70s and grew to full bloom in the 80s, so, a decade later, the seeds of applied interest in situation awareness were planted in the mid 80s, and I forecast will grow and bloom in the 90s (I'll leave the "withering and dying on the vine" part of the analogy for others to speculate). (p. 1)

Thus, we can appreciate the quality and elegance of Sternberg's model or the multiple resource model, but at the same time wonder whether the huge amount of data that has been motivated by these models has led to a commensurate gain in understanding. I hope that in studying SA, researchers are more cautious and don't let the nuances of overly refined theories lead to a restrictive perspective on the phenomenon of interest and to an empirical base that will be discarded in the wake of the next fad.

Hindsight Bias

Hindsight, of course, is 20-20. Research shows that people overestimate their ability to predict outcomes after the fact (Fischhoff, 1982). Thus, it is easy to criticize the information processing model or the workload research that was motivated by Stemberg's model in retrospect. It is interesting to read Einstein's (1954) discussions of Leibniz's and Huygens' view of space as (a) place - a property of an object versus Newton's view of space as (b) a container with an independent existence from the objects encompassed. Einstein writes:

> The concept of space was enriched and complicated by Galileo and Newton, in that space must be introduced as the independent cause of the inertial behavior of bodies if one wishes to give the classical principle of inertia (and therewith the classical law of motion) an exact meaning. To have realized this fully and clearly is in my opinion one of Newton's greatest achievements. In contrast with Leibnitz and Huygens, it was clear to Newton that the space concept (a) was not sufficient to serve as the foundation for the inertia principle and the law of motion. He came to this decision even though he actively shared the uneasiness which was the cause of the opposition of the other two: space is not only introduced as an independent thing apart from material objects, but also is assigned an absolute role in the whole causal structure of the theory. This role is absolute in the sense

that space (as an inertial system) acts on all material objects, while these do not in turn exert any reaction on space.

The fruitfulness of Newton's system silenced these scruples for several centuries. Space of type (b) was generally accepted by scientists in the precise form of the inertial system, encompassing time as well. Today one would say about the memorable discussion: Newton's decision was, in the contemporary state of science, the only possible one, and particularly the only fruitful one. But the subsequent development of the problems, proceeding in a roundabout way which no one could possibly foresee, has shown that the resistance of Leibnitz and Huygens, intuitively well founded but supported by inadequate arguments, was actually justified. (pp. xiv - xv)

Although Einstein's theory of relativity required a reassessment of Newton's assumption of absolute space and time, Einstein did not fail to appreciate Newton's contribution. It is prudent to take a similar view of the information processing model. It was the right program for its time. It was the most productive choice. The intuitively well founded concerns of Gibson (1979), Neisser (1976) and others are only now beginning to bear fruit. The problem here is not a matter of giving adequate credit to the scientists on whose shoulders we stand, but whether the success of the information processing model will be a stepping stone to a deeper awareness or will it be a girdle, restricting the expansion of awareness? My fear is that efforts to preserve the information processing model by identifying SA as yet another box in the processing stream will be an obstacle to full awareness. Perhaps, it is time to move beyond the computer and communications channel metaphors to the metaphor of adaptive, dynamic systems (e.g., Kelso, 1995). In doing so, we should not devalue the contribution of the earlier metaphors.

Unified Field Theory

Einstein and Infeld (1938) suggested that field theories represent a higher degree of sophistication than either "substance" or "mechanistic" theories. Whereas, substance and mechanistic theories emphasize the reality of particles (one electron acting on another), field theories focus on higher order constraints (i.e., fields) as the determinants of behavior and thus the objects of study. "The field thus becomes an irreducible element of physical description, irreducible in the same sense as matter in the theory of Newton" (Einstein, 1961, p. 150). Space and time exist only as dimensions of this field, not as an objective container existing independent of the field. It can be useful to think in terms of the relation between prey (fox) and predator (rabbit). Does the fox determine the rabbit population or vice versa? Neither! Population is recursively determined by the coupling between prey and predator. One is not the cause of the other, rather they are interacting constraints that together bound the trajectory of the prey/predator system. The relation between prey and predator is understood as properties of a unified system, not as an interaction of distinct particles. Thus, the interaction of two electrons is understood in the context of interacting fields that mutually constrain the behavioral trajectories. Ray and Delprato (1989) quote Kantor's (1959) description of field as an alternative to mechanistic causation:

> All creative agencies, all powers and forces, are rejected. An event is regarded as a field of factors all of which are equally necessary, or, more properly speaking, equal participants in the event. In fact, events are scientifically described by analyzing these participating factors and finding how they are related (p. 90).

Is a field theory of cognition possible? Progress is already being made. Gibson's (1979) concepts of optic flow fields and perceptual arrays and Gibson and Crook's (1938) construct of "safe field of travel" represent steps toward a field theory. Kugler and Turvey's (1987) insect nest

building metaphor and their work on coordination is an important illustration of how complex behaviors can be generated as a result of coupling force and information fields. The application of nonlinear dynamics theory to motor coordination (e.g., Kelso, 1995) continues and expands this theme. Also, work on artificial life (e.g., Langdon, 1989) is taking a similar perspective. Brunswik's (1956) Lens model represents another important description of the relational constraints that determine perception. Kirlik (1995) extends Brunswik's intuition to the problem of design. Descriptions of how constraints join to bound the space of possibilities for cognitive work is a central theme in Rasmussen, Pejtersen, and Goodstein's (1994) discussion of cognitive systems engineering. Hutchins (1995) describes distributed cognition in terms of constraint satisfaction. Additionally, Engstrom's (1993) "activity theory" approaches performance in terms of activity systems that integrate the "subject, the object [problem space], and the instruments (material tools as well as signs and symbols) into a unified whole" (p. 67).

Flach and Dominguez (1995) have suggested that it might be useful to distinguish two classes of constraint that together bound the field of possibilities for control and adaptation in dynamic work environments --- constraints on action and constraints on information. These constraints which are distributed over the human and environment will be the "objects of study" for a field theory of cognition or situation awareness. Rather than thinking in terms of the "causes" of behavior, the focus should be on the boundary conditions that limit the field of possibilities. The constraints that define these boundary conditions are the objects of design in human-machine systems. Engineering and design are processes of shaping these constraints to the demands of particular functions. A science of cognition that focuses on the information and action constraints will have a common ground for communicating with designers. Design implications will no longer be an afterthought, but will be a central and natural consideration for basic research in cognition

In Sum

> In the attempt to achieve a conceptual formulation of the confusingly immense body of observational data, the scientist makes use of a whole arsenal of concepts which he imbibed with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly ever, and in any case not seriously, to be doubted. How could he do otherwise? How would the ascent of a mountain be possible, if the use of hands, legs, and tools had to be sanctioned step by step on the basis of the science of mechanics? And yet in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them. This becomes evident especially in those situations involving development of ideas in which the consistent use of the traditional fundamental concepts leads us to paradoxes difficult to resolve (Einstein, 1954, p xi - xii).

Sometimes psychologists act as if the essence of science is "data." But Einstein's quote reminds us of the importance of ideas and assumptions and the importance of constantly reassessing the fundamental assumptions guiding the empirical work. As psychologists, we imbibed with our mother's milk notions of cause-effect, the notion of mind independent of matter, the Sternberg Task, the Analysis of Variance, the importance of control over the stimulus, suspicions of verbal reports, constructs of information processing stages, intelligence, memory trace, schemas, mental model, intelligence etc. Perhaps SA will be the paradox that leads us to question whether these concepts are loyal servants or tyrannical rulers.

In considering confirmation bias, Hutchins (1995) asks an important question: "A property of cognitive processing that prevents us complex creatures from finding better interpretations once we have a good one seems very maladaptive indeed. Why then should such a property survive?" (p. 240). Hutchins' response to this question was that this maladaptive property of the individual may actually be an adaptive virtue on the group level. Science is a group exercise. One of the great promises of SA, that I think was evident at the Daytona meeting, is that the field of SA will be an important arena where theories of cognition will be sorely tested. It is evident from the discussions that people are strongly committed to various beliefs. The resulting tensions are driving us toward the chaotic edge of cognitive science where there is the greatest promise for new levels of organization that may lead to greater degrees of awareness. The phenomenon of SA stands as both an important challenge for the basic science of cognition and as an invitation to the next plateau in our growing understanding of performance in human-environment systems.

Acknowledgments

Sincere thanks to Dan Garland and Mica Endsley for the opportunity to participate in the *International Conference on Experimental Analysis and Measurement of Situation Awareness.* It was an excellent conference and I learned a great deal from the presentations and discussions. Thanks to Cindy Dominguez for comments on an early draft of this proposal. John Flach was supported by grants from the Air Force Office of Scientific Research during preparation of this manuscript. Opinions expressed in this article are John's alone and do not represent an official position of the Air Force or any other organization.

References

Brunswik, E. (1956). *Perception and the Representative Design of Psychology Experiments.* Berkeley: University of California Press.

Dreyfus, H.L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason.* (2nd edition). Cambridge, MA: MIT Press.

Dunker, K. (1945). On problem solving. *Psychological Monographs,* 58(5), Whole no. 270.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (ed.). *Formal Representations of Human Judgment.* New York: Wiley.

Einstein, A. (1954). Forward. In M. Jammer. *Concepts of Space: The History of Theories of space in Physics.* Cambridge, MA: Harvard University Press.

Einstein, A. (1961). *Relativity.* New York: Crown Publishers, Inc.

Einstein, A. & Infeld, L. (1938). *The Evolution of Physics.* New York: Simon & Schuster.

Endsley, M. (1995). Measurement of situation awareness in dynamic systems. *Human Factors,* 37, 65 - 84.

Engstrom, Y. (1993). Developmental studies of work as a testbench of activity theory: The case of primary care medical practice. In S. Chaiklin & J. Lave (eds.). *Understanding Practice: Perspectives on Activity and Context.* Cambridge, England: Cambridge University press.

Fischhoff, B. (1982). For those condemned to study the past: Heuristics and biases in hindsight. In D. Kahneman, P. Slovic, and A. Tversky (eds.). *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge, England: Cambridge University Press.

Flach, J.M. (1994). Ruminations on mind, matter, and what matters. *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting.* (p. 531 - 535). Santa Monica: The Human Factors and Ergonomics Society.

Flach, J.M. (1995). Situation awareness: Proceed with caution. *Human Factors,* 37, 149 - 157.
Flach, J.M. & Dominguez, C.O. (1995). Use-centered design. *Ergonomics in Design,* July, 19 - 24.
Flach, J.M. & Warren, R. (1995). Active psychophysics: The relation between mind and what matters. In J.M. Flach, P.A. Hancock, J.K. Caird, & K.J. Vicente (eds). *Global Perspectives on the Ecology of Human-Machine Systems.* (pp. 189 - 209). Hillsdale, NJ: Erlbaum.
Gibson, J.J. (1979). *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.
Gibson, J.J. & Crooks, L.E. (1938). A theoretical field analysis of automobile driving. *American Journal of Psychology,* 51, 694 - 703. Also reprinted in E. Reed & R. Jones (eds.) *Reasons for Realism.* (pp. 119 - 136). Hillsdale, NJ: Erlbaum.
Hastie, R. & Kumar. P. (1979). Person memory: Personality traits as organizing principles in memory for behavior. *Journal of Personality and Social Psychology,* 37, 25 - 38.
Howard, G.S. (1994). Why do people say nasty things about self-reports? *Journal of Organizational Behavior,* 15, 399 - 404.
Hutchins, E. (1995). *Cognition in the Wild.* Cambridge, MA: *MIT* Press.
Kantor, J.R. (1959). *Interbehavioral psychology.* (2nd ed). Granville, OH: Principia Press.
Kelso, J.A.S. (1995). *Dynamic Patterns.* Cambridge, MA: MIT Press.
Kirlik, A. (1995). Requirements for psychological models to support design: Toward ecological task analysis. In J.M. Flach, P.A. Hancock, J.K. Caird, & K.J. Vicente (eds). *Global Perspectives on the Ecology of Human-Machine Systems.* (pp. 68 - 120). Hillsdale, NJ: Erlbaum.
Klayman, J. & Ha, Y.W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review,* 94, 211 - 228.
Klein, G.A. (1995). Situation Awareness Process Assessment. Panel presentation at the *International Conference on Experimental Analysis and Measurement of Situation Awareness.* Daytona Beach, FL: Embry-Riddle Aeronautical University.
Klein, G. A., Orrsanu, J., Calderwood, R., & Zsambok, C. (1993). *Decision Making in Action: Models and Methods.* Norwood, NJ: Ablex.
Kugler, P.N. & Turvey, M.T. (1987). *Information, Natural Law, and the Self-assembly of Rhythmic Movement.* Hillsdale, NJ: Erlbaum.
Langdon, C.G. (ed.) (1989). *Artificial Life.* Redwood City, CA: Addison Wesley.
Neisser, U. (1976). *Cognition and Reality.* San Francisco: Freeman.
Newell, A. (1977). You can't play 20 questions with nature and win. In W. G. Chase (Ed.). *Visual Information Processing.* Academic Press.
Norman, D.A. (1986). New vies of information processing: Implications for intelligent decision support systems. In E. Hollnagel, G. Mancini, & D.D. Woods (eds.). *Intelligent Decision Support in Process Environments.* New York: Springer-Verlag.
Rasmussen, J. (1994). Modeling complex adaptive systems for design of work support. Keynote address 1994 *Symposium on Human Interaction with Complex Systems.* Greensboro, NC: University of North Carolina A & T.
Rasmussen, J. Pejtersen, A.M. & Goodstein, L.P. (1994). *Cognitive Systems Engineering.* New York: Wiley.
Ray, R.D. & Delprato, D.J. (1989). Behavioral systems analysis: Methodological strategies and tactics. *Behavioral Science,* 34, 81 - 127.
Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research,* New York: Appleton-Century-Crofts.
Simon, H. A. (1981). *The Sciences of the Artificial.* (2nd edition). Cambridge, MA: MIT Press.
Sternberg, S. (1966). High speed scannin in human memory. *Science,* 153, 652 - 654.
Suchman, L. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication.* Cambridge University Press.
Taylor, R.M. (1995) Subjective measurement techniques. Presentation at the *International conference on Experimental Analysis and Measurement of Situation Awareness.* Daytona Beach, FL: Embry-Riddle Aeronautical University.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Hueristics and biases, *Science*, 185, 1124 - 1131.
Underwood, B.J. (1957). *Psychological Research.* Englewood Cliffs, NJ: Princeton Hall.
Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12, 129 - 140.
Watkins, M.J. (1990). Mediationism and the obfuscation of memory. *American Psychologist*, 45, 328 -335.
Wickens, C.D. (1984) *Engineering Psychology and Human Performance.* Columbus: Merrill.
Wickens, C.D. (1992). Workload and situation awareness: An analogy of history and implications. *Insight,* 14,(4), 1 - 3.
Zukav, G. (1979). *The Dancing Wu Li Masters.* New York: Quill.

Expert Performance and Situation Awareness

Neil Charness

The Florida State University

Expertise and Situation Awareness: Relationships

My task, as an outside observer, is to bring my background knowledge to bear on the construct of situation awareness (SA). My research interests focus around aging and expertise. I am particularly interested in trade-offs between these two variables for performance. In the literature concerned with individual differences, expertise and age are probably the top two factors, far surpassing more controversial ones such as sex or race. Average effect sizes range from .5 to 10 SD units. A good example is provided by the competitors in the recent chess world championship match. Garry Kasparov and Viswanathan Anand are about 8 standard deviation units above a starting tournament player, as measured by the Elo chess rating scale (Elo, 1986).

Age, usually, but not always, is associated with declines in performance for adults. If I want to predict how long it will take a 70 year-old to perform a speeded task, compared to a 20 year-old, multiply the younger performer's response time by 1.5 (e.g., Hale & Myerson, 1995).

The SA literature is not typically concerned with age[1], though there are significant relations between age, experience, and aircraft crashes (e.g., Tsang, 1992) and between age and pilot performance on navigation-related tasks (e.g., Taylor et al., 1994). Thus, I will concentrate on linking just expertise and situation awareness. My goal is to outline parallels to my own field that may help to flesh out SA. My assumption is that for almost all the tasks that have been discussed in SA, the practitioners are probably experts, at least if you use a criterion of hours of training necessary to perform the task skillfully.

Tenets of Situation Awareness

From my rather brief reading of the situation awareness (SA) literature, it is apparent that there are many different views of its definition:

- Endsley (1988): "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (p. 97).
- Sarter and Woods (1991): "the accessibility of a comprehensive and coherent situation representation which is continuously being updated in accordance with the results of recurrent situation assessments" (p. 52).
- Smith and Hancock (1995): "adaptive, externally directed consciousness" (p. 137).

[1] Endsley (1995b) reports experiments with retired pilots with a mean age of 45 to 48 (age ranges of 32-68). We can expect performance in these studies to be influenced by both age and skill.

- Adams, Tenney, and Pew (1995): "up-to-the-minute cognizance required to operate or maintain a system" (p. 85).

With such diversity in definition it is difficult to decide how to operationalize the construct. In this respect, I don't think that SA is necessarily more problematic than closely aligned constructs such as mental model (for problem solving) or situation model (for text comprehension). More recent theoretical efforts to expand the definition of SA and pinpoint its components (e.g., Endsley, 1995a) seem promising. Others more competent than I have offered cogent critiques of its definition and use (e.g., Flach, 1995). It seems necessary to avoid the problems that arise when you define the construct both too narrowly or too broadly. An example of the latter would be Searle's (1980) definition of *understanding* with his famous Chinese Room example which required the use of the additional ill-understood term *intentionality,* thereby restricting understanding to humans. You may or may not be happy with the notion of an auto-pilot system showing SA, but it is worth entertaining and expanding this notion.

The construct of SA seems to have originated in the aviation field primarily in the context of accident analysis. It appears to me to be a "default" construct. Namely, that you know it best when it fails, when someone loses SA and the result is a crash. It seems somehow easier to say that someone lacks situation awareness than that they possess it. Default definitions have their place in science. For instance, the medical categories of Alzheimer's disease and Parkinson's disease both require as part of their definition that competing disease processes are ruled out (American Psychiatric Association, 1994) before you default to these categories.

On the other hand, for me, one of the key features of tasks that are currently associated with SA is the complex and changing nature of the data stream that the operator must process. Usually the task requires splitting attention between information sources and/or spatial locations, as in the case of piloting an aircraft. Human beings are notoriously poor at parallel processing when it comes to reasoning and decision-making, so they are usually forced into serial time-sharing strategies. Humans seem more comfortable at breaking complex tasks into simpler ones via serial sub-goaling than juggling multiple contingencies simultaneously.

One way to approach the problem is to recall some earlier research on concepts. Research on categories such as "bird", distinguish between: *Characteristic* or *Prototypic* features, such as small, sings, flies (exceptions: ratite birds such as ostrich, kiwi don't fly, and the former is large), and *Defining* features, such as *has feathers*, eggs develop outside body, warm-blooded, 4-chambered heart. Defining features for SA are lacking and there may be poor agreement on prototypic ones.

In the spirit of trying to stretch the construct of SA a bit, let me pose a set of hypothetical questions, to be answered on a 1-5 scale from 1, completely unimportant, to 5, extremely important. How important is SA in:

Table 1. Activities involving different degrees of SA.

ACTIVITY	RATING (1-5)
eating a banana	
finding an open pass receiver	
breathing	
performing a 2-choice RT task	
choosing a move in a chess game	
diagnosing a disease in a patient	
piloting an aircraft	
comprehending a novel	
tying your shoelace	

Although there is perhaps little agreement on the formal definition of SA, I suspect that this quiz would reveal high reliability in assessing whether it is or is not associated with a particular set of human performance tasks. Other constructs in psychology such as attention, consciousness, and even expertise probably share this characteristic.

Another concern is that little attention has been paid to the idea of *discriminant validity* for the construct of SA. That is, how does SA differ from historically prior constructs, such as skill or expertise?

The field of expertise research offers a promising model for understanding the construct of situation awareness. I'll discuss three areas of potential overlap: definition of the field, theoretical models, and research methodology.

Tenets of Expertise: Definition of the Construct

How has the field of expertise handled similar definition problems? My colleague, Anders Ericsson, and I (e.g., Ericsson & Charness, 1994) offered the following definition for expertise: "Consistently superior performance on a specified set of representative tasks for the domain that can be administered to any subject." We went on to argue that superior performance could be bounded by the notion of "outlier", a performance level that is at least 2 standard deviations above the mean in the domain population. In chess, for instance, using the United States Chess Federation's rating system, an expert level chess player (rating of 2000-2199 rating points) is about 2 standard deviations above the mean[1] of the rated chess playing population.

The term "consistent" was chosen to help constrain expertise to those who can show superior performance on a regular basis. Thus, we would rule out one-time achievements such as a 50% annual growth rate by a particular money manager as indicative of expertise unless that return were consistently above the average market return over multiple years.

The choice of the term "representative tasks" was intended in part to ensure that the superior performance was ecologically valid and also to ensure that there would be agreement from the practitioners in the field that we are looking at the critical tasks. (We are heeding an injunction by Bryan and Harter, 1899, who urged that our enterprise be seen as valid by those who engage in the profession.)

I'm not sure that SA can be equally circumspectly defined, but it is worth trying. I do suspect that defining representative tasks from the domain would help flesh out SA.

Theoretical Models

I tend to subsume expertise in part under the more inclusive topic of problem solving (e.g., Newell & Simon, 1972). Typically, you become expert at solving the critical problems in your domain. Newell and Simon developed a useful framework for understanding the problem solving process. They envision problem solving as occurring within a symbolic information processing system. They coined the term "problem space" to describe the mental space occupied by the problem description, the goal element(s), the methods for changing from state to state, and the knowledge governing the selection of methods and the evaluation of a state. They envisioned problem solving primarily as a search process in this mental space. What makes most problems difficult is the size of the problem space, which in turn necessitates heuristic search processes since it would be almost impossible to use a generate-and-test procedure to discover the path to the goal state.

What the expertise literature has contributed to the theory of problem solving is the importance of knowledge in constraining search (viz. Feigenbaum's 1994 ACM A. M. Turing Award for expert systems work). In extreme cases, experts solve problems by recognition, rather than by search. In many cases they use knowledge to constrain search to a forward-branching process that

---

[1] The mean is a bit less than 1600 rating points, with the standard deviation intended to be about 200 rating points, the size of the rating class interval, though the SD has waxed and waned depending on the number of young players entering the rating pool in the USA.)

generates only the necessary intermediate states en route to the solution (e.g., for physics problems: Simon & Simon, 1978). If experts are forced to deal with very difficult problems with severe constraints on knowledge accumulation, such as incremental presentation of symptoms in medical diagnosis (Patel, Arocha, & Kaufman, 1994) they may be forced to do the less efficient and more memory-demanding backward reasoning (means-ends reasoning).

There are some intriguing parallels between features of SA and recognition-based problem solving. That is, someone who has SA should be able to recognize critical events in the data stream, whereas someone who lacks SA would be expected to miss those events, or fail to take appropriate action. We might want to entertain the hypothesis that skilled performers have large vocabularies of recognizable chunks that comprise the most critical features of the operator's task. That large vocabulary of "condition-action" rules may be an important component of SA. Time-sharing skill, the ability to interleave two processes, might be another important component.

Expertise by-passes the traditional limitations of information processing

One of the most interesting features of human performance is its limitations. The cognitive revolution was fomented in part by the finding that information transmission through the human operator was limited. In his justly lauded paper, George Miller (1956) suggested that the currency of the realm was chunks, not bits. Others also argued persuasively (e.g., Broadbent, Simon) that humans were limited in their ability to adapt to the environment's demands. Errors (a departure from perfect rationality) were the inevitable result of information overload. As many human factors specialists have pointed out, with the increases in the complexity of the information displays in such complex systems as nuclear reactors and aircraft cockpits, we come perilously close to overtaxing human adaptive capabilities.

What is fascinating about the expertise literature is the discovery of how people circumvent basic information processing capabilities. From Lashley's (1951) analysis of the problem of serial order for pianists generating trills to current day analyses of transcription typing (Salthouse, 1986), it has become clear that with sufficient practice skilled performers can re-organize their behavior to bypass basic limitations in information processing rates. Expert typists use prepared overlapping movements to generate inter-keystroke intervals that are shorter than those for simple reaction times. Expert memorists can accept the presentation of random digits at a rate of 1 per second and recall more than 100 digits after a year or two of practice (Ericsson, Chase & Faloon, 1980), far in excess of the 7 ± 2 digits that the rest of the population can manage. They learn to attach encoded digit groups to pre-compiled hierarchical retrieval structures.

The evidence for the extent of adaptation with practice is most graphic for the physiological adaptations made by skilled athletes in terms of muscle type, heart size, and bone size and density (Ericsson & Charness, 1994). Compare the asymmetry in arm size of elite tennis players for their racquet arm and their non-racquet arm.

Even skilled readers who comprehend text passages manage to maintain and manipulate situation models containing elements far in excess of Miller's magic number (Ericsson & Kintsch, 1995). Retrieval structures, pinpointed as the means to the memorist's ends, apparently also play a critical role in normal comprehension processes. It is probably for this reason that theorists working with ambitious cognitive architectures, such as Anderson's (1983) ACT model, or Newell's (1993) SOAR model, have endowed their working memory sub-system with capacity well in excess of seven chunks.

Such escapes from normal limits are not restricted to cognitive and motor components only. Perceptual processing also shows the impact of skill. A good example is shown in the apparent parallel extraction of chess relations by skilled chess players. We (Reingold & Charness, 1995) asked skilled chess players and novice players to decide as quickly as possible whether a King was in check on a 3 x 3 portion of chessboard that subtended a visual angle of about 9 degrees. (There may be some analogy here with a fighter pilot's rapid recognition of the threat inherent in a given volume of space.) There could be one or two potential attackers present in the diagram. Not

surprisingly, skilled players were faster and more accurate than less skilled ones, though accuracy was near ceiling.

The main point to the study, though, was to understand the micro-structure of perception through eye-tracking. We showed that when we used chess symbols (as compared to letters designating chess pieces), highly skilled players made their decision in almost 20% of the cases without moving their eye from the initial fixation point at the center of the display. They apparently extracted the check relation para-foveally. Weaker players had to make direct fixations of the pieces before making their decision. When we changed the displays to letter symbols instead of chess piece symbols, even the experts were forced to fixate the pieces directly much more often. Such results parallel the findings with skilled sports athletes. The expert can both accurately encode a game situation and prepare an appropriate response much more quickly than their less-skilled counterpart (Abernethy, 1987).

It is not always the case, though, that the expert is advantaged in managing multiple locations or multiple streams of data. An intriguing exception is seen in the work of Britton and Tesser (1982). They demonstrated that experts performed worse than non-experts in a divided attention task (tone detection as the secondary task). Experts seemed more engaged or captured (Stroop-like?) by their primary activity, such as choosing a move in a chess game, and hence were slower to respond to the secondary task than were non-experts. Contrary to expectations, the experts did not have more capacity freed up to devote to the secondary task. Could such "capture" by the current primary task be partly responsible for failures in SA by skilled pilots?

Should we use the term SA to describe a skilled chess player or an expert athlete? Or can we get by without introducing this construct? Certainly a skilled chess player is not dealing with as rapidly fluctuating a data stream as an aircraft pilot. Moves occur on average every 3 minutes, though in cases of time trouble they are made in seconds. Nonetheless, when masters make errors, they may indeed be attributable to a failure to comprehend the situation and anticipate their opponent's plans, or to a failure to recall prior analysis when they were searching for the best move to make. There are many anecdotal reports (Krogius, 1976) of gross blunders that suggest a failure to update the problem space accurately resulting in a "hallucination" about the position of a critical piece in a long sequence of planned moves.

Athletes clearly do have to know whether they should be on offense or defense in a fast-moving game such as hockey since a failure to register that state quickly can result in a breakaway by their opponents. The quarterback who fails to read a defense accurately and throws an interception can be classified as having impaired SA by way of failing to update a situation model in a timely fashion.

Models of Skill Acquisition and The Importance of Deliberate Practice

Below in Figure 1 is a framework that we have found useful (Charness, Mayr, & Krampe, in press) for understanding expertise.

One critical finding in research on expertise is that only certain types of practice are effective at developing expert performance, namely, the class of activities we call deliberate practice. Important characteristics of deliberate practice are its demanding nature, the need for full concentration (for about an hour), necessity of resting, limited duration (4 hr per day), and the fact that, at least in music, it is not perceived as inherently enjoyable[1]. As well, access to top-level instructors and coaches and training settings seems critical in some domains.

If SA turns out to be not binary, but a matter of degree, we will need to know more about its development and maintenance. In a sense, SA could be paralleling the development of the field of expertise, where earliest concerns were with expert/novice distinctions and the definition of the construct, and more recently turned to understanding the skill acquisition process.

---

[1] Exceptions seem to occur in wrestling and figure skating where access to mat time and ice time are pretty restrictive

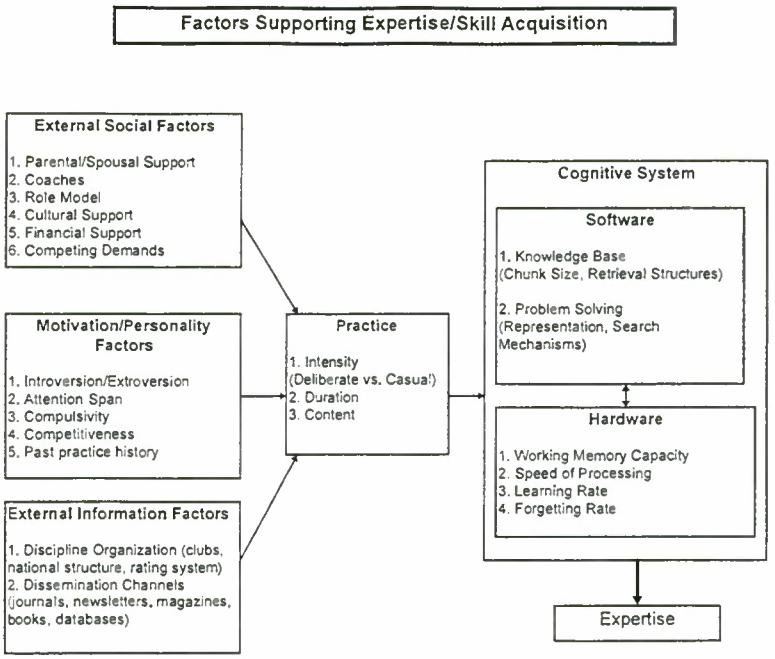


Figure 1. A framework for understanding the factors affecting expert performance.

Methodology Parallels

As outlined in Sarter and Woods (1995) (and in similar ways in Endsley, 1995b), there are three major methods of assessing SA, which they termed subjective ratings, explicit performance measures, and implicit performance measures. The subjective ratings technique asks people to rate their degree of SA and is not generally recommended since its main use, determining calibration, relies on combining self-ratings with other measures. Also, there is evidence that people cannot accurately report on subjective states, and SA surely falls into that category. People asked to indicate how close they are to solution of insight problems are often very poorly calibrated (Metcalfe, 1986). Also, there is a cogent argument from theoretical models developed in the context of protocol analysis that experts may be particularly poor at describing processes which are automated for them but are not automated in novices (Ericsson & Simon, 1993).

Explicit performance measures include stopping someone during an on-line task (e.g., flying in a simulator) and querying them on the status of variables in the current situation (e.g., Endsley, 1995b). There is some concern that querying the operator might change the nature of the processes, though negative findings in Endsley (1995b) for outcome measures such as aircraft kills argue weakly against this. The evidence is weak because of small sample sizes and a finding

of no difference. This concern is somewhat akin to concerns in the use of concurrent protocol analysis, a much favored technique in the expertise area for understanding problem solving (Ericsson & Simon, 1993). A fair amount of experimental work has shown that concurrent protocols can be given without changing the process a great deal (e.g., some slowing can be expected) when certain constraints are met, such as the easy availability of language for describing the current contents of working memory. So, it is fair to say that the literature on protocol analysis can be helpful in deciding how to organize procedures to query aspects of situation awareness without changing the ongoing process.

The final technique, implicit performance measures, generates events that probe whether the operator is sensitive to critical aspects of the situation. If they are aware they will respond differently than if they are not aware. Again, this technique has been used in the expertise literature, to show, in some instances, that experts make encoding mistakes that novices cannot (Adelson, 1984; Arkes & Freedman, 1984). Most of these techniques seem useful for converging on the construct of SA and have already proven fruitful in expertise research.

Conclusions

SA may prove to be a useful construct in understanding human performance failures in complex task environments requiring highly skilled operators. It may also prove to be a useful adjunct to research on expert performance by filling in important gaps about time-sharing abilities and their impact in rapidly changing situations, such as sports settings. Techniques for understanding SA have important counterparts in the expertise area. Both fields could benefit from cross-fertilization of theories and methods.

References

Abernethy, B. (1987). Anticipation in sport: A review. *Physical Education Review, 10*, 5-16.

Adams, M. J., Tenney, Y. J., & Pew, R. W. (1995). Situation awareness and the cognitive management of complex systems. *Human Factors, 37*, 85-104.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, Mass.: Harvard University Press.

Adelson, B. (1984). When novices surpass experts: The difficulty of a task may increase with expertise. *Journal of Experimental Psychology: Learning Memory and Cognition, 10*, 483-495.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Association.

Arkes, H. R. & Freedman, M. R. (1984). A demonstration of the costs and benefits of expertise in recognition memory. *Memory and Cognition, 12*, 84-89.

Britton, B. K. & Tesser, A. (1982). Effects of prior knowledge on use of cognitive capacity in three complex cognitive tasks. *Journal of Verbal Learning and Verbal Behavior, 21*, 421-436.

Bryan, W. L. & Harter, N. (1899). Studies in the telegraphic language. The acquisition of a hierarchy of habits. *Psychological Review, 6*, 345-375.

Charness, N., Mayr, U., & Krampe, R. (in press). The role of practice and coaching in entrepreneurial skill domains: An international comparison of life-span chess skill acquisition. In K. A. Ericsson (Ed.) *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games*. Erlbaum.

Elo, A. E. (1986). *The rating of chessplayers, past and present*, 2nd Edition. New York: Arco.

Endsley, M. A. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*, 32-64.
Endsley, M. A. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors, 37*, 65-84.
Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*, 211-245.
Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). Cambridge, MA: MIT Press.
Ericsson, K. A., Chase, W. G. & Faloon, S. (1980). Acquisition of a memory skill. *Science, 208*, 1181-1182.
Flach, J. M. (1995). Situation awareness: Proceed with caution. *Human Factors, 37*, 149-157.
Krogius, N. (1976). *Psychology in chess*. New York: RHM Press.
Lashley, R. S. (1951). The problem of serial order in behavior. In L A. Jeffress (Ed.), *Cerebral mechanisms in behavior*. New York: Wiley, 1951.
Hale, S., & Myerson, J. (1995). Fifty years older, fifty percent slower? Meta-analytic regression models and semantic context effects. *Aging and Cognition, 2*, 132-145.
Metcalfe, J. (1986). Feeling of knowing in memory and problem solving. *Journal of Experimental Psychology: Learning, memory, and Cognition, 12*, 288-294.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.
Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
Patel, V. L., Arocha, J. F. & Kaufman, D. R. (1994). Diagnostic reasoning and medical expertise. *The Psychology of Learning and Motivation, 32*, 186-252.
Reingold, E. & Charness, N. (1995). Perceptual Automaticity in Chess Skill: Evidence from Eye Movements. Poster presented at the *36th Annual Meeting of the Psychonomic Society*, Los Angeles.
Salthouse, T. A. (1986). Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological Bulletin, 99*(3), 303-319.
Sarter, N. B., & Woods, D. W. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors, 37*, 5-19.
Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*, 417-457.
Simon, D. P. & Simon, H. A. (1978). Individual differences in solving physics problems. In R. Siegler (Ed.), *Children's thinking: What develops?* (pp. 325-348). Hillsdale, NJ: Lawrence Erlbaum Associates.
Smith, K., & Hancock, P. A. (1995). Situation awareness is adaptive, externally directed consciousness, *Human Factors, 37*, 137-148.
Taylor, J. L., Yesavage, J. A., Morrow, D. G., Dohlert, N., Brooks, J. O. III, & Poon, L. W. (1994). The effects of information load and speech rate on younger and older aircraft pilots' ability to execute simulated air-traffic controller instructions. *Journal of Gerontology: Psychological Sciences, 49*, P191-P200.
Tsang, P. S. (1992). A reappraisal of aging and pilot performance. *International Journal of Aviation Psychology, 2*, 193-212.

Experimental Analysis and Measurement of Situation Awareness: A Commentary

David Meister

Introduction

It is an awesome responsibility to attempt to review an area of investigation, and one must be properly cautious in doing so. And since the topic of situation awareness (SA) has already been extensively critiqued, there is also the problem of trying to say something that has not already been said.

The reviewer's biases should also be made visible before he begins. In my case, I must say that I am not much impressed with theories and models, since in my experience (based also on some historical and epistemological research on ergonomics that I am presently conducting) they have little real as opposed to perceived impact on research and design.

I am, moreover, highly pragmatic, having spent much of my working life in design and in close contact with engineers. Hence, I have been conditioned to believe that if the theory/model/research cannot be applied in some meaningful fashion to the solution of a design problem, the theory/model/research impresses me in much the same way as I view art, to be appreciated only for its aesthetic qualities.

The SA Model

In the case of SA the theorizing and model making as described in the 1995 issue of *Human Factors* (Adams, et. al., 1995; Endsley, 1995a, b; Gaba, et. al., 1995; Sarter and Woods, 1995; and Smith and Hancock, 1995) are very elegant and from a theoretical perspective, quite convincing. If at one time I had conceptualized SA as merely a form of attention, the theories and models proposed in this symposium and in earlier papers on this topic have certainly disabused me of this notion. Attention is part of the SA model, of course, but so also are goals, perception, mental models, meaning, short and long term memory, information processing, and decision making (DM), so much so that SA seems to relate to almost all task–oriented and problem solving behavior. If one believes, as I do, that all behavior forms systems of greater or lesser comprehensiveness, I approve.

Unfortunately, the variables in the SA models (using Endsley, 1995a, as an example), are almost exclusively behavioral, at a very high level of abstraction, and represent major functional entities (e.g., perception, memory) rather than dynamic processes, so that they seem to sit very passively in the model.

This makes it difficult (but not impossible) to conceptualize how they work in dynamic situations and has a less than desirable effect on the specificity of the experimental hypotheses one can develop from them. For example, the hypotheses described by Endsley (1995a, p.58) simply emphasize the importance of the various model elements, but do not indicate how their effects are produced.

System Characteristics

The emphasis on behavioral variables ignores that fact that in all probability physical system characteristics play a fundamental role in SA effectiveness, perhaps as important a role as do behavioral variables. I have in mind such characteristics as the speed of system action (fast/slow) and system complexity (many/few components and interactive processes). The first imposes time constraints on the human reacting to an emergency, which vastly increases the requirements imposed on his/her SA; the second imposes resource demands on the operator in terms of the scope of the mental model s/he must develop. Manifestly it is much easier to develop an adequate mental model when only a few components and interactions must be conceptualized. Note that as one approaches the extreme limits of resource demands upon such entities as memory, for example, some sort of reduction of those demands will be required if the operator is to perform even less than optimally. This may require a change in the characteristics of the primary equipment and/or the provision of some sort of support system.

The nature of the system (static/dynamic; determinate/indeterminate, see Meister, 1991, for definitions) must also be considered. Although the studies have not been performed, I suspect that SA is profoundly influenced by the dynamics of the system; one therefore need not be much concerned about systems that are relatively static and determinate, like a farm, which demands relatively little SA except when weather conditions become severe.

Other hypotheses stem from the nature of the system. For example, the more hierarchically organized a system is, the more likely it is that certain processes will be less visible to system personnel, which in turn increases the difficulty of its personnel visualizing them and creating an adequate mental model.

Another potentially critical system factor is whether the system is or is likely to be in an adversary situation, in which case variables that are less predictable must be included in the operator's mental model. All systems, not merely military ones, may be in a adversary situation, when they face competition, whether that competition involves enemy fighters or playing the stock market. The adversarial situation requires prediction of future states (Endsley's level 3).

Another factor that has usually been ignored in SA theorizing is the state of uncertainty that surrounds the system. It has been suggested (Meister, 1991) that different types of uncertainty influence the decision making process (i.e., stimulus ambiguity; the difficulty of determining the likelihood of future events; lack of information about cause/effect relationships; difficulty in predicting decision outcomes). Endsley (1995a) has suggested that SA be differentiated from DM, that DM is an effect of SA, but in real world situations it is often difficult to separate the two, because events that involve SA and DM often transition to each other so rapidly, that the two cannot be distinguished. From an experimental standpoint it is important to distinguish between model elements that can be manipulated as independent variables and those that are dependent. System size, for example, is something that can be manipulated, whereas a mental model is a consequence of other processes and cannot easily be directly controlled. What makes the measurement situation more difficult is that this mental model, while a dependent variable, also has significant effects on the independent variables.

For all these reasons it would be highly desirable for SA theorists and researchers to develop a taxonomy of variables that influence SA and show how each of these can be treated as independent and dependant variables in an experiment.

SA as a General Construct

The implication of what I have been saying is that it is necessary to consider whether the SA phenomenon is one found only in aerospace or is a concept that crosses domain boundaries. It

seems to me that that decision has already been made, because SA has been investigated and found to be applicable in several areas, including anesthesiology and nuclear power. If SA were an aerospace phenomenon only, it would be easier to treat, because the number of system variations in aviation characteristics are relatively few. If, however, one throws the SA concept open to the behavioral world, as it were, then system characteristics and variables become more and more varied and important. Because of this, the performance of SA research in different domains may lead to slightly different results, because of individual differences in the various systems.

I myself am convinced that SA is a general phenomenon, because any situation that poses a problem in task performance requires and is involved with SA. The significance of SA (i.e., how important it is in affecting performance) may depend on the type of system in which it occurs. Because of the generality of the phenomenon, it should be explored as a general system factor rather than as an effect of individual domains like aerospace or automobiles. If SA is investigated as a system factor, it would be desirable to build a taxonomy that adequately describes the system conditions under which SA functions. Adding to the complexity and definitional vagueness of the SA concept is the possibility that SA in different systems may vary somewhat because of system individual differences. This possibility does not negate the validity of a general SA concept; it makes it necessary, however, to view SA in terms of its system context.

It is an open question whether SA researchers would or would not prefer to think of SA as a general phenomenon, because, while this generality adds to the importance of the concept, it increases the difficulty of dealing with it.

If SA is a general phenomenon, it might be desirable to approach its measurement on a general basis, by designing a prototypical, synthetic system as an experimental vehicle, a system whose characteristics could be varied along a number of dimensions, such as those earlier noted. Rouse and his colleagues (Henneman and Rouse, 1987) have done this with some success.

Moreover, if SA is considered as a general construct, linked to perception, information processing, and DM, for example, then at some point in SA research it will be necessary for SA investigators to indicate how the SA phenomenon fits in with other behavioral elements. It may then be possible to re–interpret research results found in other areas in terms of the SA concept and to add their databases (or at least parts of it) to the SA database. There is obviously some sort of interaction between concepts such as uncertainty, indeterminism and ambiguity, which stem from thinking about DM, and SA issues such as cue clarity and stimulus interpretation.

Under the circumstances it is easy to fall into a conceptual morass, as Flach (1995) has pointed out. Conceptual complexities there are in abundance, but I am less concerned about these than the ability to partition variables into those that are independent and manipulable and those that are dependent (and not easily controllable). It may be simplistic to bring up the S–O–R paradigm again, but if one can apportion variables and effects to the input/output elements of the paradigm, we are not likely to go seriously wrong in SA research.

How does one deal with so extensive a chunk of human performance? Does one have to control all SA model elements in arranging SA research, or can one extract only certain critical elements? The very comprehensiveness of the SA model, while otherwise highly desirable from a theoretical standpoint also makes it unwieldy.

SA and Design

Beyond SA theorizing and research, the engineering ergonomist faces the problem of translating their behavioral conclusions, implications and speculations into concrete design guidelines for hardware and software. Too many human factors people forget that this is the primary purpose of Human Factors/Ergonomics (HF/E). Unfortunately, HF/E specialists have done a miserable job of this in general, and one may ask whether it is likely that they will do better for SA. Theory and research provide a context for design guidance, but do not usually provide that guidance. The

theoretical hypotheses suggested by SA researchers are good starting places, but are rather general, although they may be refined in time.

SA and Research

I have much more confidence in the applicability of SA measurement methodology, e.g., SAGAT. It is in the nature of HF/E that we do much better in measuring human performance than in the business of translating measurement results into design guidelines. The variables that are studied in SA research do, however, present measurement problems, in part because many of those variables studied, e.g., memory, attention, symptom interpretation and DM, are covert and hence not easily measured.

If the theorizing and the research tell us no more than that attention, perception, DM, etc. are all involved in SA, they do not tell us much, because all of these *must* be involved, if SA exists. Like most behavioral phenomena, SA is circular; it is produced by certain variables, and once produced, affects these and other variables.

If SA is a given, the purpose of SA research should not be to demonstrate that it exists in a particular context but to explain, predict and control the phenomenon. Explanation requires the development of lawful relationships between inputs and outputs; prediction involves knowing when these relationships will occur; and control of SA requires the development of physical systems that aid the operator to overcome his/her SA deficiencies or that replace the operator who falls below a certain level of SA effectiveness. The fact that control is one of the purposes of SA research suggests that the research should include the development and testing of aiding systems.

Perhaps one might do better with more objective variables, those inherent in the nature of the system and the task being performed. One would wish to know, for example, about time constraints imposed by the system, i.e., fast reacting systems like aircraft and relatively slow ones like nuclear power facilities. One would expect that SA in time constrained situations is different than that in slow reacting systems, and that there are certain times in each when the human gets into trouble.

SA Aiding

Emphasis in SA research should perhaps be on extreme situations contrasted with routine ones, situations in which the operator is pressed to his/her SA limits. It is in such extreme situations that the human fails; and the whole point of the research is to find ways of preventing or mitigating such failures. It is even possible that extensive research is not needed to find ways of aiding the human. For example, if memory appears to be a problem, provide memory aids; if the human has difficulty in integrating multiple, dynamic stimuli, provide an expert computer system that can aid in the integration, either by providing suggestions to the operator or by replacing him.

The use of aiding support in highly complex, automated systems is not novel. The nuclear power industry has provided such aiding devices for its control room operators and the Halden, Norway, OECD Project laboratory has a continuing program of developing and testing such devices for nuclear power, using human performance measurement as a means of evaluating the adequacy of proposed solutions. Perhaps one could combine the design of such devices with research to determine how effective they are, in the process throwing light on the variables that affect the operator. Perhaps, one could do the same for SA. If one knew which design factors prevented SA failures, one might be able to work backward to an understanding of how SA variables function.

This process is of course very untraditional and it may also be more expensive, since aiding equipment would have to be designed and fabricated before it could be used as a research instrument. Such a process would also require behavioral researchers to work closely with engineers, something which is usually not done outside an engineering development facility.

With regard to SA research, I do not like overall objective performance measures like task completion and success rate, because one cannot tie these performance measures directly to variations in SA. Like workload, which it resembles, SA has both objective and subjective aspects and therefore it must make use of subjective measures. What makes it so dicey to measure SA is the same methodological problem encountered in workload: how does the subject know s/he has a certain amount of SA? Like workload, SA is also a hypothetical construct and the problem is to find input/output variables and relationships that reflect the construct. The trouble with any measure of a hypothetical construct, is that one must infer it from its measure, and one can never be sure that the measure is actually measuring what one hypothesizes. From a practical standpoint, I must confess to a preference for Endsley's SAGAT, but my philosophy is, the more measures the better, although if one finds measures that conflict, how does one explain the discrepancies?

Conclusion

To summarize, then: (1) SA is a hypothetical construct which is useful in explaining human performance in a variety of systems. It is also a phenomenon which reflects and affects human performance. (2) SA theory is quite elegant, but so far its research has not led to the determination of methods of aiding SA. (3) Since SA research which does not lead to concrete improvements in SA is useless, it would be highly desirable to integrate design and testing of SA–aiding systems with more conventional SA research. (4) SA theory and research have so far ignored the physical system, which probably has as much influence on SA as do behavioral variables. (5) It would be highly desirable to create a detailed taxonomy of SA variables which would include system as well as behavioral elements. (6) Overall performance and physiological measures are not likely to be as useful in SA research as subjective and SA–linked performance measures.

References

Adams, M.J., Tenney, Y.J., and Pew, R.W. (1995) Situation awareness and the cognitive management of complex systems. *Human Factors*, *37*, 85–104.

Endsley, M. R. (1995a) Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*, 32–64.

Endsley, M.R. (1995b) Measurement of situation awareness in dynamic systems. *Human Factors*, *37*, 65–84.

Flach, J.M. (1995) Situation awareness: Proceed with caution. *Human Factors*, *37*, 149–157.

Gaba, D.M., Howard, S.K., and Small, S.D. (1995) Situation awareness in anesthesiology. *Human Factors*, *37*, 20–31.

Henneman, R.L. and Rouse, W.B. (1987) Human problem solving in dynamic environments: Understanding and supporting operators in large scale, complex systems. ARI Research Note 87–51, Georgia Institute of Technology, Atlanta, GA.

Meister, D. (1991) *The Psychology of System Design*. Amsterdam, The Netherlands: Elsevier.

Sarter, N.B. and 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*, 5–19.

Smith, K. and Hancock, P.A. (1995) Situation awareness is adaptive, externally directed consciousness. *Human Factors*, *37*, 137–148.

Situation Awareness: A Cognitive Neuroscience Model Based on Specific Neurobehavioral Mechanisms

Robert S. Kennedy and J. Mark Ordy

Essex Corporation

Abstract

The ubiquitous concept of Situation Awareness (SA) has emerged in human factors in response to rapid developments in advanced technology. Specifically, as automation has been introduced into complex systems which are operated in diverse dynamic environments, the operator has been distanced from personal contact with the system. Although SA obviously involves complex theoretical and empirical issues, intial approaches have included such simple questions as: (a) to what extent is SA in the person or the situation?, (b) can SA be measured objectively?, and (c) does SA represent genetic or acquired traits? From empirical and practical standpoints, these SA questions need answers and they have focused research on such issues as measurement, selection, and training and the field continues to be in the process of clarification (Endsley, 1994; Flach, 1994; Pew, 1994). This review provides a possible extension of the concept of SA into cognitive neuroscience.

Current status of the concept of Situation Awareness in Human Factors Psychology

While the term SA is certainly intuitively appealing and has captured the imagination, human factors psychologists (Pew, 1994) are becoming increasingly concerned with what it is. It is used as the explanation for prediction, and control of human performance in complex environments involving rapid acquisition of multi sensory information, use of flexible working and long term memory for directing, attention, recognition, categorization, and rapid decision-making in dynamic settings (Endsley, 1994). Within an information processing conceptual framework, SA was originally proposed to encompass such critical components as multisensory integration, perception, interactions between working and long term memory, and rapid decision-making in dynamic systems (Endsley, 1988; Hartman & Secrest, 1991). However, there is much confusion on the use of the term and it means different things to different people. In addition to the cognitive elements, perceptual issues are included by some (Ercoline, 1994), but not others (Rimson, 1994), although visual-spatial perceptual skills in pilots, for example, are crucial for mission success and safety (Kosslyn, Flynn, Amsterdam, & Wang, 1990).

Although interest in SA among human factors scientists is expanding rapidly, SA remains an enigmatic theoretical concept for many reasons, such as attempting to use SA for selection

(Endsley & Bolstad, 1994) - which seeks to identify stable traits and at the same time discussing SA also for training which depends on the modifiability of skills (Kass, Herschler, & Companion, 1990). Also it is difficult to bound a construct when its level of focus has been claimed to go "beyond traditional information processing" (Endsley, 1994). However, that does not mean SA is not important in diverse contexts which confront human factors psychologists, engineers, and decision makers (Endsley, 1994). Folklore has it that SA is a term given to human factors by pilots (Nordwall, 1993) and because it has such wide acceptance we are probably going to continue to use it. SA nearly always entails explicit/implicit control over rapid sensory-motor integration involving spatial relationships. More recently, the construct of SA is having important implications for how theoretical and empirical research efforts are focused in order to explain, predict, and control human performance in diverse complex spatial environments, and to provide guidance on how such rapid decision-making behavior can be selected and/or improved. Individual differences in some proposed SA functions have suggested modification by selection for, and training in, SA capacity. The concept of SA has been expanded to include skillfull, adaptive, and conscious problem-solving behavior in general psychology (Smith and Hancock, 1994). In a more restricted use, the construct of SA has been used to emphasize the role of perception in rapid problem solving and decision-making in complex systems involving intricate and dynamic "coupling of perception, cognition, memory, and conscious action" (Flach, 1995). The construct of SA is most widely used in military and commercial aviation problems of skilled mission performance and safety (Endsley, 1994). However, the SA construct is also finding its way into the literature in other human factors problems and complex system designs (Adams, Tenney, & Pew, 1995). Examples include operation of aircraft, air traffic control, flexible manufacturing systems (FMS), operation of tactical and strategic test devices, driving, and many other activities that require a dynamic information update to function rapidly and effectively (Endsley, 1994).

In order to specify the theoretical and empirical utility of the concept of SA in dynamic human decision making, plausible starting points include: (1) a general definition of SA, (2) an enumeration of the alleged important role that SA may play in diverse dynamic decision making environments, (3) the heuristic role of SA in the current explicit/implicit memory dichotomy in cognitive neuroscience that may be involved in rapid sensory-motor integration in dynamic environments, and (4) exploration of whether a cognitive neuroscience approach for SA research may be useful for studying individual differences in SA, and thus enhance opportunities in selection and training.

General definition of SA in Human Factors Psychology

The following general definition of SA has been proposed: "Situation awareness is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future." To expand on this definition, the SA construct has been described in three hierarchical phases: "(a) Level 1 SA - perception of the elements in the environment; (b) Level 2 SA - comprehension of the current situation; and (c) Level 3 SA - projection of future status" (Endsley, 1994).

Basically, it has been proposed that SA is based on short term sensory memory, perception, and interaction of working memory with long term memory. In addition to extemal factors, attention and perception can be directed by the contents of both working and long-term memory in this SA model. The perception of the environment, the first level of SA is largely guided by the contents of working and long term memory stores that focus attention, recognition, and categorization of information. Once perceived, information is stored in working memory, a limited capacity system for holding and manipulating information. Active processing of SA information in the dynamic decision making environment occurs in working memory. Working memory can be activated by information from either the environment or from long term memory. Endsley (1994) maintains that working memory may constitute the main "bottleneck for situation awareness". In developing expertise in SA, a form of "automaticity" may be involved (Shiffrin & Schneider,

1977). Automatic processing tends to be fast, autonomous, effortless, and unavailable to conscious awareness in that it can occur without attention (Endsley, 1994). As discussed later, knowledge and/or memory which can influence awareness in a situation can be acquired "autonomously" or "implicitly". This knowledge so obtained may be less flexible than information in SA acquired by conscious, episodic, or explicit learning and memory (Schacter, 1994; Ungerleider, 1995), but it happens this way often enough in real world activities that any useful model of SA must be able to incorporate such explicit/implicit memory conditions.

SA in Cognitive Psychology

Despite the fact that the concept of SA has captured the enthusiasm of human factors psychologists, so far as we can tell the concept has not been associated with or incorporated into the currently most popular neurocognitive theoretical framework of memory functions. For example, it would seem appropriate to adapt into SA models these elements of modern neuroscience of memory: (1) perceptual priming, (2) explicit/implicit learning and memory, or (3) perceptual and memory functions with and without conscious awareness (Squire, 1992, 1994). In such a theoretical framework of memory, the concept of SA can be formulated in two distinct ways:

According to the dominant current view in human factors, SA is not a "passive process", or presumably implicit process (Endsley, 1995), and the skills required for developing and maintaining superior SA capacity need to be identified as conscious perceptual spatial cues, and as explicit, or declarative memory, with conscious decision making that can be improved in specific and explicit SA training programs (Squire, 1992, 1994). We would agree that conscious awareness of a complex, dynamic spatial situation requiring a rapid decision making response can be developed or expressed by conscious recollection of spatial memory that is specified by a specific set of cues, time, and context. This represents declarative, conscious, or explicit memory or knowledge (Squire, 1992, 1994). However, unconscious awareness follows from the principle of implicit perceptual learning, and is involved in picture recognition, constancy of scaling, and has been illustrated in fMRI studies of explicit/implicit learning (Pascual-Leone, Grafman, & Hallett, 1994).

Implicit nonconscious spatial perception, learning, or memory in SA may be developed or expressed by behavior that demonstrates that previous exposure to a spatial task has resulted in improved SA and rapid decision-making psychomotor performance in dynamic systems on that specific task "without the subject recalling consciously being exposed to the specific task before" (Endsley, 1994; Schacter, 1994). Also, unlike explicit measures of situation awareness that require conscious recognition and recall from spatial working memory, implicit measures of spatial memory in SA may be inferred from task performance or from neural measures such as bioelectric events or from fMRI imaging studies. These operational measures of implicit SA are less subjective and can also provide assignment of more quantitative values to the functional sensory, cognitive, and motor content of SA (Roedinger, 1990). However, admittedly, they are not easy to obtain in field studies, although this may be changing. Perhaps the most intensely studied form of implicit memory improvement in SA may be the phenomenon of repetition or direct "priming" effects, or the facilitated rapid identification of cues in complex spatial environments as a consequence of prior exposure (Schacter, 1994). It has been proposed that "priming" may reflect operations of an implicit perceptual, representational, spatial memory system that can function even independently of the explicit, episodic, or declarative spatial memory system that has been proposed to operate in SA (Endsley, 1994; Schacter, 1994).

Many researchers in cognitive psychology have been interested in whether implicit and explicit spatial representations in learning and memory are independent, whether they interact, whether there is a reciprocal conversion from explicit to implicit stages. Also, can implicit knowledge be redescribed into explicit or conscious and explicit episodic memory and knowledge (Seger, 1994). As yet, the heuristic value of the explict/implicit memory taxonomy in cognitive psychology has not been investigated in such dynamic approaches as are used in SA research. In contrast, recent

cognitive neuroscience studies have shown that different forms of spatial memory may be associated with different mechanisms of cortical neuronal circuit plasticity and that there is a reciprocal transfer of knowledge from an explicit to an implicit state in human subjects (Pasqual-Leone, Grafman, & Hallett, 1994). This is exactly the kind of relationship that would be desirable in some forms of SA training. The reverse may also occur when subjects initially learn motor tasks automatically and "unconsciously" and later switch to a conscious and explicit and more flexible mode of performance (Ungerleider, 1995). These cognitive neuroscience studies imply that neuronal plasticity is a common feature of the human cortex and that this neural plasticity may be closely linked to the perceptual, cognitive, and psychomotor performance ability to respond flexibly and promptly in SA decision making to spatial cues in dynamic systems. We believe these findings and those like them (e.g., Squire, 1992, 1994; Ungerleider, 1995) should be incorporated into current SA models and conceptualization.

Possible Heuristic and Broader Role of SA in Cognitive Neuroscience: Correlation of Some Specific SA Functions with Neural Mechanisms and Neuronal Plasticity.
Although currently not dealt with in cognitive neuroscience, or with explicit reference to the role of the brain in neurobehavioral studies, the concept of SA has been "loosely defined" as an internal representation, or cognitive map or model of the environment localized in the brain (Endlsey, 1994). Human Factors psychology has placed the concept of SA within an information processing framework, with sensory input, cognitive function, and motor performance as a series of separate and discrete conscious awareness stages (Endsley, 1994; Flach, 1995). Within this cognitive theoretical framework, SA has been placed after attention and cognitive processing, but before decision making and psychomotor performance (Endsley, 1994). There has also been increasing focus in cognitive psychology on intricate coupling among perception, cognitive function, decision, and actions, with the concept of the "perception-action cycle" playing a prominent role in different theoretical SA models (Adams et al., 1995). This perceptual action cycle approach in SA has resulted in placing "sequential boxes" in the information-processing cognitive model but without providing explicit criteria of differentiating the specific sensory, cognitive, and motor functions of the separate boxes in SA (Flach, 1994). Working memory has also been conceptualized as interface between memory and cognition (Baddeley, 1994). In contrast to this schematic, theoretical and purely descriptive and introspective framework for SA in cognitive psychology, the remarkable technological progress in functional brain imaging techniques (fMRI) in cognitive neuroscience has made it possible in future studies to correlate some specific explicit/implicit SA functions with specific neural mechanisms and neuronal plasticity. Specific correlations have been identified in: (1) perceptual priming tasks in implicit memory (Schacter, 1994), and (2) in the conscious modulation of human cortical sensory-motor coupling "maps" during development and expression of dual explicit and implicit learning, memory, and performance tasks (Pasqual-Leone et al., 1994). Human fMRI studies have shown that learning and memory may involve many of the same corticaly regions that process sensory information and also control motor output (Ungerleider, 1995).

Because SA assessment has been regarded as the evaluation of spatial information immediately available in conscious awareness (Endsley, 1988, 1994), explicit or conscious declarative measures of spatial memory have been considered to be the most direct way of SA assessment (Endlsey, 1994). It has been proposed that "Ideally, it would be desirable to install a window on the operator's mind and observe an exact picture of what is known at all times" (Endsley, 1993). In earlier views, it was concluded that known physiological methods, EEG, P300, etc., while providing some objective and useful data on SA localization in the brain, did not appear promising for specific SA measurement (Endsley, 1993). We would agree with this characterization since to be useful these methods need to provide greater temporal and spatial resolution of information processing in the brain. However, recent PET, and even more recent functional brain imaging techniques (fMRI) or "Images of the Mind" (Posner & Raichle, 1994) have fundamentally changed the technical, non-invasive cognitive neuroscience studies. Consequently, a cognitive neuroscience focus on SA may provide unique opportunities for clarifying this theoretical concept as well as for improving practical prospects of evaluating individual differences in SA capacity by

selection and/or training. For example, in SA involving working memory, it has been proposed that working memory can be activated by information from either the extemal environment or from long term memory storage. It has also been proposed that flexible working memory may represent the critical link for SA (Endsley, 1994). It is now widely recognized in neuroscience that memory is not unitary and can be classified as explicit or implicit on the bases of how and where information is stored in the brain, and how it is recalled (Squire, 1992, 1994). Explicit memory involves the temporal lobe and the hippocampus, whereas implicit memory does not require conscious recall, and involves primarily reflex pathways, as well as the amygdala and cerebellum (Squire, 1992, 1994; Ungerleider, 1995). If working memory does represent the critical link in SA (Endsley, 1994), it seems apparent that the neural basis of SA includes relations between explicit and implicit learning and memory, and the recently fMRI localized neural interactions between the cerebral cortex and hippocampus (Squire, 1992, 1994; Ungerleider, 1995) (see Figure 1).

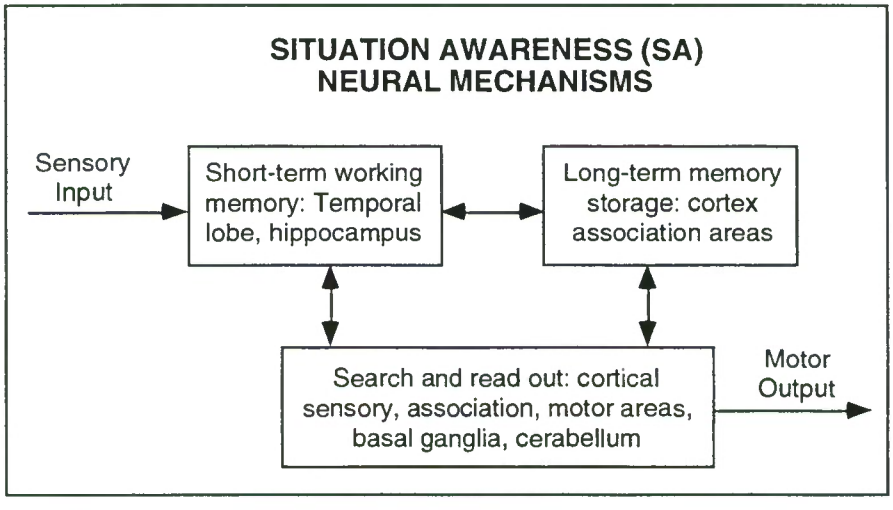


Figure 1. Schematic localization of SA neural mechanisms in terms of working - long term memory interactions associated with cortical, hippocampal, and cerebellar functions.

Using functional brain imaging (fMRI) and specific psychophysical tests, remarkable progress has been made in discovering neuronal circuit plasticity that is involved in perception, attention, cognition, and complex motor performance (Ungerleider, 1995). There is considerable evidence that SA may be closely linked to neuronal plasticity in perceptual performance, in both explicit and implicit learning and memory, and in the skilled motor ability to respond flexibly to rapid environmental changes. Human lesion, and experimental animal studies have shown that specific and different neural pathways and regions are critical for explicit and implicit forms of perception, learning, and memory (Squire, 1992, 1994; Ungerleider, 1995). Human fMRI studies have shown that conscious, explicit, declarative memory is more flexible than implicit non-conscious memory (Ungerleider, 1995). Specifically, in fMRI tests, visual targets for recognition first activate visual areas (V1), then while the target stimulus is held in "mind" or short-term working memory in the temporal lobe and hippocampus, the prefrontal cortex becomes activated because feedback projections are necessary for reactivating the stimulus, trace, or target representation of

the visual cortical association areas, possibly from long-term memory storage (Squire, 1992, 1994; Ungerleider, 1995). Also, the prefrontal cortex and cerebellum may be selectively engaged in visual search tasks in which the initial perceptual learning is explicit and conscious, with the motor cortex becoming more dominant as the task becomes implicit, non-conscious, or "autonomous" (Pasqual-Leone et al., 1994; Ungerleider, 1995). The association cortices are involved in many of the higher cognitive integrative functions that may be involved in neural plasticity and in SA (Kandel et al., 1991). See Figure 2 for a lateral view of the human brain showing sensory, association, motor cortices which can be activated differentially in fMRI studies.

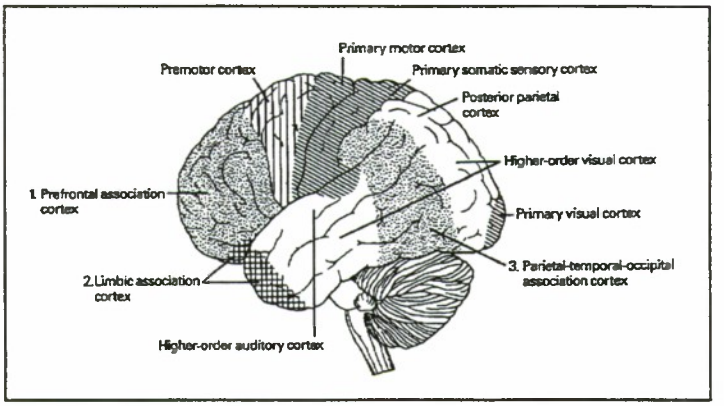


Figure 2. Lateral view of human brain showing primary sensory, motor cortices, higher-order motor and sensory cortices, and three association cortices which can be activated differentially in cognitive fMRI studies (adapted with permission from Figure 53-2, p. 825, Kandel et al, 1991).

From a cognitive neuroscience perspective of SA, it is now possible to visualize the brain, not only as a "black box" with sequential schematic boxes for sensory, cognitive, and motor functions (see Figure 1), but also as an intricately organized neural network system in which different pathways and regions interact to perform complex, and yet flexible and rapid information processing (see Figure 2).

In summary, although earlier neurophysiological methods were not promising for SA measurement and research because of their limited spatial and temporal resolution of information processing in specific regions of the brain, the recent developments in fMRI brain imaging have fundamentally changed cognitive neuroscience research (Ungerleider, 1995) and we belive it is time to begin introducing these concepts and methods into modern SA research and development.

References

Adams, M. J., Tenney, Y. J., & Pew, R. W. (1995). Situation awareness and the cognitive management of complex systems. *Human Factors, 37*(1), 85-104.

Baddeley, A. (1994). Working memory: The interface between memory and cognition. In D.L. Schacter & E. Tulvin (Eds.), *Memory Systems*. Boston, MA: MIT Press.

Endsley, M. R. (1988). *Situation awareness global assessment technique (SAGAT).* Paper presented at the National Aerospace and Electronic Conference (NAECON), Dayton, OH.
Endsley, M. R. (1993). Situation awareness and workload: Flip sides of the same coin. *Proceedings of the 7th International Symposium of Aviation Psychology.*
Endsley (1994). Situation Awareness in dynamic human decision making: Theory. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 27-58). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.
Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32-64.
Endsley, M. R. & Bolstad, C. A. (1994). Individual differences in pilot situation awareness. *The International Journal of Aviation Psychology, 4*(3), 241-264.
Ercoline, W. R. (1994). Spatial disorientation type II [Letter to the editor]. *SAFE Journal, 24*(3), p. 58.
Flach, J. M. (1994). Situation awareness: The emperor's new clothes. *Proceedings of the 1st Conference on Human Performance in Complex Automated Systems* (pp. 241-248). Hillsdale, NJ: Erlbaum.
Flach, J. M. (1995). Situation Awareness: Proceed with caution. *Human Factors, 37*(1), 149-157.
Hartman, B. O. & Secrest, G. E. (1991). Situational awareness is more than exceptional vision. *Aviation, Space, and Environmental Medicine, 62*, 1084-1089.
Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (1991). *Principles of Neural Science* (3rd Ed.). E. Norwalk, CT: Appleton & Lange.
Kass, S. J., Herschler, D. A., & Companion, M. A. (1990). Are they shooting at me?: An approach to training situational awareness. *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 1352-1356). Orlando, FL: Human Factors.
Kosslyn, S. M. , Flynn, R. A., Amsterdam, J. B., & Wang, G. (1990). Components of high-level vision: A cognitive neuroscience analysis and accounts of neurological syndromes. *Cognition, 34*, 203-277.
Nordwall, B. D. (1993). EW goal is improved situation awareness. *Aviation Week & Space Technology, 139*(1), p.59.
Pascual-Leone, A., Grafman, J., & Hallett, M. (1994, March 4). Modulation of cortical motor output maps during devlopment of implicit and explicit knowledge. *Science, 263*, 1287-1289.
Pew, R. A. (1994). In introduction to the concept of Situation Awareness. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 17-24). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.
Posner, M. I. & Raichle, M. E. (1994). *Images of the Mind.* New York: Scientific American Laboratory.
Rimson, I. J. (1994). Loss of situational awareness [Letter to the editor]. *SAFE Journal, 24*(3), p. 57.
Roedinger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45*, 1043-1056.
Schacter, D. L. (1994). Priming and multiple memory systems: Perceptual mechanisms of implicit memory. In D.L. Schacter & E. Tulvin (Eds.), *Memory Systems.* Boston, MA: MIT Press.
Seger, C. A. (1994). Implicit learning. *Psychological Bulletin, 115*(2), 163-196.
Shiffrin, R. M. & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and general theory. *Psychological Review, 84*, 127-190.
Smith, K. & Hancock, P. A. (1994). Situation Awareness is adaptive, externally-directed consciousness. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 59-68). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.
Squire L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 37*(1), 137-148.

Squire, L. R. (1994). Declaritive and nondeclaritive memory: Multiple brain systems supporting learning and memory. In D.L. Schacter & E. Tulvin (Eds.), *Memory Systems*. Boston, MA: MIT Press.

Ungerleider, L. G. (1995). Functional brain imaging studies of cortical mechanisms for memory. *Science, 270.*

The Tradeoff of Design for Routine and Unexpected Performance: Implications of Situation Awareness

Christopher D. Wickens

University of Illinois at Urbana-Champaign

Abstract

We identify tradeoffs in three domains--hazard awareness, system awareness, and task awarenessbetween design for routine operations and design for unexpected circumstances, which occur rarely, but may be more likely to have catastrophic consequences if the operator is not aware of the broader state of the environment when the unexpected occurs. We demonstrate that design for the routine is rarely optimal for the unexpected, describing in detail this tradeoff for hazard awareness. Finally we discuss the implications of this tradeoff for performance measurement in test and evaluation, and for display design for hazard awareness, contrasting four design philosophies.

Introduction

While situation awareness remains a somewhat fuzzy concept, recent efforts to define it by investigators such as Endsley (1995), Dominguez (1994), and Adams, Tenney, and Pew (1995) have allowed some degree of consensus to emerge. I have chosen to use the following, which is closely related to Endsley's definition.

> Situation awareness is the continuous extraction of information about a dynamic system or environment, the integration of this information with previously acquired knowledge to form a coherent mental picture, and the use of that picture in directing further perception of, anticipation of, and attention to future events.

The utility of any such definition however requires that the user carefully define the properties of the environment within which the "situation" in question evolves. In the aerospace community, three such environments are prominent, each with different implications for objective measurement:

- The 3D geographical space around the aircraft, occupied by hazards, such as other air traffic (friend and foe), weather and terrain (Wickens, 1995a,b),
- Internal systems within the aircraft, involving in particular, automation systems (Sarter and Woods, 1995),
- Responsibility for the array of tasks confronting the pilot, his or her crew, and various automated agents (Funk,1991).

For each of these environmental domains, the specific metrics of "the situation" will be expressed in different qualitative languages, and hence, the objective measures will be superficially quite different between them. Geometry plays a key role in (1), Boolean logic plays an important role in (2), and checklists and queues are important in (3). Yet in all cases, there are certain common themes underlying the preservation of situation awareness....themes that impact both its objective measurement, and design decisions to support its maintenance. The important theme we emphasize in this paper relates to the tradeoffs that exist between design for the routine and design for situation awareness and the relation of this tradeoff to expectancy. In the following, I will illustrate the specific nature of this tradeoff with regard to hazard awareness, then illustrate its parallels in systems and task awareness, and finally discuss some of the implications of the tradeoff for both design and performance evaluation.

Design Tradeoffs

Hazard Awareness

In our laboratory we have completed a series of studies examining the optimal format for design of piloting navigation and hazard awareness displays. In this research we have characterized the pilot as confronting two generic types of navigational tasks: local guidance and global awareness.

Local guidance is the process of maintaining precision along the flight path, and characterizes routine flight that occupies perhaps 95-99% of a pilot's flight time. The information needs for local guidance are of depictions of deviations off of the flight path; that is, information that is ego-referenced, presenting a view directly ahead of the aircraft; it is forward looking (i.e., three-dimensional), and relatively close in (i.e., the flight path within a few thousand feet ahead of the aircraft. A distance defined, in part, by the time constant of the aircraft and its speed). Hence, as we have found, the ideal display for local guidance is a forward looking ego-referenced "highway (or tunnel) in the sky" (Wickens and Prevett, 1995; Haskell and Wickens, 1993; Theunissen, 1994). Such a view is represented schematically in Figure 1a.

Yet on relatively infrequent occasions, the pilot is called upon to utilize far more global hazard awareness within a much greater volume of space, 360 degrees around the aircraft. Here the highly ego-referenced view of the tunnel in the sky is ill-suited. It either provides a very narrow "keyhole" view of the world, or, if the geometric field of view is expanded (Figure 1b; Barfield et al., 1995; Wickens and Prevett, 1995), it provides a highly distorted picture of where things (hazards) are. Instead, ideal displays for hazard awareness tend to be those that are more two-dimensional (look down), zoom out, and world-referenced. In a large number of experiments, summarized in Wickens (1995a,b), we have utilized several different measures of hazard awareness (see Olmos, Liang, and Wickens, 1995), to establish the advantage of more exocentric displays for situation awareness (Figures 1c and 1d), and therefore to identify the tradeoff of display support for the two different kinds of tasks.

Hence, the dilemma for designers: does one design for the routine, which occupies most of a pilot's time, or for the unusual, which, by definition occurs only rarely, but has potentially dangerous consequences if it is not well supported. Before we address this question, we describe analogous forms of this tradeoff in systems and task awareness.

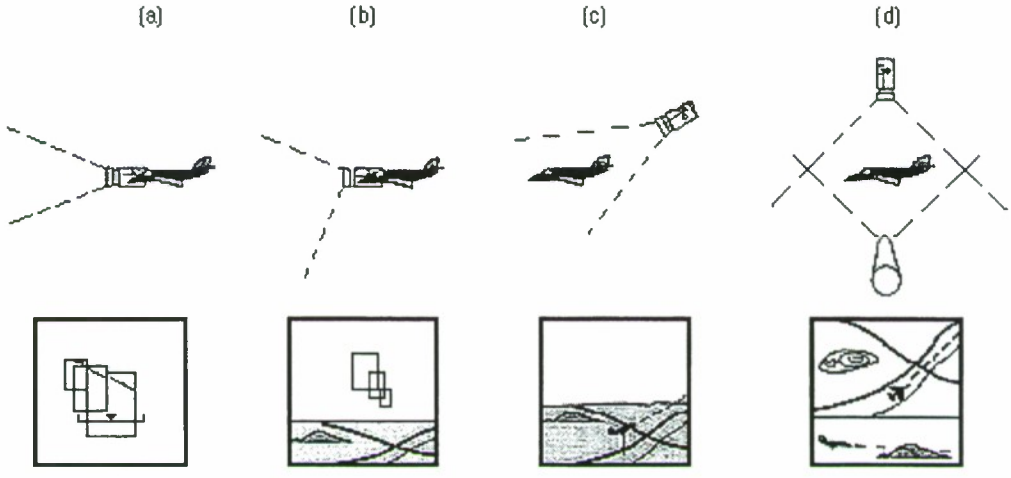


Figure 1

Systems Awareness

Emerging from early applications of cognitive engineering to the design of nuclear power control rooms (Landeweerd, 1979; Goodstein, 1981, Rasmussen, 1986), there has been a realization that system characteristics that are ideally suited for routine operation may be very poorly suited for dealing with system failures. For example, in routine operations, operators utilize skill- and rule-based behavior to assist them in causal reasoning and relatively convergent thinking. In contrast, under conditions of failure and fault management, they must use knowledge-based behavior to assist them in diagnostic reasoning, and relatively divergent thinking (Wickens, 1992). Displays and procedures designed to best support the former circumstances are not necessarily well-suited for the latter.

Applying similar logic to the design of flight deck automation, a series of investigations by Sarter and Woods (1991, 1994, 1995a,b) has revealed that the design of automated systems when things are working well (which is the situation most of the time) are poorly suited for the occasions when either the automated system fails, or it is asked to perform its duties under relatively improbable (but not impossible) circumstances. In the former case, effective design may merely require an economical display of which automation modes are in effect, and which ones are "armed" to be activated in the near future (this is analogous to the preview of the highway in the sky, for local guidance). In the latter case, there is a need for considerably more elaborate feedback regarding why the system is responding as it is; and such a need imposes a more elaborate display design challenge.

Finally, in both nuclear or process control, and flight deck management, even within the conditions of fault and failure themselves, there is evidence for a tradeoff. The procedures to be followed for relatively "routine" simple faults -- which can be addressed by following straightforward checklists -- may be quite inappropriate for the less expected, more complex multiple faults. In the latter circumstances, blindly following checklists may be a very "brittle" tactic (Roth and Woods, 1988), which can lead the fault manager down some dangerous paths, and may make a bad situation worse.

Task Awareness

A third, but less investigated example of the tradeoffs that exist may be found in task awareness. This describes the operator's awareness of what tasks have been done, need to be done, and, for the latter, in what priority (Adams, Tenney, and Pew, 1993). In either automated systems or multi-operator systems, task awareness includes not only this knowledge of task queuing, but also task responsibility: which agent is responsible for the performance of which task. Here again, however, there is a difference in design support for the routine, and for the unexpected. During the routine, checklists, whether paper or automated, provide excellent support for task awareness (Degani and Wiener, 1993). But during the unexpected or unusual operating circumstances, the checklist may no longer be as effective, since departures from pre-planned sequences, and unexpected shifts in responsibility may be required.

Implications Of Tradeoff For Test And Evaluation

We have outlined three important aerospace performance domains, in which design for optimal support for routine performance may not effectively support situation awareness, necessary to function effectively during unexpected circumstances and vice versa. This generic tradeoff poses a very real dilemma for those engaged in system (or operator) test and evaluation. On what criterion should the merits of a system be judged -- that which characterizes 90-95% of performance (the routine), or that very unexpected and small sample of time when the unexpected occurs? A logical argument could be voiced that the figure of merit of such a system should be weighted 95% on how well it supports the routine, and 5% on how it supports situation awareness in unexpected circumstances. The problem with such an argument however is that it is often during the unexpected that systems are most vulnerable to the sorts of catastrophic events that characterize incidents like the Three Mile Island nuclear power plant disaster (Reason, 1991), or the recent causes of commuter airline crashes in Indiana and North Carolina. That is, during the unanticipated events, the expected cost of inadequate system operation is far greater than it is during the anticipated and routine. While low frequency events (i.e., failures) *are* adequately sampled by test pilots in aircraft certification programs, these events are generally anticipated as part of the test flight plan, and hence cannot truly be described as "unexpected."

A second issue related to the ability of operators to handle the unexpected concerns *individual differences* in operator personnel. As we have noted, it may be argued that the system is most vulnerable in the unexpected circumstance. Furthermore, however, it is also likely that this system vulnerability will be amplified when the unexpected circumstance is encountered by operators on the low end of the distribution of skills and abilities of the pool of users of the system in question. I would argue that such behavior (low-skilled operators in unexpected circumstances for which they are ill-prepared) needs to be adequately sampled, and disproportionately weighted in system evaluation. It may contribute only a small proportion to the total system operation time; but contribute a disproportionately high expected cost of system failure. Such combination of low operator skills and inadequate training and preparation for unexpected events is certainly not incorporated in flight tests with generally highly qualified test pilots. The combination *would* appear to be sampled in traditional LOFT training. But the lessons learned from such training are rarely applied to system design, as we discuss below.

The Tradeoff In The Design Process

As we have noted elsewhere, the tradeoffs discussed above explicitly dictate that optimal designs may be very different for the routine and the unexpected. Addressing our concerns here more specifically to the depiction of the geographical environment (local guidance and hazard displays), we consider four alternative solutions to the tradeoffs that appear to exist.

The Compromise Display

As Figure 1 illustrates, there may be display schemes at the midpoint of egocentricity, between the ego-referenced display best suited for local guidance (1a), and the world-referenced display best suited for global awareness (1d); such a compromise display will "satisfice" in the sense of providing adequate support for both. Indeed our research has revealed that a rotating "tether" display concept, schematically illustrated in Figure 1c appears to achieve such a compromise (Wickens and Prevett, 1995).

The Dual Display

An alternative to the compromise is to design the optimal display for each task (guidance and awareness), and present the two simultaneously. Such a solution encounters three limitations. First, it is not an economical use of limited real estate in environments such as the flight deck. Second, it will impose added scanning demands on the operator. Third, there is often a need for the operator to mentally translate between information presented in one display and in the other. For example, the location of a dangerous hazard may be depicted on the situation awareness display, but the pilot may need to know where it is likely to be seen on the local guidance display. We have found that this "cognitive linkage" between separate display panels can, to some degree, be supported by adopting techniques of *visual momentum* (Woods, 1984). An example here is to physically depict the field of view of the local guidance display, in terms of a "wedge" depicted on the rendering of the global awareness display (Aretz, 1991; Liang, Olmos, and Wickens, 1995).

The Sequential Display

The real estate problem of the dual display can be solved by use of flexible or multifunction displays, in which different renderings, optimally suited for one task or another can be depicted at different times within the same physical viewport. Numerous examples of this strategy may be found with the various modes of depicting electronic maps (the horizontal situation display) in commercial glass cockpits, or of depicting radar coverage in combat aircraft. Whether such displays are user-chosen, or adaptively selected by automation (Moûloua and Parasuraman, 1994), they have an advantage of space economy. Yet there are three drawbacks to the implementation of a sequential strategy. First, if information on one display must be related to that on the other, working memory limitations sometimes impose on the ability of the user to carry over information from one to the other (Seidler and Wickens, 1992; 1995). Second, when the number of possible displays exceeds a handful, an added burden is imposed on manually operating whatever device choice or menu selection tool is required to "navigate" from one screen to another (Seidler and Wickens, 1992). Finally, such a situation will create a certain amount of inconsistency of representation. For example, the same physical space may, at one moment, represent a rotating map (stable aircraft), and at the next moment, represent a rotating aircraft (fixed map). Such inconsistency may be confusing, and at times even dangerous, if the operator temporarily forgets what mode he or she is viewing.

The Automated System

The final, and perhaps most radical solution to the design tradeoffs between guidance and hazard awareness, is to design primarily for the support of situation awareness (support for the unexpected), assuming that local guidance can be, will be, and often *is* easily automated. By definition, local guidance during routine operations is fairly predictable and is the sort of skill-based task that easily lends itself to effective and reliable automation. Hence, display support to the operator for such a routine task can be less than optimal. Such an approach then allows greater efforts to be focused on the design of effective displays for global awareness. Indeed, our research has indicated that the "tether display" shown schematically in Figure 1c, represents such a choice, serving global hazard awareness somewhat better than it serves human performance in local guidance. But if the guidance function were under autopilot control, such minor deficiencies in human performance would be less critical to full system performance.

However, if this approach (weight design for optimal support of awareness and assume automation of routine) is adopted, then the decision to assume that routine performance will often be automated should only be made while bearing two important considerations in mind. First, sufficient care should be given to providing the operator with some level of active choice in the navigational decisions of the system, even if the decisions are intermittent (Billings, 1996). Second, designers should NEVER proceed to assume that automation will not fail, despite any assurances that may be offered by manufacturers to the contrary. Painful lessons have revealed that such automation is rarely, if ever, entirely "failure free."

Conclusions

In conclusion, we have argued that a system-wide analysis of performance differences in routine and unexpected circumstances with low as well as high skilled operators should be undertaken, coupled with understanding of the constraints of display real estate and human cognition, before designs are implemented, and before test programs are carried out to evaluate the implemented designs. We feel that weighting for optimal design for (and measurement of) situation awareness should be shifted to a level that considerably exceeds the proportion of time that such awareness is actually required in order to handle the unexpected.

Acknowledgments

Much of the research summarized in the report was supported by a grant from the NASA Ames Res. Ctr., Moffett Field, CA (NASA NAG 2-308). Vernol Battiste and Sandra Hart were the technical monitors. Several students at the University of Illinois contributed to the thinking and data underlying this report including Bradley Boyer, Ian Haskell, Chia-Chin Liang, Oscar Olmos, Tyler Prevett, and Karen Seidler.

References

Adams, M.J., Tenney, Y.J., and Pew, R.W. (1991). *Strategic workload and the cognitive management of advanced multi-task systems* (State of the Art Report SOAR/CSERIAC 91-6). Crew System Ergonomics Information Analysis Center, Wright-Patterson AFB, OH.

Adams, M.J., Tenney, Y.J., and Pew, R.W. (1995). Situation awareness and the cognitive management of complex systems. *Human Factors, 37*(1), 66-85.

Aretz, A.J. (1991). The design of electronic map displays. *Human Factors, 33*(1), 85-101.

Barfield, W., Rosenberg, C., and Furness, T.A., III (1995). Situation awareness as a function of frame of reference, computer-graphics eyepoint elevation, and geometric field of view. *The International Journal of Aviation Psychology, 5*(3), 233-256.

Billings, C. (1996). *Toward a Human Centered Approach To Automation*. Englewood Cliffs, NJ: Lawrence Erlbaum Associates.

Degani, A., and Wiener, E.L. (1993). Cockpit checklists: Concepts, design, and use. *Human Factors, 35*(4), 330-345.

Dominguez, C. (1994). "Can SA be defined?", in M. Vidulich, C. Dominguez, E. Vogel, and G. McMillan, *Situation awareness: Papers and annotated bibliography (U)*," Interim Report AL/CF-TR-1994-0085, Armstrong Laboratory, Air Force Materiel Command, Wright-Patterson Air Force Base, Ohio.

Endsley, M.R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors, 37*(1), 65-84.

Funk, K. (1991). Cockpit task management: Preliminary definitions, normative theory, error taxonomy, and design recommendation. *The International Journal of Aviation Psychology, 1*(4), 271-286.

Goodstein, L.P. (1981). Discriminative display support for process operations. In J. Rasmussen & W.B. Rouse (Eds.), *Human Detection And Diagnosis Of System Failures*. New York: Plenum Press.

Haskell, I.D., and 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.

Landeweerd, J.A. (1979). Internal representation of a process fault diagnosis and fault correction. *Ergonomics, 22*, 1343-1351.

Liang, C.C., Wickens, C.D., and Olmos, O. (1995). Perspective electronic map evaluation in visual flight. In R. Jensen (Ed.), *Proceedings of the 8th International Symposium on Aviation Psychology*. Columbus, OH: Dept. of Aviation, Ohio State University.

Mouloua, M., and Parasuraman, R. (Eds.) (1994). *Human Performance In Automated Systems: Current Research And Trends*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Olmos, O., Liang, C.C., and Wickens, C.D. (1995). Construct validity of situation awareness measurements related to display design. *Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness*. Daytona Beach, FL, Oct.

Rasmussen, J. (1986). *Information Processing And Human-Machine Interaction: An Approach To Cognitive Engineering*. New York: North Holland.

Reason, J. (1990). *Human Error*. New York: Cambridge University Press.

Roth, E.M., and Woods, D.D. (1988). Aiding human performance I: Cognitive analysis. *Le Travail human, 51*, 39-64.

Sarter, N.B., and Woods, D.D. (1992). Pilot interaction with cockpit automation: Operational experiences with the flight management system. *The International Journal of Aviation Psychology, 2*, 303-322.

Sarter, N.B., and Woods, D.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.

Sarter, N.B., and 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.

Seidler, K.S., and Wickens, C.D. (1992). Distance and organization in multifunction displays. *Human Factors, 34*, 555-569.

Seidler, K.S., and Wickens, C.D. (1995). The effects of task and multifunction display characteristics on pilot viewport allocation strategy. *Proceedings of the 39th Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: Human Factors & Ergonomics Society.

Theunissen, E. (1994). Factors influencing the design of perspective flight path displays for guidance and navigation. *Displays, 15*(4), 241-254.

Wickens, C.D. (1992). *Engineering Psychology And Human Performance* (2nd ed.). New York: HarperCollins.

Wickens, C.D. (1995a). Situation awareness: Impact of automation and display technology. *Keynote address, NATO AGARD Aerospace Medical Panel Symposium on Situation Awareness*, Brussels, Belgium, April.

Wickens, C.D. (1995b). *Integration of navigational information for flight*. University of Illinois Institute of Aviation Final Technical Report (ARL-95-11/NASA-95-5). Savoy, IL: Aviation Res. Lab.

Wickens, C.D., and Prevett, T. (1995). Exploring the dimensions of egocentricity in aircraft navigation displays: Influences on local guidance and global situation awareness. *Journal of Experimental Psychology: Applied, 1*, 110-135.

Woods, D.D. (1984). Visual momentum: A concept to improve the cognitive coupling of person and computer. *International Journal of Man-Machine Studies, 21*, 229-244.

Performance Measures and Situational Awareness: How Strong the Link?

John M. Reising

Wright-Patterson Air Force Base

Introduction

Endsley (1988a, p.792) defines situational awareness (SA) as the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and their status in the near future". The measurement of SA has gone along three paths: subjective, objective, and performance based. While the main focus of this paper is on the performance based measures, examples of subjective and objective measures will be given in order to contrast them with the performance based.

Subjective Measures of SA

An example of a subjective measure of SA is the Situational Awareness Rating Technique (SART), (Taylor, 1989). SART is a questionnaire method which concentrates on measuring the operator's knowledge in three areas: 1) Demands on attentional resources, 2) Supply of attentional resources, and 3) Understanding of the situation. The reason that SART measures three different components [there is also a 10 dimensional version] is that the SART developers believe that , like workload, SA is a complex construct; therefore, to measure SA in all its aspects, separate measurement dimensions are required. Because information processing and decision making are inextricably bound with SA [since SA involves primarily cognitive rather that physical workload], SART has been tested in the context of Rassumessen's Model of skill, rule and knowledge based behavior. Selcon and Taylor (1989) conducted separate studies looking at the relationship between SART, and rule and knowledge based decisions respectively. The results showed that SART ratings appear to provide diagnosticity in that they were significantly related to performance measures of the two types of decision making.

Objective measures of SA

One of the most well known objective measure of SA is the Situational Awareness Global Assessment Technique (SAGAT), (Endsley, 1988b). SAGAT was developed for the simulation environment of a military cockpit but could be generalized to other systems. During random times, the simulation is stopped and the pilot is asked questions to determine SA at that particular point. A random subset of the possible SA questions [the air-to-air version has 36 questions] are asked of the pilot during a particular simulation. Answers to the questions raised at various times during the

simulation are then stored in a micro-computer (Endsley and Cayley, 1990). Following the completion of the simulation, the answers are compared with the correct answers in the simulation computer data base. "The comparison of the real and perceived situation provides an objective measure of pilot SA", (Endsley, 1988b, p.101). This same technique could be used with any complex system that is simulated, be it a nuclear power plant control room or the engine room of a ship.

Performance Based Measures of SA

Although both the subjective and objective measures appear to have the ability to measure SA, there is not always clear a relationship between SA and performance based measures. In a recently completed study (Reising, Liggett, Solz & Hartsock, 1995) conventional head up display (HUD) symbology was compared to Pathway-the-Sky HUD symbology in order to see which symbology type enabled pilots to fly curved landing approaches more easily. It turned out that the Pathway HUD resulted in significantly better pilot performance (smaller RMS error for lateral, vertical, and airspeed deviation). A number of the pilots described the Pathway as providing instantaneous situational awareness. Therefore, we have a strong relationship between performance and SA, and here is a clear case of a performance based measure of SA -- or is it?

In a second study, military instrument approach procedures, which are currently displayed on paper as pages of a book, were graphically generated on a cathode ray tube (CRT). Because previous research (Mykityshyn, Kuchar & Hansman 1994) comparing various electronic versions of paper approach procedures (Jeppesen Charts) had shown little objective performance differences, it was decided to focus on SA in addition to flight performance measures. The objective of this study was to compare four different versions of the electronic plates (North Up, Track Up, Color, and Monochrome). The size of the approach chart generated on the CRT was the same size as the paper version (6 x 8). The resolution of the CRT is less than the printed paper, and pilots had previously noted the difficulty of reading the electronic versions of the charts (Huntley, 1992); therefore, a continuous zoom feature was available to the pilots in order to solve this problem. During the study a series of probe questions were asked to obtain an indication of the pilots SA; for example, Is the highest obstacle east of the airfield? It was thought that the *number of times* the pilots changed the zoom range, as well as the *different* zoom ranges chosen when SA probe questions were asked, might be a performance based measure of SA. At the time of this writing, 7 of 16 pilots have completed the experiment; none of the zoom-related measures show any relation ship to the SA probe questions.

Conclusion

Based on the research just discussed, there does not appear to be any direct relationship between performance based measures and SA. The basic problem in both of these studies is the conceptual distance between the motor skill performance based measures and the high level of understanding implied in the SA. As mentioned previously, Selcon and Taylor (1989) related SART to two of Rassmussens three levels of decision making: rule based and knowledge based. They did not relate SART to the lowest level: skill based. The skill based behavior is automatic and requires very little complex decision making. The performance measures of flight performance and zoom control fall into the skill based level and are far removed from the complex decision making involved in achieving SA.

References

Endsley, M. R. (1988a) "Situational Awareness Global Assessment Technique (SAGAT)". In *Proceedings of the National Aerospace and Electronics Conference*, (pp. 789-795). Dayton, Ohio.

Endsley, M. R. (1988b) Design and evaluation for situation awareness enhancement, In *Proceedings of the 32nd Annual Meeting of the Human Factors Society*, (pp. 97-101). Santa Monica, CA: Human Factors Society.

Endsley, M. and Cayley, P. (1990) Using the Macintosh to measure pilot situation awareness. *The Review of Macintosh Applications*, Winter.

Mykityshyn, Kuchar & Hansman (1994) Experimental Study of Electronically Based Instrument Approach Plates, *The International Journal of Aviation Psychology, Vol(2),* Lawrence Earlbaum Associates: Hillsdale, NJ

Reising, J.M., Liggett, K.K., Solz, T.J. and Hartsock, D.M. (1995) A Comparison Of Two Head Up Display Formats Used To Fly Curved Instrument Approaches, In *Proceedings of the 39th Annual Meeting of the Human Factors Society*, Santa Monica, CA: Human Factors Society.

Selcon, S.J. and Taylor, R.M. (1989 Evaluation of the situational awareness rating technique (sart) as a tool for aircrew systems design. In *AGARD Conference Proceedings No. 478*, 7 Rue Ancelle 92200 Neuilly Sur Seine, France.

Taylor, R.M. (1989) Situational awareness: aircrew constructs for subject estimation. In *Proceedings of the 68th Aerospace Medical Panel Meeting of NATO AGARD*, Copenhagen, Denmark, 2 - 6 Oct.

The Role of Scope as a Feature of Situation Awareness Metrics

Michael A. Vidulich

Wright-Patterson Air Force Base

Abstract

This paper discusses "scope" as one of the desirable attributes of a situation awareness (SA) metric. In this context, the term "scope" refers to the extent or range of view of a SA metric. The basic premise is that SA metrics must be designed to either directly measure enough aspects of the subject's awareness of the situation to possess sufficient scope or indirectly measure the outcome of a process that integrates across enough aspects of the situation. The scope typically found in several common SA metrics is discussed. Also, the Global Implicit Measure (GIM) approach is introduced as a new metric approach that may combine the strengths of several current SA measures, including an optimal scope of measurement.

Introduction

Situation awareness (SA) refers to the pilot's cognitive understanding of the current situation and its implications. SA has become a framework concept for interpreting the impact of a wide variety of influences on mission success. However, the utility of the SA concept to the design and test and evaluation communities has been hampered by the relative paucity of practical and generally-accepted metric tools. In particular, there is a lack of good objective performance measures that can be relied upon as indicators of pilot SA.

Considerable imagination has been applied to designing and proposing various SA metrics that have been tested to varying degrees. Typically these tests emphasize such things as demonstrating sensitivity to manipulations expected to influence SA, or the reliability of SA metrics. Another common issue in metric evaluations has been the practicality of different measurement techniques to different environments (e.g., simulation studies versus flight tests). All of these tests and considerations are certainly valuable in the search for good SA metrics, and the present paper is not intended as a critique of these previous efforts. The purpose of the present paper is to discuss the issue of "scope" as a desirable attribute of SA metrics.

Scope refers to the extent or range of view of a SA metric. In general, a SA metric is intended to measure the subject's awareness of the current situation. Environments in which SA has become an important topic are usually characterized by multiple and complex sources of information (e.g., military or commercial aviation, nuclear power plant control, process control, etc.). In performing in such an environment, it is assumed that the effective operator must sample and maintain awareness of many aspects of the environment. A SA metric with a wide scope would test the extent of the subject's range of view across these many sources of information. A SA metric with limited scope would test the subject's awareness of a very limited set of features (possibly only one feature) from the environment. The present paper contends that, in most cases,

SA metrics should be designed, or selected, to incorporate a wide scope. However, as will be seen, some metrics can conceivably suffer from excessive scope. Next, several categories of SA metrics will be discussed in terms of scope.

Scope and Some Selected SA Metrics

Measures of Effectiveness (MOEs)

In some cases researchers have attempted to use measures of effectiveness (MOEs) as a direct indicator of SA. A MOE is usually an outcome measure of mission success. In some cases it might be an estimated probability of mission success based on observing the outcome of numerous simulated sorties. Or, it might be a quantification of some outcome directly correlated with mission success (e.g., number of enemy aircraft shot down). The major benefit of a MOE as a metric tool is its obvious relevance to real-world mission performance. However, a MOE is generally a highly contaminated metric. In any realistic simulation a number of parameters will vary across scenarios and even a good mission strategy will sometimes fail due to the chance combinations of events that might occur. Thus, even though we might expect improved SA to show up in an improved MOE score in the long-run, it might very well be swamped by random variability in the relatively limited number of data-collections runs used in the typical simulator tests. Finding a significant effect of cockpit configuration in MOE data is impressive, but the lack of such an effect could often be attributed to the poor statistical properties of many MOEs. Also, while better SA might be the cause of an improved MOE score, it might be from other influences. For example, improving the probability of kill of a missile may improve the average number of kills without directly influencing pilot SA.

In other words, considered as a SA metrics, MOEs might suffer from excessive scope. A good MOE would not only implicitly represent the overall adequacy of a subject's SA but also the influence of any critical non-SA variables (e.g., equipment effectiveness or reliability, weather conditions, luck, etc.). MOEs will and should remain a valuable tool for system evaluation, but trying to interpret them as a direct indicator of SA will be problematical.

Measures of Performance (MOPs)

A second approach is to use a more focused measure of performance (MOP). A MOP will typically be developed around a specific task that lends itself to careful measurement. For example, average error of the flight path around a designated landing path could be a MOP for evaluating the quality of cockpit displays. Or, the average reaction time to respond correctly to a warning signal could be a measure of that warning signal's strength and interpretability. Well-designed MOPs often have much more attractive statistical properties than do available MOEs, but suffer in terms of their interpretation. Unlike MOEs, the MOP measures are not necessarily obviously linked to overall mission success. And like MOEs, the MOPs are not necessarily linked to SA either. In most cases a MOP will only serve as a measure of one of numerous tasks that the subject must perform. Other tasks will necessarily be ignored by any given task's MOP. Thus, the scope of a MOP will often be too limited to serve as an overall SA score.

Memory Probes

A third form of performance measure is the performance of the subject on unexpected probes of his/her memory for details of the current situation. These memory probes are typically collected

during a stoppage of a simulated task. The advantage to this approach is that it is specifically designed to assess SA. It also seems obvious that a greater understanding of the current situation (represented by good performance in answering the memory probe questions) would be a likely outcome of an improved crew station. The scope of a memory probe metric is determined by the specific questions that are asked. In some cases, the subject might be asked a single focused question on a critical portion of a task (e.g., Vidulich, Stratton, Crabtree, and Wilson, 1994). Alternatively, a wide-variety of information might be asked to determine the breadth of the subject's current SA. For example, in the popular Situation Awareness Global Assessment Technique (SAGAT) a careful analysis is performed to specify the SA requirements of the task, the probe questions are then designed to test as much of that information as possible (see Endsley, 1991, for an example). Each approach has its advantages and disadvantages as a measurement approach, but in general, memory probe evaluations designed to assess a wide scope seem to be more successful as SA measurements.

The major problems with memory probe SA measures are practical. It is an exceedingly intrusive technique since it entails the actual stopping of a trial. Also, in order to get full coverage of the situation, the pool of possible questions to be asked is usually very large. At any given data collection point the specific questions to be asked will generally be randomly selected from the larger possible pool. This means that several stops must occur in every experimental condition to ensure that a sufficient number of all available questions are responded to. The number of trials required might be prohibitive in some evaluations.

Implicit Measures

As an alternative to the memory probe SA metrics some researchers (e.g., Fracker and Davis, 1991) have suggested a specific form of a MOP called implicit measures. An implicit measure of SA is a focused look at whether or not a pilot is aware of some specific critical event. The event is selected to be one that demands a timely and accurate response from the subject. One advantage of an implicit measure is its precise measurement. Another advantage is that the implicit probes are part of the normal task and are therefore not intrusive to the realism of the simulated task. Also, the resulting data is often interpretable within Signal Detection Theory. This provides strong analysis tools for the interpretation of the data. The disadvantage of traditional implicit measures is that they might be over-focused. They might provide an excellent analysis of the pilot's SA for a specific component of the task, but do not appear to have the scope to encompass the pilot's overall SA for the entire mission situation.

Subjective Ratings

Another very popular approach for measuring SA is the use of subjective ratings. Defining the scope of subjective ratings is difficult, but it is easy to suspect that subjects probably integrate a wide variety of impressions in producing their ratings, which could be considered a wide-scope. Furthermore, the use of multi-dimensional rating tools could be considered a method for widening the scope of subjective ratings. Such multidimensional tools have been demonstrated to be more sensitive than unidimensional SA ratings (e.g., Vidulich, Crabtree, and McCoy, 1993).

On the other hand, subjective ratings will naturally be insensitive to gaps in the subject's understanding of the current situation that the subject is unaware of. For example, a pilot may generate a high SA rating just before being surprised and shot down by an unobserved enemy.

A New Approach: The Global Implicit Measure (GIM)

Looking over all of the common approaches to SA measurement, there seems to be an interesting set of trade-offs between the scope of the techniques and other good or bad aspects of the technique. In particular, the implicit approach possesses such desirable qualities as unintrusiveness and well-quantified measurement, yet appears to suffer from a lack of sufficient scope. A reasonable question is whether this lack of scope could be addressed by incorporating some of the procedures common to the more global memory probe SA measures (such as SAGAT) into the implicit measurement approach.

A project that is currently underway at Armstrong Laboratory is evaluating just such a combination (see Brickman, Hettinger, Roe, Stautberg, Vidulich, Haas, and Shaw, this volume, for a much more detailed description of this project). This project will attempt to develop a new procedure for a performance-based metric of SA. Overall, the approach is to expand the concept of an implicit measure by tying a wide variety of implicit style probes to a careful task analysis of all of the goals the pilot is trying to achieve in each segment of a simulated air-to-air combat mission. The fact that the metric description process starts with a goal-oriented task analysis should provide metrics with the strong face validity enjoyed by MOEs. The implicit measures described will be assessed on a moment-by-moment basis throughout the simulated sortie, which should provide a density of data collection comparable to some of the better MOPs, this should benefit the statistical power associated with analyses of the data. By selecting implicit measures associated with a wide variety of the pilot's tasks the overall scope should be competitive to that achieved in memory probe assessments. While gaining all of these benefits, it is expected that the precision and unintrusiveness normally associated with implicit measures will be maintained.

Since the procedure involves integrating data from implicit probes present across and throughout the entire simulated sortie, it can be called the Global Implicit Measure (GIM).

A General Description of the GIM Procedure

As a starting point the use of a GIM approach for measuring SA assumes that the task environment is a complex one. The current study uses a simulated air-to-air combat mission to ensure a sufficiently complex environment. In order to have the prerequisite variety of implicit probes available it is necessary that there be a combination of goals competing for attention and a pilot with the appropriate expertise to deal with the mission complexity. To some degree, of course, these goals are designated by the experimenter and communicated to the pilot through explicit rules of engagement that will designate the pilot's current goals and constrain the pilot's tactics for achieving those goals. This task analysis and description of the rules of engagement will serve to parse the overall mission into identifiable mission segments. The transition from one segment to another is characterized by events that cause a major shift in the pilot's current goals. Segment identification is an important precursor to GIM measurement since the implicit probes will be assessed in reference to the pilot's current goals.

Within each segment, implicit probes will be defined to measure how well the pilot is moving towards achieving the segment goals within the constraints of the rules of engagement. For example, the rules of engagement might specify which of several possible targets is the appropriate one to designate. If the pilot has that target designated the current score on that implicit probe item for that moment is 1. If the pilot did not have that target designated the score for that implicit probe for that moment would be 0. The scoring would be conducted in a continuous stream at a rate equivalent to the frame rate of the simulation. For the entire segment, or for any designated time period within the segment, a proportion score for success on that element would be calculated by dividing the sum of all of the observations by the total number of observations. The score on this segment could then be combined with comparable data from similar implicit probes associated

with other segment goals as defined by the rules of engagement (e.g., weapon status, radar settings, aircraft course, aircraft attitude, etc.).

Due to the fact that all of the data will be based on the same type of proportion(success) score, sets of implicit measures associated with different goals within a segment can be averaged independently in order to provide a diagnostic SA profile for the segment. Alternatively, all of the scores for a segment could be averaged together for an overall segment SA score. Or, to carry the logic even further, the data from all segments within a mission could be averaged together to generate an overall mission SA score. During the averaging process it may be desirable to use a weighting of the different individual implicit probes to reflect expert opinion about the relative importance of the various sub-goals.

Implemented in this fashion the GIM is expected to be a sensitive, diagnostic, and flexible SA metric tool. In particular, the linking of the GIM to continuously assessing the pilot's performance relative to all relevant goals should ensure that the scope of the measure is optimal.

GIM Development Goals

The development of the GIM procedure is expected to proceed through certain steps. Success in the early steps will provide the expertise to move on to the later steps. At each step the utility of the GIM should be expanded. Briefly, the program can be expected to move through three stages: near-term, middle-term, and long-term.

In the near-term, the first set of goals is to develop the steps for implementing the GIM process, to determine the algorithms for generating the GIM scores from simulator data, and to provide the initial validation of the approach within a cockpit evaluation. In this first validation, the GIM scoring will be performed post-hoc on data collected during a current laboratory cockpit evaluation. The validity of the procedure will be tested by comparing its sensitivity to the cockpit configuration manipulation to other measurement tools. The near-term design of the GIM scoring algorithms will be designed with the middle-term goals in mind.

In the middle-term the GIM scoring algorithms (if validated in the near-term study) will be implemented in real time during the running of a future laboratory evaluation. This would provide experimenter and observers with a real-time diagnostic display of SA during a pilot's simulated sortie. This would be useful in giving the experimenter a richer basis for the feedback to be presented to the pilot. The final overall mission SA score should also provide good motivational feedback to the participating pilots.

In the long run (if the diagnostic attributes of the GIM are validated in the real-time measures) the real-time GIM could provide one type of input data for real-time inferencing engines to control adaptive interfaces.

Conclusion

The combination of the unintrusiveness and potential real-time measurement of implicit SA measures with the global approach previous employed in some memory probe SA measurement is being developed and tested. A powerful motivation for this development is to create a practical tool for SA measurement that can be optimized in terms of the scope of the measure.

References

Endsley, M.R. (1991, July). *Situation awareness in an advanced strategic mission* (Tech. Report No. AL-TR-1991-0083). Wright Patterson AFB, OH: Armstrong Laboratory. (AD-B161348)

Fracker, M.L., and Davis, S.A. (1991, October). *Explicit, implicit, and subjective rating measures of situation awareness in a monitoring task* (Tech. Report No. AL-TR-1991-0091). Wright-Patterson AFB, OH: Armstrong Laboratory. (AD-A262702)

Vidulich, M.A., Crabtree, M.S., and McCoy, A.L. (1993). Developing subjective and objective metrics of pilot situation awareness. In *Proceedings of the Seventh International Symposium on Aviation Psychology* (Volume 2, pp. 896-900). Columbus, OH: The Ohio State University.

Vidulich, M.A., Stratton, M., Crabtree, M., and Wilson, G. (1994). Performance-based and physiological measures of situation awareness. *Aviation, Space, and Environmental Medicine, 65*(5, Suppl), A7-A12.

Use of Testable Responses for Performance-Based Measurement of Situation Awareness

A. R. Pritchett, R.J. Hansman, and E.N. Johnson

MIT Aeronautical Systems Laboratory

Introduction

The use of testable responses as a performance based measurement of situation awareness is a valuable measurement technique for testing of a wide-range of systems. Unlike measurement techniques that attempt to ascertain the subject's mental model of the situation at different times throughout an experiment, performance based testing focuses solely on the subject's outputs. This quality makes it ideal for comparing the desired and achieved performance of a human-machine system, and for ascertaining weak points of the subject's situation awareness.

This paper will focus on the use of situations with testable responses during simulations. During the simulation runs, the subjects are presented with situations. The situations are designed such that, if the subject has sufficient situation awareness, an action is required. This provides an unambiguous accounting of the types of tasks for which the pilots had sufficient situation awareness.

First, this method of assessing situation awareness will be briefly compared with other methods. The use of situations with testable responses in a representative flight simulator study will be detailed. Then, because the subject's responses depend heavily on the precision with which the situations are generated, techniques for robust generation of pre-determined situations will be discussed, and the performance of a current implementation will be discussed.

A Comparison of Performance Based Measurement with Other Methods of Situation Awareness Assessment

Performance-Based Measurement of Situation Awareness has taken several forms. Some techniques measure the overall final performance of the human-in-the-loop system in any or all of its tasks (Endsley, 1995). This paper focuses on the use of Testable Responses for evaluating situation awareness, where the subjects are presented with realistic situations during the simulation runs which, if they have sufficient situation awareness, require decisive and identifiable actions.

Several other methods of testing situation awareness have been documented (Endsley, 1995; Adams, Tenney & Pew, 1995). Several complex techniques exist which attempt to determine or model the subject's knowledge of the situation at different times throughout the simulation runs. For example, the Situation Awareness Global Assessment Technique (SAGAT) freezes the simulator screens at random times during the runs, and queries the subjects about their knowledge of the environment. This knowledge can be at several levels of cognition, from the most basic of facts to complicated predictions of future states.

Several causal factors affect the actions of the subject, as shown in Figure 1. Comparing knowledge-based and performance-based techniques of evaluating situation awareness, we find

they take measurements at different points in the process of user cognition. This illustrates the different purposes for these two measurement techniques.

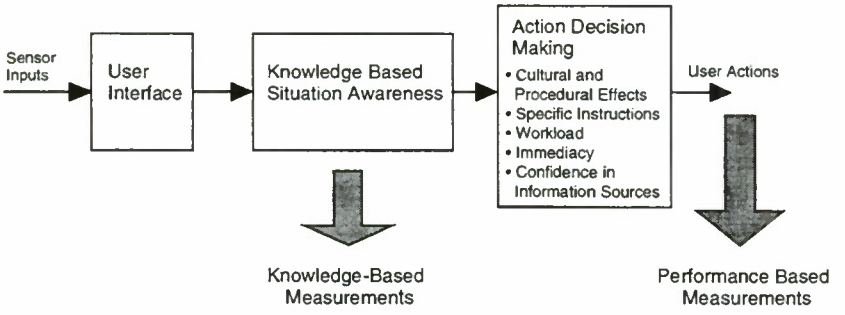


Figure 1. A Comparison of Measurement Points Between Knowledge-Based and Performance-Based Measurement Techniques

For providing a detailed, theoretical assessment of the subject's situation awareness, the knowledge based techniques are more accurate, as they measure these variables directly. Performance-based measurement can only make inferences based upon the particular information the subject acted upon, and how it was interpreted.

However, performance-based measurements can satisfy several goals that knowledge-based techniques can not. The most apparent is its ability to ascertain the timing and substance of a user's reaction to realistic situations. For testing of systems, final decisions must be based on whether the user will be provided with sufficient situation awareness to perform the correct actions, which performance-based techniques measure directly. Knowledge-based measurement techniques, on the other hand, can only make reasonable guesses about the likely user's actions given their knowledge state.

In addition, performance-based measurement provides measures of situation awareness that are not otherwise easily achievable. It can identify constraints on a user, arising from their training and standard procedures, that would not be anticipated by a strict knowledge-based model of situation awareness. For example, in a flight simulator study by Midkiff and Hansman, ATC neglected to turn the subject towards the landing runway although the subjects could overhear the aircraft before and after them being giving the proper instructions; although the subjects' actions indicated they were aware of the situation, they did not take a strong reaction because of their reticence to assume the Air Traffic controller had made an error (Midkiff & Hansman, 1993). A knowledge-based measurement of the pilots situation awareness also would have provided a measurement, in this case, of the pilot's awareness of the problem; only performance-based measurement, however, could ascertain how the pilots would act upon this information within an established set of Air Traffic Control procedures.

Performance-based measurement is also able to determine perceived reliability of the knowledge users gather from any of a multitude of sources. For example, the same simulation study by Midkiff and Hansman found pilots were often unwilling to act upon information only overheard on ATC voice frequencies because they did not have confidence in the mental model it provided (Midkiff & Hansman, 1993). The study was therefore able to ascertain whether pilots had sufficient confidence in their mental model to take action. A knowledge-based measurement, in the same study, might have concluded that the pilots had correct knowledge, but might not realize the pilots would refuse to act upon it in the same manner as if they had verifiable, correct knowledge.

Finally, performance-based measurement works well in time-critical situations to find the real-time response, rather than planned or thought-through response. Subtle variations in situation awareness or current conditions may be causal factors in different actions by the user, as shown in autopilot mode-awareness simulation, where the pilot's actual, real-time reactions often varied significantly from those they named as 'what they would do' during non-time critical questioning afterwards (Johnson & Pritchett, 1995).

In summary, performance-based measurement is complementary to knowledge-based measurement in the development of a human-in-the-loop system. Each is useful at different times, and for different purposes, throughout the design process. For final testing of a system, performance-based measurement is very useful because of its ability to ascertain the resulting performance of the entire system, and to point to areas of situation awareness that are deficient. Although performance-based measurement does not provide as pure a measurement of a user's knowledge base as other techniques, it is able to illustrate the inter-relationship between the user's knowledge and the manner in which they use it.

Use of Situations with Testable Responses in a Representative Flight Simulation Study

This section shall use a recent flight simulator study to demonstrate the use of testable responses in measuring situation awareness and overall system performance. Both the development and performance of the measurement techniques shall be discussed.

The flight simulator study by Midkiff & Hansman was conducted to evaluate pilot utilization of the Party Line Information they can overhear on shared Air Traffic Control frequencies (Midkiff & Hansman, 1993). Two-pilot air transport flight crews, using the NASA Ames Man-Vehicle System Research Facility (MVSRF), flew a 3 leg flight, during which they were exposed to nine different situations.

The design and scripting of the situations is the most crucial aspect of the experiment design. The situations must be designed to have several traits. Most importantly, the situations must be designed such that, should the user have sufficient situation awareness, a clear and unambiguous response is mandated. As illustrated in Figure 2, the task of the experimenter is to expose the user to situations which force a measurable action, without attempting to examine the specifics of the 'inner' workings of the subject, such a their knowledge state.

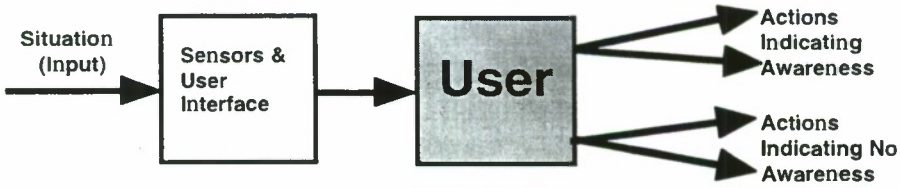


Figure 2. Use of Testable Responses to Situations

When expert-users, such as airline pilots, are used as subjects, situations can be chosen for which standard operational criteria demand a certain response. For example, one situation in the

Midkiff and Hansman simulator allowed pilots to overhear communications which suggested that another aircraft had not departed the runway the subjects were very close to landing on. In this case, action was required to avert a collision; a lack of action by the pilots could be considered to represent a lack of pilot situation awareness.

In addition, the situations should be chosen to cover the domain of important situations in which the system is expected to perform. For example, in the Midkiff and Hansman simulator study, the nine situations tested were the testable situations which had received the highest importance ratings in a pilot survey of Party Line Information importance. Testing of a final prototype system may include situations which test all conditions given in the system design specifications.

Finally, the situations must represent believable and recognizable occurrences to which the subject can be expected to react as they would in the real, non-simulated environment. For example, in the Midkiff and Hansman study, the subjects were flying an air transport simulator and believed they were over-hearing other air transport aircraft. Therefore, the 'Potential Collision' situations were staged to happen at a rate which was physically reasonable and were carefully scripted to portray to the subject a believable scenario of pilot confusion and/or mechanical failure on the part of the intruding aircraft.

The testable responses should be capable of examining the range of all probable actions and in-actions by the subject throughout the experiment. Care must be taken to look for actions which are different, less severe or incorrect in addition to just looking for the expected or desired result. For example, the response to the situation "Aircraft on Landing Runway" might be expected to be an immediate go-around. However, the subject's actions were often less severe, with pilots instead attempting to query ATC or each other to verify the knowledge they had gained from Party Line Information.

The strong reactions can be considered an indication of good situation awareness; correspondingly, the lack of any indication of awareness can be considered an indication of insufficient situation awareness. As discussed earlier in this paper, the uncertain or weak responses are also valuable measurements. They may illustrate problem areas such as lack of pilot confidence in information, feelings by the subjects that the expected reaction would defy accepted procedures, or other such unexpected impediments to action.

Performance-based measurement does not preclude other concurrent methods of assessing situation awareness. For example, Midkiff and Hansman also debriefed their subjects in an attempt to get pilot opinions on their situation awareness during the experiment.

Generating Repeatable Situations

When the purpose of an experiment is to test subjects' responses to specific situations involving multiple agents, there is a need to repeatably generate these situations across multiple trials. This is often complicated since subjects may not act consistently or as expected before the desired situation. As an illustration, consider the creation of an aircraft collision hazard. If the subject does not fly at exactly the speeds that were expected, the resulting situation can be completely different than that desired, or, as in this example as depicted in Figure 3, not occur at all.

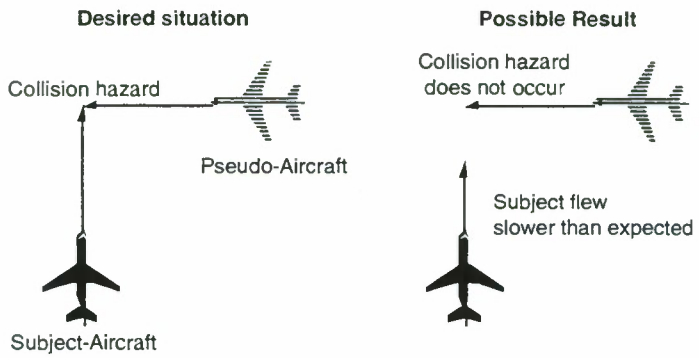


Figure 3. Situations are Dependent on Subject Actions

In order to make situations repeatable, some form of feedback of system state must be used to control the pseudo-agents (agents other than the subject), constantly controlling their actions to create the desired situations. Traditionally, this has been achieved by using experimenters to control pseudo-agents, in real-time, during the simulation run. A Robust Situation Generation architecture has been developed (Johnson & Hansman, 1995) whereby system state information is used to automatically generate scripted situations for a human subject, shown in Figure 4.

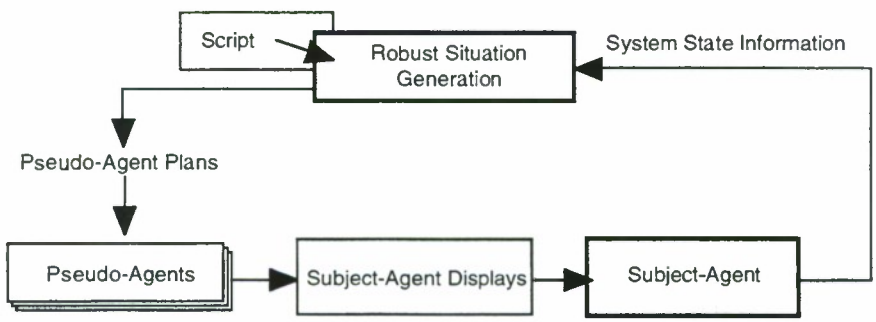


Figure 4. Overview of Robust Situation Generation

Pseudo-agents have plans that consist of a desired trajectory specified by waypoints and a discrete action plan. System state is utilized in three fundamental ways: pseudo-agent waypoints specified as relative to the subject, discrete actions of pseudo agents triggered by a cue, and cued amendments to pseudo-agent flight plans. Instances of these features are specified in a pre-determined script.

A Robust Situation Generation system has been implemented as part of the MIT Aeronautical Systems Laboratory (ASL) Advanced Cockpit Simulator (ACS), illustrated in Figure 5. A single workstation is used to simulate the pseudo-agents, consisting primarily of aircraft and controllers, and is referred to as the experimenter's station. Pseudo-aircraft state and digitally pre-recorded radio transmissions are presented to a subject operating the cockpit simulator. The scripts can be

designed interactively in preliminary simulation runs using the experimenter's station, and are then stored and used as often as required.

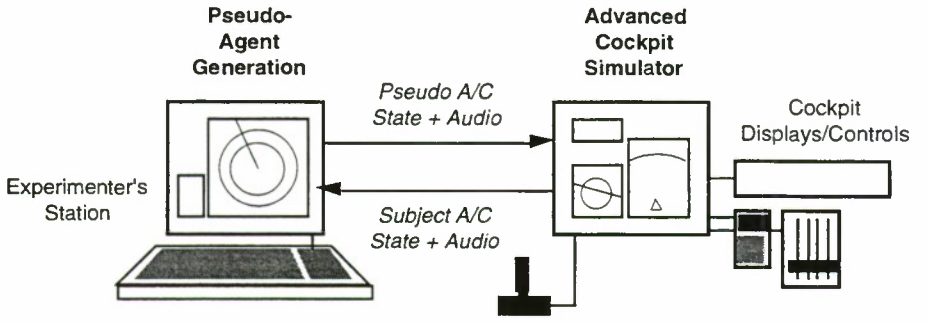


Figure 5. Implementation of Robust Situation Generation

The achieved robustness of the system, i.e. the maximum subject variation that can occur while still producing scripted situations, has been tested by varying subject-aircraft speed and position, as well as testing blunders by the subject, such as missing a turn. Unless the subject operates at an extreme limit of performance, situations were demonstrated to occur repeatably. The level of robustness depends on the level of fore thought and detail in the script, which can be made to react an arbitrary amount of subject variation, as required by the simulation.

Conclusion

Performance based measurement of situation awareness is a powerful tool for measuring the performance of a human-in-the-loop system and for identifying areas of inadequate situation awareness. The use of situations with testable responses can provide valuable insight into the user's situation awareness and how the user will act upon it.

The development of automatic robust situation generation has created a reliable mechanism for repeatable, consistent situations, making performance based measurement more reliable and easy to implement. Although the current implementation has been designed specifically for flight simulator experiments, Robust Situation Generation can also be implemented for any simulation involving multiple controllable agents.

References

Adams, M.J. et al (1995) "Situation Awareness and the Cognitive Management of Complex Systems" *Human Factors 37(1)* 85-104

Endsley, M.R. (1995) "Measurement of Situation Awareness in Dynamic Systems" *Human Factors 37(1)* 65-84

Johnson, E.N. & Pritchett, A.R. (1995) "Experimental Study of Vertical Flight Path Mode Awareness" Paper Presented at Sixth IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design and Evaluation of Man-Machine Systems, Cambridge MA, June 1995

Johnson, E.N. & Hansman, R.J. (1995) "Multi-Agent Flight Simulation with Robust Situation Generation" MIT Aeronautical Systems Laboratory Report ASL-95-2

Midkiff, A.H. & Hansman, R.J.(1993) "Identification of Important 'Party Line' Information Elements and Implications for Situational Awareness in the Datalink Environment" *Air Traffic Control Quarterly, 1*

Experiential Measures: Performance-Based Self Ratings of Situational Awareness

R.M. Taylor

DRA Centre for Human Sciences

A brief review is provided of the issues and lessons learnt in the development of SART, a practical self-rating scale tool for the subjective measurement of situational awareness (SA). The need to develop additional measures of cognitive quality in human systems is identified, that are based on task performance, for the purpose of aiding systems design and evaluation. The development of a validated set of rating scales intended to provide experiential measures of system cognitive compatibility (CC) is described, based on task-elicited personal constructs. The resultant new tool, named CC-SART, is proposed to complement SART, by providing further characterisation of the nature of cognition and understanding involved in situational awareness.

Starting with SART

Measurement is needed for systematic improvement of human performance, either through training, or by systems design. The Situational Awareness Rating Technique (SART) was developed at Farnborough (Taylor, 1990) to provide a validated and practical subjective rating tool for the measurement of situational awareness (SA). SART provides ratings of subjective dimensions associated with SA, based on 10 aircrew personal constructs. The constructs dimensions were elicited from aircrew, utilising the Repertory Grid technique, through consideration of aircrew descriptions of tactical flight situations involving SA. The structure of the construct dimensions has been interpreted as comprising three related conceptual groups, which form the principle dimensions of SART. These construct group dimensions have been identified as associated with the *Demand (D)* and *Supply (S)* of attentional resources, and with *Understanding (U)* of the situation. Three SART dimensions are associated with situational D factors *(complexity, variability, instability)*, four are concerned with attentional S factors *(arousal, concentration, division of attention, spare mental capacity)*, and three rate more cognitive U factors *(information quality, information quantity, familiarity)*. Subsequent studies provide evidence that the SART D and S dimensions are associated with task demand manipulations, and with subjective dimensions for workload assessment, specifically, the NASA TLX technique; SART U dimensions concern factors which have a strong association with task performance and the provision of SA information (Taylor et al 1995; Vidulich et al, 1995).

Understanding Understanding

The role understanding in SA is probably less well understood than the management attentional resources, and associated workload issues. The SART U dimensions distinguish between the information provided in the situation, and the familiarity brought to the situation. This is a fundamental psychological distinction, with roots in system design and training, respectively. It

reflects the interacting contributions of information obtained directly from the environment, and provided by the aircraft systems, and of the individual operators past experience, knowledge, or expertise. A formula proposed to derive a global estimate of SA from the individual SART scores, reflects the assumed pivotal role of understanding. This unitary SA index is obtained by combining the rating means on the three principle SART dimensions (3-d SART), or the equivalent rating means from the 10 personal construct dimensions (10-d SART), using a simple additive algorithm:

SA (Calculated) = Understanding - (Demand - Supply) or **SA (c) = U - (D - S)**

Studies have shown a reasonably strong association between SA(c) and task performance (Crabtree et al 1993; Vidulich et al 1993). SA(c) has been shown to have some useful *a priori* predictive power, accounting for some 30-40% of the variance in performance data (Taylor et al 1994).

We believe that it is particularly useful to think of SA as the skill associated with the management and control of perception. Perception has long been considered as an active process. Thus, the SA(c) formula derives from the proposition that SA is principally concerned with knowledge of critical features, fundamental differences, and important relationships, and of the status of important variables, in the situation. This knowledge provides the potential for exercising skilful control of perception, situation assessment, or good SA. It is considered that SART U ratings reflect knowledge of important relationships between situation variables, and that this knowledge largely determines SA. The ratings of SART D and S indicate the matching of attentional resources to changes in the situation variables. This attentional matching provides information on the current status of the variables. It acts as a modifier of SA, independent of knowledge of the important relationships, providing refinement and updating of the internal situation model in accordance with the changing status of the variables. Attentional matching increases SA when the available resources are sufficient (S>D), and reduces SA when the resources are insufficient (D>S).

Taken together, the 10-d, 3-d, and SA(c) measures provide a hierarchy of subjective measures of SA, with increasing global characterisation up through a pyramidal structure, with increasing semantic compression, and presumably with associated, but as yet largely un-proven, differences in sensitivity, predictive validity, and diagnostic power. Since SART is intended as a practical measurement tool, differences in intrusiveness, and in ease of implementation and interpretation, largely determine the most appropriate form of SART for a given application; validity tends to be an accepted, in-built assumption by practitioners, on account of the methodological rigour adopted to derive the dimensional structure.

Limitations

The limitations on SART mostly arise from generic constraints on the validity of subjective measurement (i.e. reliability, sensitivity, diagnostic power), and in particular, dissociations between subjective ratings and performance (ODonnell and Eggemeier 1986). Some fundamental limitations arise from the form and scope of the subjective dimensions. SART was developed from a broad, agreed aircrew working definition of SA, which invited consideration of related factors and variables. This approach was taken since SA was initially a term introduced by aircrew, to characterise a problem with advanced aircrew systems. The origins of the term were without any formal basis in psychological theory. Thus, aircrew knowledge elicitation seemed an appropriate method of investigation. This indicated a range of factors associated with SA, based around attention management, information and expertise. The inclusion of workload-related factors has been seen by some as a weakness, rather than a strength, of the SART technique (Endsley, 1993). This argument assumes a dissociation between workload and SA, which necessitates separate measurement. In this criticism, the presence of workload-related dimensions are considered as confounding factors, adding complication to the interpretation of the results, and

requiring caution in the inferences that can be drawn. SART provides some separation through the U and D-S ratings; but these are integrated in the global SA(c) index. The alternative seems to be metrics governed by a limiting, theoretical definition of SA, such as a knowledge-based or understanding-only proposition, that excludes consideration of other factors, not directly related to the defined concept, such as dimensions of workload. The problem with this approach lies in the circularity of defining SA in a limiting sense, and then only accepting evidence consistent with that definition. Arguably, working definitions provide a less restrictive, and more creative starting point for investigation of a problem. Like all good theories, definitions should be generative of hypotheses that enable assumptions to be tested, accepted or rejected, and refined.

The dimensions of SART have considerable generality. This generalisability arises from their basis in aircrew personal constructs, rather than in any system, or technology, specific terminology. Some critics have noted a lack of plain English in the description of the dimensions as a potential weakness (McGuinness, 1995). But equally, one might argue that in practice, it is the generality of the construct dimensions that has enabled SART to have wide application, beyond the field of aviation. Indeed, when considered from a theoretical rather than practical stand-point, the dimensions can be perceived as lacking specificity and resolution of factors involved in perception and cognition.

The additive SA(c) formula is obviously highly simplistic, particularly when compared with sophisticated models of human information processing. But the global index is intended to be indicative only of the general level of SA in situations, which may be sufficient for many assessment purposes. Weighting of the dimensions (e.g. SWAT conjoint scaling) may improve sensitivity to individual situation contexts and individual rating styles. It may need revision to account for limiting conditions, such as when there is extremely low attentional demand, or low workload, leading to boredom, complacency and reduced SA. The formula is also simplistic in that it does not reflect the complex interaction between the control of attention and higher mental functions associated with knowledge and understanding. The quality of attentional matching is likely to be enhanced in familiar situations with high rated *Understanding*, through conscious or unconscious directing of attention to important events and changes that can be anticipated and predicted based on past experience.

The difference between information quantity and quality offers some general diagnostic value for system design. But, given the key role of SART *Understanding*, greater resolution of the subjective components might be useful, particularly in relation to the perceptual and cognitive processes, since these could provide aid in the design and assessment of human systems. It was this consideration that has led to interest in the theory and development of measures of cognitive quality, and in particular cognitive compatibility (CC).

Understanding Compatibility

Concern with compatibility in systems design originates from the ideas on stimulus-response compatibility. McCormick and Sanders (1982) define compatibility in relation to human engineering design, as follows: *"... the spatial, movement, or conceptual relationships of stimuli and responses, individually or in combination, which are consistent with human expectations."* These authors go on to describe a taxonomy of different types of compatibility. *Conceptual compatibility* is described as referring to conceptual associations, intrinsic in the use of codes and symbols, or culturally acquired associations. More recently, Wickens (1984) has discussed the importance of compatibility (or congruence) between levels of representation which form the basis for understanding systems, namely the physical system (which is analogue), the *internal representation* or "*mental model*" (which should be analogue), and the interface between the two. Compatibility between the real system and the mental representation he argues is clearly a matter of training. Thus, standardisation and consistency are at least as important. If both the physical

system and the mental representation are analogue, as they should be for correct understanding, then it is important that the display should be formatted in a way which is compatible with the other two. Andre and Wickens (1992) use the notations S for stimulus, C for *comprehension or cognitive understanding*, and R for response, suggesting that S-C *mappings* are concerns of cognitive compatibility and S-R mappings are concerns of physical compatibility.

Pictorial and schema-based display formats provide practical examples of how these ideas are applied in systems design. The research literature on human-computer interaction and usability provides guidance for design of intuitively useful features, similar to ideas of cognitive compatibility. Intuitive features of graphical user interfaces are considered to include *WYSIWIG* (what you see is what you get), and simplicity of mapping between representation and product. Usability principles for learnability include predictability, sythesizability, familiarity, generalisability, and consistency. But the problem is that no formal methods are available for specifying or measuring intuitive features, pictorial quality, or schema-basedness.

Theoretical Assumptions

From a consideration of the literature, we propose the following expression as a working model for investigating the relationships between types of CC:

$$CC = K((Md)(Sp)(Mv)(Cn)(Ns)$$

(CC is Cognitive Compatibility score; K represent constants; Md is modality compatibility; Sp is spatial compatibility; Mv is movement compatibility; Cn is conceptual compatibility; Ns is not yet specified compatibility e.g. social and organisational dimensions).

On the basis of above working model, we have set out to consider the extent to which cognitive compatibility (CC), like mental workload and SA, can be considered to be a measurable cognitive condition or state. Measurement of knowledge structure using network analysis has been proposed for estimating the cognitive complexity of displays (Chechile 1992). But more commonly, the quality of CC has to be inferred indirectly from objective measures of performance, whilst making assumptions about the underlying and hidden cognitive structures and processes. As SART has shown, measurement of unobservable cognitive states, not directly available for analysis, presents practical and theoretical difficulties which affect the validity and reliability of the data. The approach we have taken rests on the fundamental theoretical assertions that degrees of CC are experienced at levels of awareness, that awareness is mediated by working memory, and that the associated mental activity leaves a memory trace.

Theories of SA stress the importance of the pilots continuously updated, mental representation or cognitive model of the situation. Maintenance of this internal model is affected by limitations on attention, associated with working memory, and by the availability of knowledge of critical features and important relationships, stored in semantic and episodic memory in the form of schemas and scripts. Recently, Baddeley (1993) has argued that conscious awareness is a means of co-ordinating information from a number of sources, including the present, specific episodes from the past, and projections as to the future, using a system operating through working memory. Gardiner and Java (1993) have argued the utility of subjective, *experiential* measures, compared with conventional measures of accuracy and performance, in distinguishing between the different states of awareness. The experiential approach focuses on measurement of the actual mental activity experienced by the individual, mediated by memory. At one level, there seems to be evidence of a fleeting awareness of recent experiences and events, mediated by primary memory operations. Memory studies of longer retention intervals show that conscious recollection of events, associated with explicit remember responses and episodic memory operations, seem to involve a different state of awareness than having feelings of familiarity, and more implicit know responses, associated with semantic memory operations. Following the same paradigm, being unaware is associated with being unremembered and unknown. Memory theory suggests that the type of processing involved in mental activity may determine awareness and the strength of the

memory trace. For example, data-driven automatic processing seems to leave a weak memory trace, whereas conceptually driven, intentional processing leaves a strong memory trace. A further important distinction occurs between implicit learning involved in the performance of automatic skilled behaviour, and explicit learning associated with the following of rules and procedures. It seems that implicit learning operates without awareness and with only a weak memory trace. Thus, subjective ratings of explicit dimensions of cognition, or of experience, are unlikely to be sensitive to implicit learning. Implicit learning may need implicit tests and implicit rating dimensions, in order to intuit, or to reason, that implicit learning has taken place. Behaviour, cognition, and experience are not necessarily correlated; the relationships need to be determined by empirical investigation. SART sets out to measure the operators appreciation or awareness of the quality of work, which may or may not be associated with the standard of the actual performance. The intention of SART is to complement performance measurement, and to provide insights into the nature of work that performance measurement might not encompass. A stronger association with performance might be obtained from subjective rating dimensions based more directly on the experience of work than the dimensions of SART which arose from consideration of SA scenarios, rather than actual experience. The following table is an attempt to summarise the kinds of information that may be available from memory, based on actual experience, and therefore may be candidates for self-report experiential measurement.

Table 1. Characteristics of Memory Stores

MEMORY STORE	FUNCTION	CHARACTERISTICS
Iconic	Immediate; Sensory; Visual	0.25>1s. Decay. 4-5 Item recall limit
Preperceptual Auditory Memory	Immediate; Sensory; Auditory	250msec. Decay
Short Term Memory	Information filter for L.T.M. Transferral involves rehearsal/ consolidation.	Minute to hours decay. Span = 7 +or-2 Items. Comprises specialised perceptual stores.
Short Term Visual Store	Visual information received from iconic store, passed to L.T.M. via repeated exposure.	Robust recognition: <48 hours. Limited capacity: New information displaces old. 0> 30s. No forgetting if free to rehearse.
Short Term Auditory Store	Auditory information received from preperceptual auditory store and passed to L.T.M. via rehearsal.	3- 40s. Decay. Selective attention. Clear recency effect in recall.
Long Term Memory (L.T.M.)	Semantic	Large body of general knowledge/ facts/ rules. Very hard to erase, but schema modification occurs via new information. Know responses
Long Term Memory	Episodic	Autobiographical; Memory for events. Constructivist rather than strictly veridical. State and context dependant recall.Remember responses
Forgetting	Allows the recording of abstractions from events, without recording the events as such.	1) Storage loss; 2) Retrieval failure; 3) Encoding deficiency; 4) Cues lacking/ overload.

Smarter than SART: Development of Experiential Measures

Scale Development Environment

In order to investigate the subjective dimensions of CC, we have sought to develop an experiential measures approach, using a task created to manipulate CC variables, based on the working model described earlier (Taylor et al, 1995). The task was a highly abstract, computer-based simulation of flying an aircraft in tactical situations. Subjects were required to provide directional (left/right) responses to a multi-modal display of situational information. Information on other aircraft locations was presented visually and auditorily, and subjects were required to make orienting responses (fly towards or away from) in relation to those locations. Spatial reasoning, visualisation, mental rotation and decision-making were key cognitive task components. The CC variables directly manipulated by the task were modality (MD), spatial (SP), movement (MV), and conceptual (CN) compatibility. The presentation of the information was designed to provide correlated and uncorrelated task cues demonstrative of varying MD, SP, MV, and CN compatibility. A total of 60 CC task situations were created demonstrating degrees of compatibility and incompatibility, with an equal number of correlated and uncorrelated task cue combinations.

Construct Elicitation

In the initial phase of the study, 30 non-aircrew subjects were presented with sub-sets of the CC task situations, and personal constructs associated with CC were elicited using a Repertory Grid procedure, similar to that used to develop SART. Subjects were guided by the following broad dictionary-based, and somewhat procrastinating, working definition for CC: *"...the facilitation of goal achievement through the display of information in a manner which is consistent with internal mental processes and knowledge, in the widest sense, including sensation, perception, thinking, conceiving, and reasoning."* They then provided subjective ratings on dimensions of the elicited constructs for 22 of the CC task situations. 56 construct dimensions (32 unique) were elicited and rated in this way. Guided by the strength of correlations of the ratings with task response times (RTs), and by the inter-correlations of the subjective ratings, the set of construct dimensions was reduced to 22.

Identification of Structure

In a second phase, 16 of the task situations, chosen to represent a range of task difficulty, were presented to 20 non-aircrew subjects, and ratings were obtained on the reduced set of 22 construct dimensions. From analysis of the ratings obtained, and of their association with the task RTs, and from consideration of the structure of the semantic associations, 10 construct dimensions were identified as characterising the main CC variability in the task. The ratings of these 10 construct dimensions appeared to be organised in three main statistical clusters and three principle components dimensions. Following group discussions with subjects who provided the personal constructs, the three principal groupings of personal constructs were identified as follows:

a. *Depth of Processing (DoP)* dimensions, *automatic*, and mostly S-R compatibility, with associated internal processes, namely: *naturalness, automaticity, association, intuitiveness.*
b. *Ease (or difficulty) of Reasoning (EoR)* dimensions, associated with working memory, *intellectual* and mostly S-C compatibility, namely: *straightforward, confusability, understandability, contradiction.*

c. *Activation of Knowledge (AoK)* dimensions, associated with learning and experience, and *schema* compatibility, namely: *recognisability, familiarity.*

Initial Validation

In the third phase of the study, the full set of 60 situations were presented to 30 non-aircrew subjects, in a balanced experimental design, with response times, accuracies, and ratings obtained on dimensions of the 10 personal constructs, and on the 3 derivative DoP, EoR, and AoK construct group dimensions. Semantic compression was used to define the 13 dimensions in terms of the meanings of associated constructs. Analysis of the results showed significant main and interaction effects on RTs and ratings of the compatibility variables, and provided evidence for the validity of the derivative primary constructs *EoR* and *AoK*. However, the *DoP* dimension generated a different pattern of ratings than the group sub-constructs. This was consistent with subjects reported difficulty in understanding the meaning of Low and High *Depth* in the context of the task and scale definition. Subsequently, this derivative group dimension has been redefined as *Level of Processing (LoP),* to clarify the intended meaning. The resultant set of 3-d, 10-d, and composite 13-d rating scales has been proposed as the basis for experiential measurement of CC, and designated as CC-SART (Cognitive Compatibility - Situational Awareness Rating Technique), to maintain the association with SA.

CC-SART

CC-SART is made up of thirteen constructs, three primary and ten subsidiary. The primary constructs supporting cognitive compatibility are defined as follows:

a . Level of Processing. *The degree to which the situation involves, at the lower score level, natural, automatic, intuitive and associated processing or , at the higher score level, analytic, considered, conceptual and abstract processing.*
b. Ease of Reasoning. *The degree to which the situation, at the lower score level, is confusing and contradictory or, at the higher score level, is straightforward and understandable.*
c. Activation of Knowledge. *The degree to which the situation, at the lower score level, is strange and unusual or, at the higher score level, is recognisable and familiar.*

Further, a hypothesis for the relationship between the CC-SART dimensions has been proposed as follows:

$$S_fCC \;\; a \;\; (K_1{*}EoR)(K_2{*}AoK) \;/\; (K_3{*}LoP)$$

(S_f is the situation identifier; CC is Cognitive Compatibility score; Kn represents anticipated constants; EoR is the Ease of Reasoning score; AoK is the Activation of Knowledge score; LoP is the Level of Processing score).

As a working model, it has been suggested that the following simple additive formula be used to derive a single index of CC(c).

$$CC(c) = AoK + EoR - LoP$$

For descriptive purposes, it seems reasonable to propose that, on the basis of this work, and thus with some ecological validity, CC could be more simply defined as follows: *...ease of processing with appropriate expectations.*

CC-SART is available from the author in hard copy and disc formats (Apple Mac Hyper-Card application, with automatic data recording to Mac Write). The following table provides examples to aid the interpretation of the CC-SART dimensions.

Table 2. Examples of CC-SART Dimensions

CC-SART DIMENSIONS	RATING	EVERYDAY LIFE EXAMPLES	FLYING AN AIRCRAFT EXAMPLES
Levels of Processing	Low	Riding a bicycle	Instrument scan.
	Medium	Playing cards	Standard radio calls.
	High	Debating an argument	Handling an emergency.
Ease of Reasoning	Low	Solving a crossword puzzle	Fuel calculations.
	Medium	Checking a telephone bill.	Operating an FMS.
	High	Following traffic lights	Selecting a Nav Aid.
Activation of Knowledge	Low	First car driving lesson	First conversion to new a/c type.
	Medium	Cooking a favourite meal.	Monitor climb.
	High	Phoning home.	Intercom call.

Consequentials

a. What is the relationship between SART and CC-SART?
b. How does SART U correlate with CC-SART?
c. How does SART (D-S) correlate with CC-SART LoP?
d. Is task performance directly related to subjective CC (10-d, 3-d, CC(c))?
e. Can performance be predicted from subjective estimates of CC (What CC-SART predictive power; develop/test Pro CC-SART)?
f. Are there relationships between system design variables and CC-SART measures?
g. What LoP, EoR, AoK weightings and combinations give best predictions of performance, and why ?
h. What is the sensitivity of CC-SART to novice/expert differences, and implicit/explicit learning differences?

References

Andre AD and Wickens CD. Compatibility and consistency in display-control systems: Implications for aircraft decision aid design. *Human Factors, 1992, 34* 6, pp 639-653.

Baddeley A. Working Memory and Conscious Awareness, in *Theories of Memory Collins AF*, Gathercole S, Conway M, and Morris P (Eds), Hove, Erlbaum, 1993.

Crabtree M.S., Marcelo R.A., McCoy A.L. and Vidulich, M.A., " An examination of a subjective awareness measure during training on a tactical operations simulator", in Jenson, R. & Neumeister, D. (eds), *Proceedings of the 7th International Symposium on Aviation Psychology,* 1993, OSU, Columbus.

Chechile RA. A review of ANNETS; A model of cognitive complexity of displays. *in Proceedings of HFS 36th Annual Meeting 1992,* pp 1176-1180.

Endsley, M.R. Situation Awareness and Workload: Flip Sides of the Same Coin. in Jenson R.S. and Neumeister D. (Eds). *Proceedings of the 7th International Symposium on Aviation Psychology,* Columbus, Ohio. April, 1993, pp 906-910.1.

Gardiner JM, and Java RI. Recognising and Remembering in *Theories of Memory Collins AF,* Gathercole S, Conway M, and Morris P (Eds), Hove, Erlbaum, 1993.

McCormick, E. and Sanders, M. *Human Factors In Engineering and Design,* New York: McGraw-Hill, 1982.

McGuinness, B. Situational Awareness: Limitations and Enhancements in the Aviation Environment. In *Situation Awareness: Limitations and Enhancement in the Aviation Environment.* 79th AGARD AMP Symposium. AGARD Conference Proceedings. AGARD, Neuilly-sur-Seine. April 1995 (In press)

ODonnell R.D. and Eggemeier F.T. Workload Assessment Methodology. in *Handbook of Perception and Human Performance, Vol II, Cognitive Processes and Performance.* Chapter 42, pp 42-1 to 46. 1986.

Taylor R.M. Situation awareness rating technique (SART): The development of a tool for aircrew systems design, in AGARD CP 478, *Situation Awareness in Aerospace Operations.* 1990. AGARD, Neuilly sur Seine.

Taylor, R.M., Selcon, S.J., and Swinden, A.D. "Measurement of Situational Awareness and Performance: A Unitary SART Index Predicts Performance on a Simulated ATC Task", *in Human Factors in Aviation Operations.* Fuller R., Johnston N., and McDonald N. (Eds) Proceedings of the 21st Conference of the European Association for Aviation Psychology (EAAP), Vol. 3. 1994, Aldershot, Avebury Aviation.

Taylor R.M., Shadrake, R, Haugh J, and Bunting A. Situational Awareness, Trust and Cognitive Compatibility. In *Situation Awareness: Limitations and Enhancement in the Aviation Environment.* 79th AGARD AMP Symposium. Conference Proceedings. AGARD, Neuilly-sur-Seine. April 1995 (In press).

Vidulich, M.A., Crabtree M.S. and McCoy, A.L.. " Developing subjective and objective metrics of pilot situation awareness", in Jenson, R. and Neumeister, D. (Eds), *Proceedings of the 7th International Symposium on Aviation Psychology,* 1993, OSU, Columbus.

Vidulich, M.A., McCoy, A.L and Crabtree M.S. "Attentional Control and Sitautional Awareness in a Complex Air Combat Simulation. In *Situation Awareness: Limitations and Enhancement in the Aviation Environment.* 79th AGARD AMP Symposium. Conference Proceedings. AGARD, Neuilly-sur-Seine. April 1995 (In press).

Wickens C.D. *Engineering Psychology and Human Performance,* Columbus: Merill, 1984.

Using Observer Ratings to Assess Situational Awareness in Tactical Air Environments

Herbert H. Bell and Wayne L. Waag

Armstrong Laboratory, Aircrew Training Research Division

Introduction

In 1991, the Air Force Chief of Staff asked a series of questions about situational awareness (SA). These questions included: What is SA? Can we measure SA? Can we select individuals for pilot training based on their SA potential? What impact does training have on SA? In response to these questions, Armstrong Laboratory initiated an SA research program. This paper summarizes our initial attempts to measure SA in operational fighter squadrons and in multiship air combat simulations. It then discusses the general problem of using subjective measures to assess performance.

Our initial efforts have focused on three issues. The first issue concerns the definition of SA. The second issue is the degree to which pilots can reliably judge their fellow pilots in terms of SA. The third issue is whether or not there is a relationship between such judgments and mission performance.

In response to the question, "What is SA?," the Air Staff provided a working definition that links SA to mission performance. This definition, written from the operator's perspective, defines SA as "A pilot's continuous perception of self and aircraft in relation to the dynamic environment of flight, threats, and mission, and the ability to forecast, then execute tasks based on that perception (Carroll, 1992)." Although there are a number of other definitions of SA available (e.g., Endsley, 1995b; Sarter and Woods, 1991; Tenney, Adams, Pew, Huggins, and Rogers, 1992), we are using this Air Staff definition as the basis for our research efforts. This definition reflects the importance of SA in mission accomplishment thus capturing the richness and complexity of the pilot's world. It emphasizes perceiving what is important and then using that perception to guide the selection and performance of appropriate behaviors. Unfortunately, it is also very complex because it combines processes, tasks, and the linkages between them into a single construct. Consequently, it is very difficult to separate SA from the other aspects of skilled performance that determine combat proficiency.

Measuring SA in Operational Fighter Squadrons

In order to determine whether or not pilots could reliably classify fellow pilots based upon SA, we limited our investigation to mission-ready F-15C pilots. With the assistance of instructor pilots and other subject matter experts (SMEs), we developed a list of 31 behavioral elements of SA. Our SMEs felt these elements reflected SA and were important to mission success. Table 1 lists these 31 elements and the eight categories of mission performance they represent.

Table 1. Elements of Situational Awareness

General Traits	*Information Interpretation*
Discipline	Interpreting VSD
Decisiveness	Interpreting RWR
Tactical knowledge	Ability to use AWACS/GCI
Time-sharing ability	Integrating overall information
Reasoning ability	Radar sorting
Spatial ability	Analyzing engagement geometry
Flight management	Treat prioritization
Tactical Game Plan	*System Operation*
Developing plan	Radar
Executing plan	TEWS
Adjusting plan on-the-fly	Overall weapons system proficiency
Communication	*Tactical Employment-BVR*
Quality (brevity, accuracy, timeliness)	Targeting decisions
Ability to effectively use information	Fire-point selection
Tactical Employment-General	*Tactical Employment-WVR*
Assessing offensiveness/defensiveness	Maintain track of bogeys/friendlies
Lookout (VSD, RWR, visual)	Threat evaluation
Defensive reaction (chaff, flares, maneuvering)	Weapons employment
Mutual support	

SA Instruments

The laboratory developed four different instruments to measure SA in operational F-15C squadrons based on the 31 elements listed in Table 1. The first instrument required respondents to provide their personal definition of SA. Using their personal definition of SA, each respondent then rated the importance of the 31 elements using a 6-point Likert scale.

The other three instruments, or SA Rating Scales (SARS), measured SA from three different perspectives: self, supervisory, and peer. All sample respondents completed the self-report and peer SARS. The self-report SARS and supervisory SARS required the respondents to rate either themselves or their subordinates on each of the 31 items. Both SARS used a 6-point scale and the ratings were made relative to other F-15C pilots. The scale anchors were "Acceptable," and "Outstanding because all respondents were on flying status and mission ready. The Squadron Commander, Operations Officer, Assistant Operations Officer, Weapons Officer, and Standardization-Evaluation Flight Examiner completed the supervisor SARS on the pilots within their squadron. In addition, squadron flight commanders completed supervisor SARS on the pilots within their flight. The peer SARS required respondents to rate the other mission-ready pilots in the squadron on general fighter pilot ability and SA ability and then to rank order them on their SA ability. Both the peer and supervisory SARS allowed respondents to omit rating a particular pilot if they felt they did not have enough information to accurately rate that individual.

Results

We obtained SA data from 238 mission-ready F-15 pilots from 11 squadrons stationed at four different Air Force bases. Two hundred and six of the respondents provided written definitions of SA. The first column in Table 2 lists the seven phases most frequently used by the respondents in defining SA. The second column shows the seven most highly rated elements of SA. There is considerable agreement between the phases used to define SA and the element ratings. In addition,

both the phases and the element ratings indicate that a significant component of SA involves assimilating and using information to guide action.

Table 2. Phases Used to Define SA and Importance of SA Elements

Most Commonly Used Phases to Define SA	*Most Highly Rated Elements for SA*
Composite 3-D image of entire situation	Use of communication information
Assimilation of information from multiple sources	Information integration from multiple sources
Knowledge of spatial position or geometric relationships among tactical entities	Time-sharing ability
Periodic mental update of dynamic situation	Maintaining track of bogies and friendlies
Prioritization of information and actions	Adjusting plan on-the-fly
Decision making quality	Spatial ability to mentally picture engagement
Projection of situation in time	Lookout for threats from visual, RWR, VSD

Analyses of the peer and supervisory SARS indicated that the pilots can reliably classify their fellow pilots in terms of SA. Internal consistency was computed for all 31 items on the supervisory SARS. The resulting measure, Cronbach's coefficient a, was 0.99. Inter-rater reliability was also estimated for the supervisor and peer SARS using an analysis of variance procedure (Guilford, 1954). For the supervisor SARS, these analyses indicated that the average reliability of each supervisor's ratings was 0.50 and the average reliability of the pooled supervisor ratings was 0.88. Similarly, the peer SARS showed an individual reliability of 0.60 and a combined reliability of 0.97. Additional detail concerning the analyses of the SARS data is available in Waag and Houck (1994).

As shown in Table 3, there was substantial agreement between supervisor and peer SARS. Table 3 also indicates that there is noticeably less agreement between the self-report SARS and the other SARS.

Table 3. SARS Intercorrelations (N = 238).

	1	2	3	4	5
1. Supervisor SARS	–				
2. Peer -- Fighter pilot ability	.89	–			
3. Peer -- SA ability	.91	.98	–		
4. Peer -- Rank order	.92	.91	.92	–	
5. Self-report SARS	.45	.56	.57	.49	–

Measuring SA in Simulated Air Combat Missions

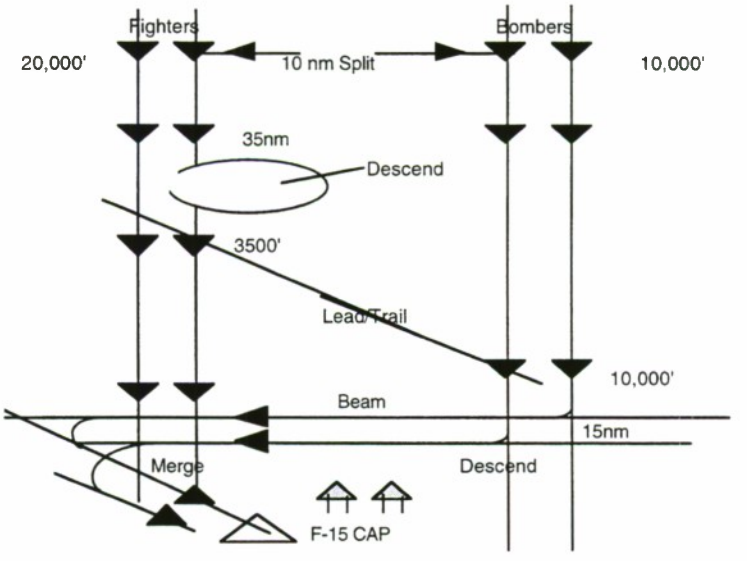
Although the SARS data indicate fairly high reliability and consistency between raters, they are not empirically linked to pilot performance in air combat missions. In an attempt to determine the relation between SA and mission performance, a composite SA score scaled with a mean 100 and a standard deviation 20 was computed for each of the 238 respondents. Based on this composite score, a sample of 40 mission-ready flight leads was selected to fly a series of multiship air-to-air combat simulations. The selected pilots covered the range of SA scores obtained for flight leads.

An additional 23 mission-ready pilots flew as wingmen during the experiment. During each week-long SA simulation, the pilots flew nine sorties with four engagements per sortie. Sorties increased in complexity throughout the week.

Scenario Design

Figure 1 illustrates a typical scenario. In this defensive counterair mission, the two F-15s are defending an airfield. The attackers consist of two bombers escorted by two fighters. The simulation begins with the enemy force 80 nautical miles (nm) away from the airfield. The enemy fighters are flying at 20,000 ft and the bombers are at 10,000 ft. There is a lateral separation of 10 nm between the fighters and the bombers. At 35 nm, the fighters maneuver rapidly and descend to 3500 ft. At 15 nm, the bombers perform a hard right turn and descend to 2500 ft. The purpose of these maneuvers is to momentarily break the F-15s' radar contact and to disrupt the F-15 pilots' ability to identify, target, or engage the enemy aircraft.

Scenarios such as these contain events that "trigger" specific goal-directed behaviors necessary for mission accomplishment. We believe that SA can be inferred based on the pilot's reaction to such trigger events. In essence, these trigger events serve as SA probes in a naturalistic environment.



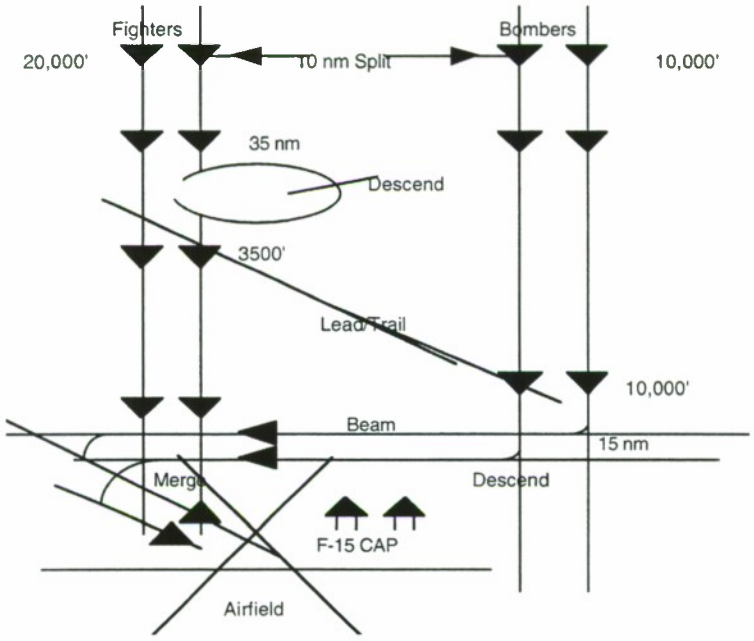


Figure 1. Defensive Counterair Mission Scenario

Rating Mission Performance

The basic approach taken toward SA measurement was through scenario manipulation and performance observation as suggested by Tenney, Adams, Pew, Huggins, and Rogers (1992). Other approaches, such as explicit probes and the Situation Awareness Global Assessment Technique (Endsley, 1995a), were considered. These other approaches were rejected because we needed measures that could be used during operational training either in simulators or actual aircraft.

As Kelly (1988) points out, measuring air combat skills presents a number of challenges. The fluid, dynamic nature of air combat, combined with the number of alternative tactics and techniques available to the pilot, make objective performance measurement extremely difficult. Even when objective data is available, it is often difficult to interpret the significance of that data. Because of the difficulties involved in interpreting air combat data, our approach is based on behavioral observation by SMEs who are unaware of the SA scores of the pilots they were observing. Two SMEs, retired fighter pilots with extensive experience in air combat and training, watched each engagement in real time and independently completed an observational checklist. To assist them in evaluating pilot performance, cockpit's instruments, intraflight communications, and a plan view display of the engagement were available throughout the engagement. After each simulator session, the two SMEs discussed each engagement and completed a consensus performance rating scale containing 24 behavioral indicators based on the SARS. In addition, the SMEs also wrote a critical event analysis for each mission that identified events that were critical to the outcome of the mission and indicative of the pilot's SA.

Results

Figure 2 shows the relationship between the composite SA scores obtained from the SARS and the mean SA score assigned by the SMEs based on their observation of performance during simulated air combat. The Pearson product moment correlation between these scores is 0.56. These data indicate that there is a significant relationship between squadron ratings of SA and performance in simulated air combat missions.



Figure 2. Simulator SA Scores and Squadron SA Scores

Discussion

We are encouraged by our initial results in developing measures of SA that can be used in a squadron's operational training environment. These results indicate that SA is a construct that has meaning and can be used by both peers and supervisors to classify mission-ready pilots. They also indicate that squadron ratings of SA are correlated with mission success in simulated air combat missions.

Although our approach to measurement may be classified as subjective rather than objective, we believe this is an oversimplification. All measurement approaches ultimately involve assigning numbers to events according to an explicit set of rules (Stevens, 1951). The distinction between objective and subjective measures simply indicates whether or not a human observer is an integral component of the measurement instrument. Objective measurement involves datum that is generated independently of the human observer. Ideally, this datum is generated, recorded, and scored without the intervention of a human observer. Subjective measurement on the other hand, requires human observers to generate the datum itself. Although Muckler (1977) argues that there is no such thing as objective measurement in the strict sense, the distinction continues to be made and "so-called" objective measures are often preferred to subjective measures. The reason for this preference is that subjective measures are frequently seen as being contaminated by the human

observers during the act of measurement. Since objective measures, on the other hand, are relatively independent of human observers, they are seen as "truer" measures of the construct under study.

Unfortunately, objective measures often fail to capture the richness and complexity of human performance (Kelly, 1988; Meister, 1989; Vreuls and Obermayer, 1985). One reason for this is that objective measures are essentially reductionistic and are therefore best suited for recording the fundamental dimensions of performance (e.g., latency, amount, and deviation). While these fundamental measures provide us with data that is less subject to error, they also frequently fail to provide us with information concerning the contextual nature of skilled performance. Subjective measures, on the other hand, seem more closely related to higher order psychological constructs. The datum they produce appears to reflect a synthesis of the more molecular behaviors and to reflect more global dimensions such as interpreting, judging, and deciding--the very essence of SA.

Obviously both measurement approaches are necessary if we are to develop our understanding of SA. The critical measurement issues are how do we refine our definition of SA and our measurement approaches and which measurements provide the best information for designing and evaluating aircrew training.

References

Carroll, L. A. (1992). Desperately seeking SA. *TAC Attack (TAC SP 127-1)*, 32, pp 5-6.

Endsley, M. R. (1995a). Measurement of situation awareness in dynamic systems. *Human Factors*, 37, pp 65-84.

Endsley, M. R. (1995b). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37, pp 32-64.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw Hill.

Kelly, M. J. (1988). Performance measurement during simulated air-to-air combat. *Human Factors*, 30, pp 495-506.

Meister, D. (1989). *Conceptual aspects of human factors*. Baltimore, MD: Johns Hopkins University Press.

Muckler, F. A. (1977). Selecting performance measures: "Objective" versus "subjective" measurement. In L. T. Pope and D. Meister (Eds.), *Symposium proceedings: Productivity enhancement: Personnel performance assessment in Navy systems*. San Diego, CA: Navy Personnel Research and Development Center.

Sarter, N.B., and Woods, D.D. (1991). Situation awareness: A critical but ill-defined phenomenon. *International Journal of Aviation Psychology*, 1, pp 45-57.

Stevens, S. (1951). Mathematics, measurement, and psychophysics. In S. Stevens (Ed.), *Handbook of Experimental Psychology*. New York: Wiley.

Tenney, Y. J., Adams, J. J., Pew, R. W., Huggins, A. W. F., and Rogers, W. H. (1992). *A principled approach to the measurement of situation awareness in commercial aviation.* National Aeronautics and Space Administration, NASA Contractor Report 4451.

Vreuls, D., and Obermayer, R. W. (1985). Human-system performance measurement in training simulators. *Human Factors*, 27, pp 241-250.

Waag, W. L., and Bell, H. H. (1994). A study of situation assessment and decision making in skilled fighter pilots. Paper presented at the Second Conference on Naturalistic Decision Making, Dayton, OH, June.

Waag, W. L., and Houck, M. R. (1994). Tools for assessing situational awareness in an operational fighter environment. *Aviation, Space, and Environmental Medicine*, 65 (5, Suppl.), pp A13-A19.

SA Measurement: Lessons Learned from Workload

Gary B. Reid

Wright-Patterson Air Force Base

Introduction

It is almost impossible to observe the debate over a definition of situation awareness (SA) without recalling a very similar debate over a definition for mental workload. Both mental workload and situation awareness are hypothetical constructs that must be inferred rather than directly observed. Hypothetical constructs, are defined as theoretical concepts that are used to describe knowledge about observable behaviors and to provide implications about new behaviors. A hypothetical construct should not be thought of as correct or incorrect. Instead, it should be considered more or less useful for explaining present knowledge and for suggesting new relationships to be empirically verified. This notion of usefulness is particularly appropriate for SA and mental workload because both of these constructs, for the most part, came from the operational community. Operational users largely feel like they can recognize good or poor SA or high and low mental workload. If users of modern complex systems feel like these are useful constructs and discuss the human machine interface in these terms, then it is self-defeating for human factors practitioners, who must work designing and testing these systems, to *arbitrarily* refuse to use the terms. Design and test decisions will be made based on these constructs with or without human factors professionals. We can better serve by working to refine the constructs and their measurement than we can by withdrawing behind a screen of scientific purity while leaving the important systems decisions to disciplines whose practitioners are not so timid about exercising their judgment.

I think that is where the research community finds itself today in regard to situation awareness and I am certain that is where we found ourselves in the early 1980's with regard to mental workload. I think that we should be better able to cope with such a fuzzy, ambiguous construct because of our experiences in dealing with workload. I don't mean to imply that the mental workload problem is "solved," but I do feel that we have made considerable progress in understanding and measuring workload and our work, in turn, has had an impact on the design of new systems.

Mental Workload

When I started in workload measurement in 1980, our program at the Human Engineering Division of the Armstrong Laboratory was composed of two research thrusts; a performance measurement thrust and a physiological measurement thrust. It was common wisdom, at that time, that subjective measures were the most commonly used measurement approaches and, therefore, we assumed them to be the most well developed. Repeatedly, in consultations with people from the test and evaluation community, we discovered that what would best fit the constraints of their

test programs was some kind of subjective approach. Still, thinking that the measurement tools were "out there", we conducted an extensive review of current approaches with the aim of developing recommendations for subjective measures based upon current literature and research experience. What we found was that subjective measurement approaches *were* in wide use, but, in general, they were measures by fiat, developed specifically for each test program or experiment. We didn't find any measurement approaches that were systematically developed and documented so that other users would know how to apply the same approach in new situations and know what to expect from the measure. In response to what we perceived to be a research deficiency, we established a third research thrust in subjective measurement aimed at testing various approaches that were currently in use and developing new subjective approaches where appropriate. What seemed to be needed was a program of development and evaluation similar to programs directed toward measurement of other physiological phenomena that would clearly define the methods, procedures, and measurement attributes.

Measurement Attributes

What kind of measurement attributes should we be concerned about? Generally, everyone wants to know whether or not the tool measures what it is intended to measure and will it provide consistent results. That sounds simple enough, but in practice, especially for measures of an emerging inferred hypothetical construct, it is very difficult.

Validity

The problem of validity arises because psychological measurement is indirect. Evidence for how difficult this indirect characteristic makes things is that the relatively simple concept of validity is divided into several different types of validity in measurement textbooks. The particular type that is most appropriate for this discussion is called construct validity. When evidence that has been gathered, has implications for or depends upon the existence of a hypothetical construct, it is referred to as construct validity. For a construct like SA, a program aimed toward establishing construct validity of the measure also serves to fine tune the construct definition. To establish construct validity, we must design task conditions that vary in some orderly way. If a measure shows differences in situations where they should, and fails to show differences where they should not, then there is evidence to support the construct and this measure's ability to index it. If the measure fails this test, then either the measure is not an index of the construct, or something is incorrect about the structure of the construct, or both. This, of course, can be a very frustrating and demanding endeavor for the researcher, but in the long run, this iterative process provides a better understanding of the construct and produces measures that have an empirical basis as indices of the construct. In practice, I have often been asked questions like,"What good is my workload measure when all it does is confirm the obvious?" The answer to that criticism, of course, is that a measure had *better* confirm the obvious when there is a reason to expect certain results. If it doesnt then there isnt sufficient evidence to justify confidence in the measure as an index for workload when the results are not so obvious. Additionally, a measurement technique is beneficial for stating levels of the construct in quantitative terms which makes accumulation and description of differences easier. It should be apparent from this discussion, that it is unlikely that a measure will ever be really "validated". Rather, like we said about a construct, a measure can be shown to be more or less useful under specified circumstances. Through an iterative process of application and modification, both will be refined and partially validated until they converge.

Reliability

The measurement characteristic of reliability also poses special problems. Reliability is an indication of the extent to which a measure contains variable errors. Within normal psychometric theory, the psychological construct (e.g., intelligence or mathematical aptitude) being measured is considered a stable attribute of the individual. The challenge is to take repeated measurements under identical conditions with the expectation of obtaining the same measured result. Construct such as mental workload have internal components, like resource allocation, or frustration that may vary even when all external conditions are held constant. Assuming that the construct is formulated correctly, and that variation in one of these internal components is really what is going on, then measurements may vary from one administration to the next because of true variation as well as measurement error. Normal correlational approaches to reliability must be performed, but the researcher needs to be aware that variation may not be totally measurement error. For example, in a study investigating the reliability of the Subjective Workload Assessment Technique (SWAT), a subject performed a laboratory task on a Friday and again on a subsequent Monday. The second administration was scored as much higher mental workload than the first. When the subject was debriefed, it was discovered that the subject had spent the weekend sick in bed and still was not feeling well. A skeptic would say that the measurement was biased by the subject's state of health. A true believer would say that there was much greater mental workload associated with task performance on day two due to the subject's state of health. Whichever is correct, the point is that this poses special problems for establishing measurement reliability for those internally based constructs. For this reason, I have previously argued that a measure of variance accounted for is at least as important as a measurement characteristic as reliability.

Measurement Selection Criteria

Tom Eggemeier (1984) has proposed an additional set of criteria that should be used to select a workload measure. I think that these criteria are equally appropriate for SA. These criteria are (1) sensitivity, (2) diagnosticity, (3) intrusiveness, (4) implementation requirements, and (5) operator acceptance.

A frequent recommendation of researchers for selecting mental workload measures is to use multiple measures whenever possible. There are several reasons why this is true. One is that more information is always better. A second that is more related to this discussion is that due to the various threats and constraints associated with measurement in operational situations, individual measures will have strengths and weaknesses. Careful selection and application of a number of measures will hopefully provide converging information about the condition under investigation. If the measures fail to complement one another, then valuable information about the limitations of the construct and/or the measurement technique will be the result, but probably at the expense of the real test objectives. Hopefully, sufficient research data about the candidate measure will keep this from happening.

Measurement Development Process

The process that has been used for development of mental workload measures has been for research groups to develop and define measures including detailed administration procedures, stated theoretical underpinnings, and frequently, analysis packages. These measures have then been evaluated and refined in controlled environments to obtain the data necessary to establish the

measures utilized and identify shortcomings. The measures have then been released to other researchers for evaluation and application. This process provides human factors practitioners with alternatives and specifications for selection of the measure, or measures, that fit his or her specific application. The work on measurement of mental workload has demonstrated that this is an effective approach and one that now appears to be appropriate for the study of situation awareness.

Conclusion

In conclusion, even though all mental workload issues have not been resolved, I think that it is appropriate to say that workload research and measurement has been a success. Workload measures in general and subjective measures specifically, have been widely used for evaluation of human systems interfaces for new systems. I think that it is very important for the research community to apply similar approaches to the study of Situation Awareness in order to empirically determine the level of usefulness of this construct.

References

Eggemeier, F. Thomas (1984). Workload metrics for system evaluation. In*: Proceedings of the DRG Panel VIII Workshop Applications of Systems Ergonomics to Weapon System Development.* Shrivenham, England.

(NOR DOC 87-83). Hawthorne, CA: Northrop Corporation.

Endsley, M. R. (1988). Situation Awareness Global Assessment Technique (SAGAT). *Proceedings of the National Aerospace and Electronics Conference (NAECON)* (pp. 789-795). New York: IEEE.

Endsley, M. R. (1989a). *Final report: Situation awareness in an advanced strategic mission* (NOR DOC 89-32). Hawthorne, CA: Northrop Corporation.

Endsley, M. R. (1989b*). Tactical simulation 3 test report: Addendum 1 situation awareness evaluations* (81203033R). Hawthorne, CA: Northrop Corporation.

Endsley, M. R. (1990a). Predictive utility of an objective measure of situation awareness. *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 41-45). Santa Monica, CA: Human Factors Society.

Endsley, M. R. (1990b*). Situation awareness in dynamic human decision making: Theory and measurement.* Unpublished doctoral dissertation, University of Southern California, Los Angeles, CA.

Endsley, M. R. (1993a). Situation awareness and workload: Flip sides of the same coin. In R. S. Jensen and D. Neumeister (Ed.*), Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 906-911). Columbus, OH: Department of Aviation, The Ohio State University.

Endsley, M. R. (1993b). A survey of situation awareness requirements in air-to-air combat fighters. *International Journal of Aviation Psychology*, 3(2), 157-168.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors, 37*(1), 65-84.

Endsley, M. R., and Bolstad, C. A. (1994). Individual differences in pilot situation awareness. *International Journal of Aviation Psychology, 4(3),* 241-264.

Endsley, M. R., and Rodgers, M. D. (1994*). Situation awareness information requirements for en route air traffic control* (DOT/FAA/AM-94/27). Washington, D.C.: Federal Aviation Administration Office of Aviation Medicine.

Hogg, D. N., Torralba, B., and Volden, F. S. (1993). A *situation awareness methodology for the evaluation of process control systems: Studies of feasibility and the implication of use* (1993-03-05). Storefjell, Norway: OECD Halden Reactor Project.

Sheehy, E. J., Davey, E. C., Fiegel, T. T., and Guo, K. Q. (1993, April). *Usability benchmark for CANDU annunciation - lessons learned.* Paper presented at the ANS Topical Meeting on Nuclear Plant Instrumentation, Control and Man-Machine Interface Technology, Oak Ridge, TN.

Direct Measurement of Situation Awareness in Simulations of Dynamic Systems: Validity and Use of SAGAT

Mica R. Endsley

Texas Tech University

Introduction

The Situation Awareness Global Assessment Technique (SAGAT), is a global tool developed to assess SA across all of its elements based on a comprehensive assessment of operator SA requirements (Endsley, 1987; Endsley, 1988; Endsley, 1990b). Using SAGAT, a simulation employing a system of interest is frozen at randomly selected times and operators are queried as to their perceptions of the situation at that time. The system displays are blanked and the simulation is suspended while subjects quickly answer questions about their current perceptions of the situation. As a global measure, SAGAT includes queries about all operator SA requirements, including Level 1 (perception of data), Level 2 (comprehension of meaning) and Level 3 (projection of the near future) components. This includes a consideration of system functioning and status as well as relevant features of the external environment.

SAGAT queries allow for detailed information about subject SA to be collected on an element by element basis that can be evaluated against reality, thus providing an objective assessment of operator SA. This type of assessment is a direct measure of SA - it taps into the operator's perceptions rather than infers them from behaviors that may be influenced by many other factors besides SA. Furthermore it does not require subjects or observers to make judgments about situation knowledge on the basis of incomplete information, as subjective assessments do. By collecting samples of SA data in this manner, perceptions can be collected immediately (while fresh in the operators' minds), reducing numerous problems incurred when collecting data on mental events after the fact, but not incurring intrusiveness problems associated with on-line questioning. By including queries across the full spectrum of an operator's SA requirements, this approach also minimizes possible biasing of attention, as subjects cannot prepare for the queries in advance since they could be queried over almost every aspect of the situation to which they would normally attend. The method is not without some costs, however, as a detailed analysis of SA requirements is required in order to develop the battery of queries to be administered.

The SAGAT technique has thus far been shown to have a high degree of validity for measuring SA. SAGAT has been shown to have predictive validity, with SAGAT scores indicative of pilot performance in a combat simulation (Endsley, 1990a). Content validity was also established, showing the queries used to be relevant to SA in a fighter aircraft domain (Endsley, 1990b). Empirical validity has been demonstrated through several studies which have shown that a temporary freeze in the simulation to collect SAGAT data did not impact performance and that such data could be collected for up to 5 or 6 minutes during a freeze without running into memory decay problems (Endsley, 1990b; Endsley, 1995). A certain degree of measurement reliability has been demonstrated in a study that found high reliability of SAGAT scores for four individuals who participated in two sets of simulation trials (Endsley and Bolstad, 1994).

SAGAT has been used to perform evaluations of avionics systems (Endsley, 1988), display designs (Bolstad and Endsley, 1990; Endsley, 1989b), and display hardware configurations

(Endsley, 1989b), supporting test and evaluation during design concept development across a variety of considerations. In addition, it has been useful in conducting research on factors related to SA, including an investigation of the relationship between SA and workload (Endsley, 1993a) and an investigation of factors leading to individual differences in SA (Endsley and Bolstad, 1994).

Despite previous studies which have shown no ill effect of inserting freezes in a simulation on subject performance, the greatest concern regarding the use of SAGAT for measuring SA has focused on its possible intrusiveness. To further explore this possibility, a study was conducted to investigate whether operator performance is affected by merely the threat of a stop to collect SAGAT data. That is, are operators somehow altering their behavior during simulation trials in which they feel they may be stopped and tested on their SA? To answer this question, a study was conducted so that performance on trials in which subjects were told that only performance would be measured could be compared to trials in which subjects were told that a stop to collect SAGAT data might occur. In the later case, SAGAT stops occurred only half of the time. Any effect of the actual SAGAT stop could therefore be differentiated from merely the threat of the stop, and compared to trials in which subjects knew they would not be stopped.

Method

Procedure

A set of trials was conducted of an air-to-air fighter sweep mission. The subject, flying as the pilot of single aircraft, was to penetrate enemy territory, maximizing kills of enemy fighters while maintaining a high degree of survivability. Four digital aircraft were the adversaries in these engagements. Subject instructions were manipulated during the test. In one-third of the trials, subjects were told that only performance would be measured. In the other two-thirds of the trials, subjects were told that there might be a stop to collect SAGAT data in addition to performance measurement. Half of these trials actually were stopped once at a random point in the trial for two minutes to collect SAGAT data. Each of six subjects completed five trials in each of the three conditions: no stop/none expected, no stop/stop expected, stop/stop expected. The conditions were presented in a random order. A total of 90 trials were completed. Pilot performance in terms of kills and losses was collected as the dependent measure.

Facilities

The test was completed using a medium fidelity mission simulation on a Silicon Graphics 4D-220 computer. The system has a high-resolution, 19" color display monitor and realistic stick and throttle controls. A simulated head-up display, tactical situation display, vertical situation display, fuel gage and thrust gage were provided.

Subjects

Six subjects participated in this test. The subjects were all experienced former military fighter pilots. The mean subject age was 43.6 years (range of 33 to 57). They had an average of 2803 hours (range of 1500 to 3850) and an average of 15.2 years (range of 7 to 25) of military flight experience. Two of the six subjects had combat experience.

Results

Analysis of variance was used to evaluate the effect of the test condition (no stop/not expected, no stop/stop expected, and stop/stop expected) on each of the two performance measures: aircraft kills and losses. The test condition had no significant impact on either performance measure, $F(2, 87) = .15$, $p = .861$, $F(2, 87) = 1.53$, $p = .223$, shown in Figures 1 and 2. In viewing the data, it can be seen that the number of kills was almost identical, independent of whether subjects expected a stop or not and independent of whether they actually experienced a stop. While the subject died slightly more often in the trials where they expected a stop but did not receive one, this difference was not significant. This data supports the null hypothesis, indicating that a stop or even the threat of a stop to collect SAGAT data does not have a significant impact on performance.

Discussion

The results of this study confirm previous findings which have not found a demonstrable effect on performance of freezes in a simulation to collect SAGAT data. It furthermore expands on these studies to reveal that even the threat of a stop does not significantly impact performance. Subjective comments by the subjects after the study confirm this. They reported that the information about whether to expect a SAGAT stop was irrelevant to them. At least on a conscious level, they were not preparing in any way for the SAGAT test. The results of this study indicate that they were not doing so unconsciously either. Overall, the results of this study indicate that using SAGAT to collect data on situation awareness is not intrusive on subject performance, and therefore provides an additional indication of the validity of the method for directly measuring subject SA during simulations.

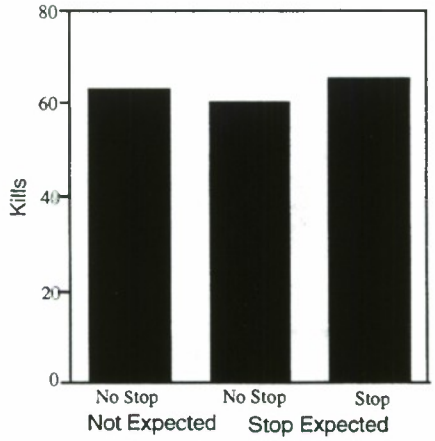


Figure 1. Aircraft Kills

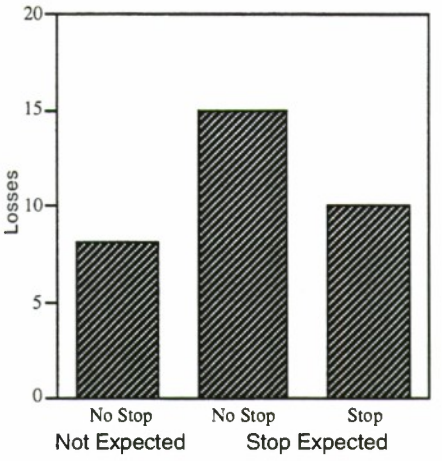


Figure 2. Aircraft Losses

Implementation of Recomendations

Several recommendations for SAGAT administration have been made based on previous experience in using the procedure (Endsley, 1995).

Training

An explanation of the SAGAT procedures and detailed instructions for answering each query should be provided to subjects before testing. Several training trials should be conducted in which the simulator is halted frequently to allow subjects ample opportunity to practice responding to the SAGAT queries. Usually three to five samplings are adequate for a subject to become comfortable with the procedure and to clear up any uncertainties in how to answer the queries.

Test Design

SAGAT requires no special test considerations. The same principles of experimental design and administration apply to SAGAT as to any other dependent measure. Measures of subject performance and workload may be collected concurrently with SAGAT, as no ill effect from the insertion of breaks has been shown. To be cautious, however, half of the trials may be conducted without any breaks for SAGAT so that a check is provided for this contingency, if performance measures are to be collected simultaneously with SAGAT data.

Procedures

Subjects should be instructed to attend to their tasks as they normally would, with the SAGAT queries considered as secondary. No displays or other visual aids should be visible while subjects are answering the queries. If subjects do not know or are uncertain about the answer to a given query, they should be encouraged to make their best guess. There is no penalty for guessing, allowing for consideration of the default values and other wisdom gained from experience that subjects normally use in decision making. If subjects do not feel comfortable enough to make a guess, they may go on to the next question. Talking or sharing of information between subjects should not be permitted. If multiple subjects are involved in the same simulation, all subjects should be queried simultaneously and the simulation resumed for all subjects at the same time.

Which Queries to Use

Random selection
As it may be impossible to query subjects about all of their SA requirements in a given stop due to time constraints, a portion of the SA queries may be randomly selected and asked each time. A random sampling provides consistency and statistical validity, thus allowing SA scores to be easily compared across trials, subjects, systems and scenarios.

Due to attentional narrowing or a lack of information, certain questions may seem unimportant to a subject at the time of a given stop. It is important to stress that they should attempt to answer all queries anyway. This is because (a) even though they think it unimportant, the information may have at least secondary importance, (b) they may not be aware of information that makes a question very important (e.g. the location of a pop-up aircraft), and (c) if only questions of the highest priority were asked, subjects might be inadvertently provided with artificial cues about the situation that will direct their attention when the simulation is resumed. Therefore a random selection from a constant set of queries is recommended at each stop.

Experimenter controlled
In certain tests it may be desirable to have some queries omitted, due to limitations of the simulation or characteristics of the scenarios. For instance if the simulation does not incorporate aircraft malfunctions, the query related to this issue may be omitted. In addition, with particular test designs it may be desirable to insure that certain queries are presented every time. When this occurs, it is important that subjects also be queried on a random sampling from all SA requirements and not just on those related to a specific area of interest to the evaluation being conducted. This is due to the ability of subjects to shift attention to the information they know they will be tested on. What may appear to be an improvement in SA in one area may just be a shift of attention from another area. When the SAGAT queries cover all of the SA requirements, no such artificial cueing can occur.

When to Collect SAGAT Data

It is recommended that the timing of each freeze for SAGAT administration be randomly determined and unpredictable enough so that subjects can not prepare for them in advance. If the freeze occurrence is associated with the occurrence of specific events, or at specific times across trials, prior studies have shown that the subjects will be able to figure this out (Endsley, 1988), allowing them to prepare for them or actually improve SA through the artificiality of the freeze cues. An informal rule has been to insure that no freezes occur earlier than three to five minutes into a trial to allow subjects to build up a picture of the situation and that no two freezes occur within one minute of each other.

The result of this approach is that the activities occurring at the time of the stops will be randomly selected. Some stops may occur during very important activities that are of interest to the

experimenter, others when no critical activities are occurring. This gives a good sampling of the subjects' SA in a variety of situations During analysis the experimenter may want to stratify the data to take these variations into account.

How Much SAGAT Data to Collect

The number of trials necessary will depend upon the variability present in the dependent variables being collected and the number of data samples taken during a trial. This will vary with different subjects and designs, but between 30 and 60 samplings per SA query with each design option have previously been adequate in a within subjects test design.

Multiple SAGAT stops may be taken within each trial. There is no known limit to the number of times the simulator can be frozen during a given trial. Experiment two found no ill effects of as many as three stops during a 15 minute trial. In general, it is recommended that a stop last until a certain amount of time has elapsed and then the trial is resumed, regardless of how may questions have been answered. Stops as long as two minutes in duration were used with no undue difficulty or effect on subsequent performance. Stops as long as five minutes were shown to allow subjects access to SA information without memory decay in experiment one.

Data Collection

The simulator computer should be programmed to collect objective data corresponding to the queries at the time of each freeze. Since some queries will pertain to higher level SA requirements that may be unavailable in the computer, an expert judgment of the correct answer may be made by an experienced observer who is privy to all information, reflecting the SA of a person with perfect knowledge. A comparison of the subjects' perceptions of the situation (as input into SAGAT) to the actual status of each variable (as collected per the simulator computer and expert judgment) results in an objective measure of subject SA. Questions asked of the subject but not answered should be considered incorrect. No evaluation should be made of questions not asked during a given stop.

It is recommended that answers to each query be scored as correct or incorrect based upon whether it falls into an acceptable tolerance band around the actual value. For example, it may be acceptable for a subject to be 10 MPH off of actual groundspeed. This method of scoring poses less difficulty than dealing with absolute error (see Marshak et. al., 1987). A tabulation of the frequency of correctness can then be made within each test condition for each SA element. As data scored as correct or incorrect are binomial, the conditions for analysis of variance are violated. A correction factor ($Y' = \arcsin(Y)$) can be applied to binomial data, however, which allows analysis of variance to be used. In addition, a chi-square, Cochran's Q, or binomial t-test (depending on the test design) can be used to evaluate the statistical significance of differences in SA between test conditions.

Limitations and Applicability for Use

This technique has primarily been used within the confines of high-fidelity and medium-fidelity part-task simulations. This provides experimenter control over freezes and data collection without any danger to the subject or processes involved in the domain. It may be possible to use the technique during actual task performance if multiple operators are present to insure safety. For example, it might be possible to verbally query one pilot in flight while another assumes flight control. Such an endeavor should be undertaken with extreme caution, however, and may not be appropriate for certain domains.

A recent effort (Sheehy, et al., 1993) employed an adaptation of this technique by making video-tapes of an ongoing situation in a nuclear power plant control room. These tapes were then

replayed to naive subjects with freezes for SAGAT queries employed. It is not known how different the SA of subjects passively viewing a situation may be from subjects actually engaged in task performance, however this approach may yield some useful data.

Most known uses of SAGAT have involved fighter aircraft simulations. In general, it can be employed in any domain where a reasonable simulation of task performance exists and an analysis of SA requirements has been made in order to develop the queries. SAGAT queries have been developed for advanced bomber aircraft (Endsley, 1989a) and recently for en route air traffic control (Endsley and Rodgers, 1994). Several researchers have begun to use the technique in evaluating operator SA in nuclear control room studies (Hogg, et al., 1993; Sheehy, et al., 1993). The technique has also been employed in a simulated control task to study adaptive automation (Carmody and Gluckman, 1993). Potentially it could also be used in studies involving automobile driving, supervisory control of manufacturing systems, teleoperations and operation of other types of dynamic systems.

References

Bolstad, C. A., and Endsley, M. R. (1990). *Single versus dual scale range display investigation* (NOR DOC 90-90). Hawthorne, CA: Northrop Corporation.

Carmody, M. A., and Gluckman, J. P. (1993). Task specific effects of automation and automation failure on performance, workload and situational awareness. In R. S. Jensen and D. Neumeister (Eds.), *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 167-171). Columbus, OH: Department of Aviation, The Ohio State University.

Endsley, M. R. (1987). *SAGAT: A methodology for the measurement of situation awareness*

SACRI: A Measure of Situation Awareness for Nuclear Power Plant Control Rooms

Stephen G. Collier and Knut Follesø

Institutt for energiteknikk

We describe a technique ('SACRI') for measuring situation awareness in nuclear power plant control rooms. We intend SACRI to be supplementary to other operator performance measures we use for evaluations of nuclear power plant MMI. We have developed and begun validation of the technique with four methodological studies using our PWR simulator and a training simulator at a commercial PWR power plant.

Introduction

The operating crew within the central control room of a nuclear power plant must maintain a comprehension of the status and changes of key plant parameters; this allows them to take timely and appropriate decisions. That is, they must maintain 'situation awareness' (SA). Their SA affects the diagnosis of any disturbances or accidents and the planning of control actions (Endsley, 1993; Wirstad, 1988). In maintaining SA, the operators integrate their overall knowledge of the process and process dynamics with information received from.the control room displays. Since the quality of the operators' SA can affect their overall task performance, it follows that the MMI should support SA (O'Hara, 1993; Roth et al., 1993).

Operator performance measures within simulator evaluations often include fault detection time and diagnostic accuracy. However, these measures do not directly assess the system's ability to enhance SA; they tap the initial and final stages of information processing, as opposed to an intermediate stage of maintaining SA (Blackman et al., 1992; Hogg et al., 1994, 1995). As far as we are aware, there is no measure of SA adapted for process control research. We therefore initiated a research project to develop such a measure to supplement others in system design evaluations. This paper summarises four methodological studies to develop SACRI. These are described at greater length in Hogg et al. (1995) and in full in a project report Hogg et al. (1994).

The Situation Awareness Control Room Inventory (SACRI)

We developed the 'Situation Awareness Control Room Inventory' (SACRI) within the Halden Man-Machine Laboratory (HAMMLAB). This contains a full-scope simulation of a nuclear power plant, based on a pressurised water reactor at Loviisa, Finland. The MMI in HAMMLAB is fully computerisedthere are no traditional hard-wired panels or controls. Various operator support systems developed at Halden can be coupled to the simulator and a range of disturbance scenarios simulated. There are facilities to log process parameters, alarms, and operator actions.

We adapted SACRI from the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1993) developed for application to pilot performance and aviation. The questions cover a range of parameters affected by disturbance situations. We selected them with the help of operators and other process experts from the Loviisa PWR plant.

The questions are phrased to ask about process trends, not absolute values. Responses are fixed choices, for example: 'increase', 'same' and 'decrease'. The latest version of SACRI covers 35 process parameters in three time frames, giving 105 questions in total. For example:

- *Past:* 'In comparison with the recent past, how has the level in the pressuriser developed?'
- *Present:* 'In comparison with the normal status, how would you describe the current level in the pressuriser?'
- *Future:* 'In comparison with now, how will the level in the pressuriser develop over the next few minutes?'

Four Methodological Experiments

So far, we have carried out four methodological studies to develop and evaluate SACRI:

- *Studies 1 & 2:* HAMMLAB, two licensed operators from the Halden Boiling Water Reactor who were cross-trained on the PWR simulator.
- *Study 3:* HAMMLAB, internal Halden research staff who had varying degrees of knowledge of the simulated process
- *Study 4:* Loviisa plant training simulator in Finland, using a licensed Loviisa operating crew.

Typical Experimental Procedure

Typically, our simulated fault scenarios last around two hours, containing several disturbances of varied severity. At unpredictable points during the scenarios we freeze the simulator, turn subjects away from the displays, and ask 12 questions randomly chosen by computer from SACRI. In the latest version, the selection is structured so as to ensure an even distribution across the plant processes and across the different time-frames. We decide in advance when and how often to freeze the simulator in each scenario. For example, one scenario used 1_ hours of run time, during which we froze the simulator and administered SACRI on 13 occasions.

Finally, on every simulator freeze, we add in questions concerning how the subjects perceive the current situation in relation to task goals. These free responses are not included in SACRI scoring but aid interpretation later.

Scoring

We score the operator's responses to each question by comparison with time-tagged trend logs of the parameters. Each response is either:

- a hit (significant parameter drift reported) or
- a miss (significant parameter drift not reported) or
- a correct rejection (no significant parameter drift, none reported) or
- a false alarm (no significant parameter drift, but the subject reports one).

This classification allows calculation of the Signal Detection Theory measures of sensitivity and response bias (A´ and R:S ratio; Wickens, 1992) . These measures are non-parametric, in the sense that their calculation does not make assumptions about the normality of the signal and noise distributions.

Results

SACRI was able to detect differences in subject competence and alarm interfaces, and could detect variations in SA during scenarios. These findings give preliminary indications of SACRI's sensitivity, reliability and validity, and suggest directions for further development.

Sensitivity

A precondition for sensitivity is that the response data do not show floor or ceiling effects. No such effects were observed in any of the studies. The inventory of questions appeared to be set at an appropriate level of difficulty for the subjects used and the situations simulated.

Sensitivity to Differences in Subject Competence
There was some evidence that SACRI is sensitive to differences in operator competence:

- *Studies 1 & 2:* one of the two subjects achieved better A´ scores on all six disturbances than the other, with this difference corresponding to results of a pre-test of process knowledge and the accuracy of their diagnoses (Figure 1).
- *Study 3:* differences between the six subjects predicted before the study were confirmed by the response data; the *a priori* rank order of competence correlated with A´ scores (Spearman's $R = -0.81$, $p = 0.05$). Also, the A´ score negatively correlated with the variance (Pearson's $r = -0.94$, $p < 0.01$).
- *Study 4:* the licensed Loviisa crew performed significantly better than the non-licensed Halden test subjects ($t = 3.01$, $p < 0.01$).

Sensitivity to Process Changes
There was also some evidence that SACRI is sensitive to changes in subjects' awareness as situations change due to process disturbances:

Study 2: two scenarios were constructed, each containing an easy, an intermediate and a difficult disturbance. Figure 1 shows the subject scores for these. The performance in each disturbance follows the same pattern for both subjects, the A´' scores becoming lower as the disturbance became more difficult

Study 3: A´ scores were compared from immediately before and after the introduction of process disturbances. A´ scores dropped after a disturbance ($t = 2.24$, $p < 0.05$).

Sensitivity to MMI Changes
There was some identifiable change in SA as a result of interface changes. In study 3 we investigated sensitivity to interface differences through the manipulation of an alarm system interface. The system has two alarm displays, both VDU-based: a chronological text list and a graphical overview. The text list was experimentally frozen (prevented from updating) for some periods in the scenarios; the graphical overview was available throughout the simulated scenarios.

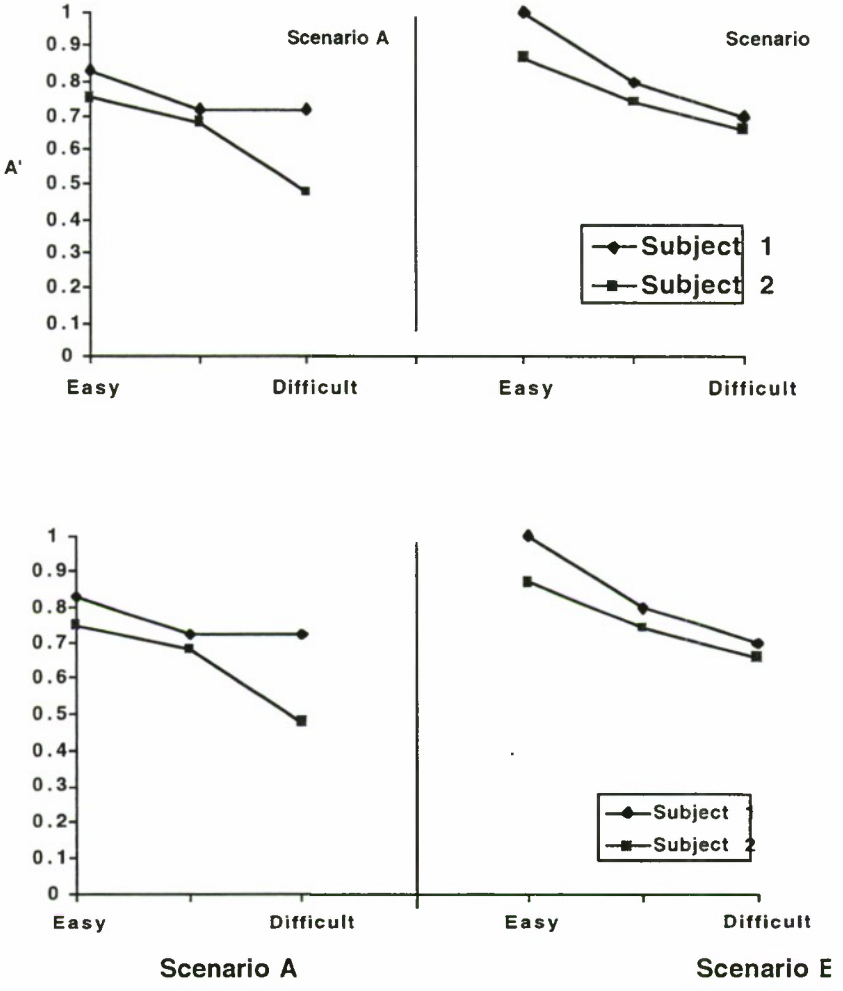


Figure 1. SA Scores for Each of Three Disturbances (Easy, Intermediate & Difficult) in Two Scenarios (A, B) Second Study.

There was little difference between the two interface configurations when the data from all six subjects and all SACRI administrations were aggregated. When examined individually, one subject proved significantly better with a *frozen* (not updating) alarm text than a live list ($t = -2.13$, $df = 19$, $p < 0.05$). Line graphs for each subject appeared to show a reduction in A´ whenever the alarm list changed from updating to not updating. Figure 2 gives an example for one subject.

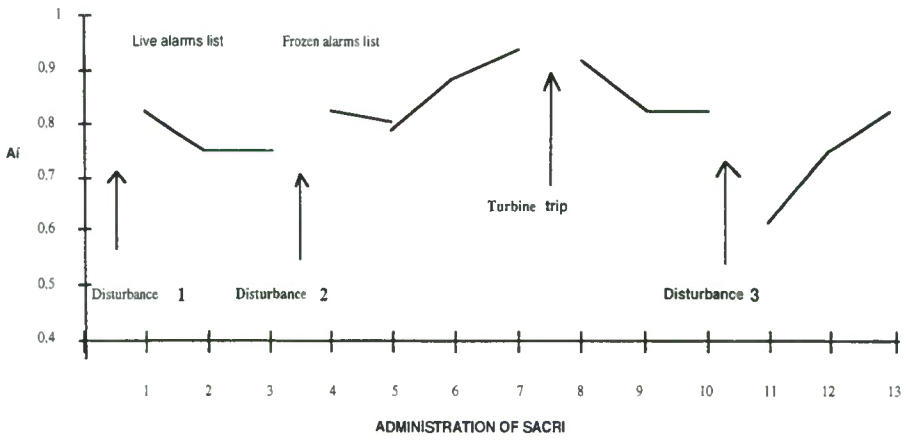


Figure 2. A′ Trends for Subject 1, Scenario A, Third Study.

To investigate this further, we used an *a priori* classification of subjects into a higher-scoring (n = 3) and a lower-scoring group (n = 3), on the basis of engineering/nuclear qualifications and experience in maintaining the process simulation. The higher-scoring group appeared sensitive to interface differences; A′ for the SACRI administrations immediately before and after alarm list changes was consistently lower when the alarm list was unfrozen after being frozen. There was no specific directional effect when the list was frozen after being live (Table 1). The lower-scoring group did not display this difference.

Table 1. Comparison of A′ Scores Before and After Alarm Text List Changes, Third Study

Subject Group	Alarm List Change	Mean, frozen list	Mean, live list	Value of t	df	Significance level
Upper 3	Frozen (live	0.81	0.62	2.9	6	$p < 0.05$
Upper 3	Live (frozen	0.70	0.70	.03	6	NS
Lower 3	Frozen (live	0.56	0.69	-.80	7	NS
Lower 3	Live (frozen	0.33	0.58	1.4	6	NS

This was unexpected. The availability of both the live text list *and* the graphical overview may have provided too much information or detail for the higher-scoring group, reducing their speed of assimilation of information and therefore their SA. There is a need to confirm the finding and investigate the causes.

The main finding for the purpose of this study is that the A′ measure has shown the potential of being sensitive to changes in the alarm system interface. This suggests it would be of value in assessing system designs.

Reliability & Validity

Although we have not yet carried out much work on SACRI's validity and reliability as a test, we have some indications:

- *Content validity:* We have put considerable effort into compiling SACRI, culminating in an item-by-item review with the Loviisa licensed operating crew. The crew concluded that the current version of the question inventory covers key indicators of process changes over a range of difference disturbance situations.
- *Predictive validity:* The *a priori* rank ordering of subjects' competence in the third study was the same as the rank order of the A´ scores.
- *Reliability:* We found initial indications of SACRI's reliability; there were consistent score differences between the two operators in the first two studies.

Time Frames Within SACRI

In study 3 we also analysed the relationship between A´ scores calculated separately for each time frame (past, present, future) within SACRI. These are shown in Table 2.

Table 2. Pearson Product Moment Correlations Between A´ Scores in SACRI Time-frames, Third Study.

Inventory Time-frame	Past	Present	Future	Overall Score
Past	-0.26	-0.26	-0.22	0.39
Present			0.20	0.46
Future				0.78(

($p < 0.01$, $df = 10$)

No correlation coefficients were significant, except for that between A´ for items relative to the future and the overall A´ score. We take this as evidence that SACRI is measuring separate components within SA.

Conclusion And Future Work

We will continue to develop SACRI as a measure of SA. Some further evaluation of SACRI is required, such as repeating the results obtained so far with a larger number of licensed PWR operators. However, we have reached the stage where we can apply it to a system evaluation. A first application will be to the designs of two alternatives for a Computerised Alarm System (CASH) for HAMMLAB (Miazza, 1993). This is currently under development. SACRI will be used with other performance indicators to provide an evaluation of the effectiveness of each design proposal.

We have recently utilised SACRI in a study of staffing in nuclear power plant control rooms. Operators' performance with different crew sizes and plant configurations were compared. SACRI was modified to provide the SA measure for each operator's primary area of responsibility, together with the general SA for the entire plant. We will publish the results of this study shortly.

References

Blackman, H., Nelson, W., Hahn, H. (1992). Measurement of human performance in complex task environments. *Human Performance, 5*(4), pp 329-351.

Endsley, M.R. (1993) Situation awareness in dynamic human decision making: theory. *Proceedings of the 1st International Conference on Situational Awareness in Complex Systems,* Orlando, February 1993.

Endsley, M.R. (1993) Situation awareness in dynamic human decision making: measurement. *Proceedings of the 1st International Conference on Situational Awareness in Complex Systems,* Orlando, February 1993.

Hogg, D.N., Follesø, K., Volden, F.S., & Torralba, B. (1994) *Measurement Of The Operator's Situation Awareness For Use Within Process Control Research.* Halden, Norway: OECD Halden Reactor Project (HWR-377).

Hogg, D.N., Follesø, K., Strand-Volden, F., Torralba, B. (1995) Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms. *Ergonomics 38*(11), pp 2394-2413

Miazza, P., Torrabla, B., Kårstad, T., Moum, B., Follesø, K. (1993). *CASH: Computerised Alarm System For HAMMLAB. An Outline Of Required Functions.* Halden, Norway: OECD Halden Reactor Project (*HWR-362).*

O'Hara, J.M. (1993) The effects of advanced technology systems on human performance and reliability. In: *Proceedings of the Topical Meeting on Nuclear Plant Instrumentation, Control and Man-Machine Interface Technologies* (pp. 253-259). La Grange Park, IL: American Nuclear Society.

Roth, E.M., Randall, J.M., Stubler, W.F. (1993) Human factors evaluation issues for advanced control rooms: a research agenda. In: *Proceedings of the IEEE Fifth Conference on Human Factors and Power plants.* New York: IEEE.

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

Wirstad, J. (1988) On knowledge structures for process operators. In: Goodstein, L.P., Andersen, H.B. & Olsen S.E. (eds.) *Tasks, Errors and Mental Models.* London: Taylor & Francis.

Measurement and Analysis of Situation Awareness in Anesthesiology

Stephen D. Small

Harvard Medical School

Situation awareness (SA) has not been measured or fully described in the medical domain although the topic has been briefly introduced by investigators working with simulators that attempt to reproduce the task environment of the anesthesiologist (Gaba 1995). Perioperative clinical activities seem well-suited to characterization by SA terminology. Performance in this context is highly time-constrained, tightly coupled to other workers and task-oriented, although many unregulated degrees of freedom exist for exercise of individual preferences. Whole task units, or cases, are embedded in a complex health care delivery system that presents numerous organizational constraints, subtle and latent subcomponent interactions, significant risk, time pressure, and multiple players with conflicting goals. There are shells of individual and team situation awareness (Endsley 1994a, Salas 1994, Bowers 1994). Examples include recognizing and managing near-awake states during light anesthesia in healthy patients to handling shock conditions during multitrauma cases, major vascular procedures or extirpative cancer operations in patients with serious infections and multiple organ dysfunction. Healthy patients scheduled for simple, elective operations can still present management issues due to latent disease, lost or uncaptured information or innumerable systems interactions that trigger unexpected, undesirable problems. Complex patients moving rapidly across the boundaries of several health care teams during periods of busy, routine work or reduced staffing periods present significant challenges at all operational levels in information management, decision making and the taking of optimal action.

The need to measure SA in anesthesiology arose initially from the desire to perform reliable, valid, and relevant physician performance representation in the context of realistic enactment of standardized scenarios in a perioperative simulation environment. The substrate for application of the performance tool would be on-line behaviors and action or videotapes. This data would be used to test theories of the effectiveness of simulator training - ideally advancing at some point to controlled studies of actual patient care and clinician management of actual critical incidents. In addition, it made sense to attempt to describe multiple levels of skill acquisition and create benchmarks to train for specific expert behaviors.

Attempts to use checklists, Likert scales, and documentation of algorithm execution proved unsatisfying during real-time action. Too much happened too quickly. Review of videotapes proved essential. Yet, although the surface of the scenario could be quickly mapped, the deep structure of the integrated activity defied superficial or point-by -point analysis. It also appeared that a wide range of worker preferences were probably acceptable. Simulator artifacts existed, such as hypervigilance, unfamiliarity with the environment, and misperception or difficulty processing certain cues important to the formation of correct mental models for particular scenarios. It was difficult to eliminate judging bias when faced with a variety of worker styles and even harder to say anything about outcome in the face of marked interpatient variability, poorly substantiated claims and controversies in the medical literature about pathophysiology, and lack of robust outcome data from the field. To be able to say something meaningful about higher levels of skill or to adequately analyze suboptimal strategies for feedback and training, better benchmark scenario-independent tools to describe perception, comprehension, integration, processing ability, social and organizational skills, decision-making and risk-coping were needed.

During the running of weekly simulator training sessions at the Boston Anesthesia Simulation Center (BASC), a five Harvard hospital collaborative project facility opened in early 1994, it became apparent that the word "awareness" (in a colloquial sense) played a key role in articulating trainee behaviors during facilitated small group debriefing sessions. The core curriculum run at the center has been adapted from Gaba's pioneering work in grafting cockpit resource management onto the perioperative domain (Gaba 1988, Howard 1992, Gaba 1994). Participants play one of three roles: the hot seat or primary anesthesia care-giver, the first responder or blinded helper called to assist in a crisis, and the hidden, passive observer. A fourth role, that of an actor or confederate in the scenario, has grown in importance as the immediacy of being in the action has been reported to enrichen learning and discovery for participants. Since this last part requires role-playing skills, familiarity with scenario details and goals, and the assignment of closely observing the action (high multitasking ability) it is usually reserved for more experienced participants. A compromise can be reached by creating a low-profile but realistic intraoperative role such as a visitor or company technician.

Observers behind one-way glass and trainees critiquing their own behaviors on videotape immediately after their scenarios uniformly questioned aspects of the action from their multiple perspectives during debriefings. Clearly, hidden observers had the advantage of minimal stress or time pressure, informal colleague consultation and more rapid detection of subtle cues. Without being subjected to the workload of actual task accomplishment and communication, observers often maintained a big picture vantage point and anecdotally succumbed less frequently to fixation or expectations than actual players. Hot seat trainees, however, argued persuasively for their mental models of the situation (dependent partly on their degree of assertiveness and ability to reflect and articulate quickly). Even relatively unskilled practitioners experienced a ratio-visceral sense of understanding , with varying degrees of subtlety, the presence (McCoy 1994), level of directed arousal (Taylor 1994) or "fit" of activated knowledge (Sarter 1994) of the hot seat trainee.

Anecdotal experience with covert manipulation of actual field workspaces of anesthesiologists revealed a wide variety of levels of awareness of changes in evolving situations. Some could have their pockets picked blind of registered narcotics or be oblivious of instructor-induced EKG lead faults, altered intravenous fluid rates, or changes in intraoperative personnel. Others might make a mental note of a trivial problem, seem to forget it, and then smoothly check and resolve it after variable periods of time and distractions. Certain individuals seemed to have developed highly efficient iterative reevaluation behaviors and be able to maintain steady levels of wariness without paranoia; others showed evidence of occasional inability to quickly redirect priorities when offered a key but slightly subtle cue. Some with superior processing abilities were so engaged in the whole task that literally no change in the environment went unnoticed. Their decision-making capabilities might be limited by experience and fewer known choice of options or successful pattern solutions, but their tracking followed the dynamic task closely. Further study of SA seemed warranted as a candidate concept that might more clearly represent these convictions and observations. It was felt that aspects of SA could serve as a foundation for comparing naturalistic behaviors in a multivalent performance tool that also contained social psychological measures (team awareness and efficiency) and weighted scores for accomplishment of critical actions demonstrating core competency. It also seemed useful to distinguish between dynamic mental tracking of reality and decision-making ability. One might have a good idea of what was going on, and what might happen, but be unable to efficiently and assertively develop convictions, plans, and enact them. A deeper evaluation of core knowledge competency seemed dependent on mapping the mental models which activate, modify, and transfer
that knowledge from assessment to assessment.

In the search for methods to measure SA, it became clear that many of Endsley's observations applied (Endsley 1994b). Trainees' actions did not necessarily mirror what they were thinking on several levels simultaneously. Post hoc subjective reasoning during debriefing represented summed rationalizations for past actions combined with altered or deficient memory traces for events. Global performance measures were suspect due to contributions of factors other than SA and affected by difficulty weighting suspension of trainee disbelief, missing cues, unnatural sequences and other simulation effects. Assignment of outcomes to scenarios could also be

viewed as arbitrary given the degree of patient plasticity, lack of performance benchmarks, and heresay status of some applied medical knowledge. Disguised probes embedded in scenarios elicited a variety of responses. Some of these were self-conscious and artifactual, some were abbreviated remarks which might be interpreted as knowledge or comprehension deficiencies but actually represented a workload coping strategy. Relatively innocent questions during action also seemed to have the unpredictable effect of altering the field of trainee attentional distribution, thus changing the variable being measured. Peer ratings appeared colored by personal style, attitude, convictions, and idiosyncratic knowledge of the literature, as well as being influenced by recent case experiences in practice.

Pilot experiments were then designed based on the Situation Awareness Global Assessment Technique (SAGAT) (Endsley 1994b, 1987). Preliminary expedited approval of the Subcommittee for Human Subjects of the Massachusetts General Hospital was obtained. Aims of the project include studying the feasability of adapting SAGAT to another domain, refining the SAGAT tool for anesthesia simulation, and attempting to measure physician SA objectively. The design of the BASC facility lends itself well to pausing scenarios and quelling data displays within seconds. Currently, the curriculum is run as a part-task trainer focused on the anesthesiologist since there are no realistically fulfilling tasks for surgeons to perform; actors/instructors playing surgeons and circulating nurses operate with a mix of independence and audio headset direction.During minimally intrusive scenario interruption, other "team" members have fallen silent and stayed in character when the action resumed. The other option of having the team continue routine work during SAGAT interruption was attempted; subjects fedback that this technique distracted from SAGAT measurement, as they were concerned they might be missing important action while being questioned. Based on the work and observations of Endsley, four to six interruptions of one to three minutes have been used during 30-45 minute scenarios. During prototype measurement periods, subjects have been asked both brief questions and given short data sheets to complete. Data is elicited in a manner designed to release case-specific knowledge with analogs of charts, history and physical forms, anesthetic records, and pictures of anesthesia machines. Videotapes of all scenarios will be available for analysis. A post-hoc questionnaire has also been designed to collect data on self-reporting of situation awareness and decision making. At some point, objective assessment from the workers' point of view of the culture, values and myths of the work space must be taken into account as one of the final arbiters of performance (Rasmussen 1983).

Subjects instinctively and quickly learned to check data displays when the SAGAT investigator entered the environment. This response was so rapid that even though analog waveforms disappeared immediately the subjects were able to pick up the slower quelled time-averaged numerics from monitors. Queries were scored as relatively non-intrusive, and easier to incorporate into the scenario flow with practice. Initially data for query was selected randomly to limit directing subject attention to key scenario elements. Anecdotally it appears that this approach may falsely bias towards the finding of ncomplete mental models as skilled subjects may not be able to quickly recall non-activated knowledge in an irrelevant context. Only past or present data points have been queried.

Methods of capturing higher levels of integration and projection have not been attempted; narrative queries and a greater level of intrusiveness during complex social interactions have been shunned. While SAGAT has been criticized as likely altering the very phenomenon it is supposed to be measuring, it is encouraging to note that the variance thus introduced may be small enough to lack significance if one considers the strong dominance of patterns of work of skilled professionals engaged in a task (Flach, communication). It is not known if SAGAT can be successfully performed in a true perioperative team environment, i.e., totally spontaneous context with all players and no instructors.

In closing, it should be noted that obstacles to measurement of SA perioperatively include the problems of risk and privacy associated with videotaping and on-line data collection in real patient care settings, and the low frequency, unpredictability and nonreproducibility of real high-workload, non-routine situations. The impact on actual patient care delivery and expense of realistic simulation sessions that require skilled physicians as operators and subjects can also be

prohibitive. Two conditions put this value judgement in further perspective. First, qualitative modeling using cognitive and social psychological methods command only a fraction of the funding and attention allocated to basic science endeavors and technology development. Second, few large, detailed databanks exist in the relatively unregulated medical domain which demonstrate the saliency of human performance and situation awareness issues in specific outcomes, cases, or system efficiency. Several recent sizable studies[1] and the current regulatory and payor focus on downsizing the medical industry while increasing quality may help to rectify these conditions

References

Bowers C.A., Braun C., Kline P.B. (1994) Communication and Team Situational Awareness. In R.D. Gilson, D.J. Garland, J.M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 305-311). Daytona Beach, Florida: Embry-Riddle Aeronautical University Press.

Endsley M.R. (1994a) Situation Awareness in Dynamic Human Decision Making: Theory. In R.D. Gilson, D.J. Garland, J.M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 27-58). Daytona Beach, Florida: Embry-Riddle Aeronautical University Press.

Endsley M.R.(1994b) Situation Awareness in Dynamic Human Decision Making: Measurement. In R.D. Gilson, D.J. Garland, J.M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp.79-97). Daytona Beach, Florida: Embry-Riddle Aeronautical University Press.

Endsley M.R. (1987). *SAGAT: A methodology for the measurement of situation awareness* NOR DOC 87-83). Hawthorne, CA: Northrop Corporation

Flach J., communication.

Gaba DM, DeAnda A (1988): A comprehensive anesthesia simulator environment: Recreating the operating room for research and teaching. Anesthesiology 69:387-394

Gaba D., Howard S., , Fish K. (1994). *Anesthesia crisis resource management.* Churchill-Livingstone.

Gaba D., Howard S., Small SD. (1995) Situation Awareness in Anesthesiology. *Human Factors* 37:25

Howard S, Gaba D, Fish K (1992): Anesthesia crisis resource management training: teaching anesthesiologists to handle critical incidents. Aviat Space Environ Med 63:763

McCoy C.E., Woleben J.K. (1994). Individual Differences in Weather Situation Awareness and Assessment. In R.D. Gilson, D.J. Garland, J.M. Koonce (Eds.), *Situational Awareness in Complex Systems* (p.239). Daytona Beach, Florida: Embry-Riddle Aeronautical University Press.

Rasmussen J. (1983). Skills, Rules and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models. *IEEE Transactions on Systems, Man, and Cybernetics,* SMC-13, pp. 61-70.

Salas E., Stout R.J., Cannon-Bowers J.A. (1994). The Role of Shared Mental Models in Developing Shared Situational Awareness. In R.D. Gilson, D.J. Garland, J.M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 297-304). Daytona Beach, Florida: Embry-Riddle Aeronautical University Press.

Sarter N.B., Woods D.D. (1994) "How in the world did I ever get into that mode?" Mode Error and Awareness in Supervisory Control. In R.D. Gilson, D.J. Garland, J.M. Koonce (Eds.), *Situational Awareness in Complex Systems* (p. 118). Daytona Beach, Florida: Embry-Riddle Aeronautical University Press.

1 The Harvard Medical Practice Study I, II; Australian Incident Monitoring Study; Adverse Drug Event Study I,II; The ASA Closed Claims Project - references available on request

Taylor R.M. and Selcon S.J.(1994) Situation in Mind: Theory, Application, and Measurement of Situational Awareness. In R.D. Gilson, D.J. Garland, J.M. Koonce (Eds.), *Situational Awareness in Complex Systems* (p.70). Daytona Beach, Florida: Embry-Riddle Aeronautical University Press.

Team Situation Awareness Research: Many Paths to a Destination

Carolyn Prince[1], Eduardo Salas[1], Clint Bowers[2], and Florian Jentsch[2]

[1] Naval Air Warfare Center Training Systems Division
[2] University of Central Florida

Attempts to measure a construct are valuable and necessary contributions to its understanding. Measurement helps in building accurate and valid models of a construct and in the articulation and identification of representative instances of it (Brannick, Prince, Prince, and Salas, in press). Several approaches to measurement that attempt to delimit the highly complex construct of team situation awareness are beginning to reveal its shape and definition. These varied approaches to understanding team situation awareness, including query technique, communication content analysis, self report, and communication pattern analysis, have all been incorporated into a theory-based, coherent research effort. This effort is directed toward developing a conceptualization of team situation awareness that can be used to design training strategies for enhancing the situation awareness of the team.

Background

This research effort has started with a framework for team situation awareness (Salas, Prince, Baker and Shrestha, 1995) that recognizes the interplay between the situation awareness of each individual crew member, the team's processes, and the context. According to Salas et al. (1995), because team situation awareness has elements of both individual situation awareness and certain team processes, it logically follows that measurement needs to be applied to both. They also pointed out that several researchers have suggested that mental models are important to situation awareness (see for example, Robertson and Endsley, 1994; Stout, Cannon-Bowers and Salas, 1994; and Wellens, 1993) and because this suggestion needs to be tested, evidence about the relevant mental models of the crew members should be measured as well. There are problems in the measurement of indications of each of these constructs (individual situation awareness, team processes, and mental models) since knowledge about each is limited. In the area of individual situation awareness, all existing measurement instruments have their own particular shortcomings (Fracker, 1991). For team process, there are a number of instruments that have been developed in aviation alone, but none have been evaluated sufficiently to determine their relative worth. Finally, attempts to measure mental models have been especially problematic (Salas et al., 1995). Therefore, it is clear that a pragmatic approach to the measurement of team situation awareness will necessarily include a variety of approaches.

Toward an Understanding of Team Situation Awareness

In earlier research, behaviors that were important to team performance were identified through interviews with over 200 military aviators (Prince and Salas, 1993). These behaviors were categorized by team researchers and aviation subject matter experts as representing indications of team aeronautical decision making, leadership, adaptability, communications, assertiveness, planning, and situation assessment or awareness. The behaviors that were specifically related to team situation assessment and awareness were used in addition to the Salas et al. (1995) framework to form the core, or base, upon which the program of research is being constructed.

To begin building an understanding of team situation awareness from the perspective of a crewmember, this research was started by conducting a series of interviews with aviators (Prince and Stout, 1995). Although self report based on reflection is generally considered an unreliable measurement method, it can yield important information that is not readily available from any other source. For example, professional pilots have experience in their work environment that is distinct from the experiences of those in many other work settings. This experience occurs in a job setting that is not easily accessible for many hours of naturalistic observation. For these reasons, research efforts were invested in asking aviators to respond to questions about their experiences with team situation awareness.

Because research has suggested that experts and novices in many skill areas process information differently, aviators with a wide range of hours in the cockpit were chosen to be interviewed (i.e., with 400-14,000 hours). Additionally, aviators from dissimilar backgrounds (general aviation, the military, and the air carriers) were selected to see if specific background experience would affect pilots' responses. Fifty aviators were interviewed, 15 general aviation pilots, 20 military aviators and 15 air carrier crewmembers. Using questions derived from a synthesis of the situation awareness research (Shrestha, Prince, Baker and Salas, 1995) a standard form was developed and used in each interview. The form contained 25 questions (e.g., "How would you define team situation awareness?", "How do you know when you have 'lost' situation awareness?"). In addition, crewmembers were encouraged to talk freely of events and incidents that they felt related to the questions.

Analysis of the interview responses has shown that there are some differences that do appear to be based on number of hours of experience, some to specific organizational experiences, while others are similar for all aviators. For example, aviators with less than 1000 hours experience cited behaviors that could be classified as being communication and planning behaviors as those that most promote situation awareness. Crewmembers with more than 1000 hours added behaviors that could be classified best as leadership and adaptability behaviors as those they felt were important contributors to the building and maintaining of team situation awareness. The interviewed aviators all believe that they can recognize when a team member has lost situation awareness and each has techniques that they use to help their fellow crewmembers regain it. The cues used differ based on the experience level of the crewmember that the individual is accustomed to assessing. For very inexperienced crewmembers, wide eyes, and fixation on a single instrument are used as indicators that individual has "lost situation awareness." These cues are far more subtle with more experienced aviators and often center on changes in communication patterns. Some of the most experienced aviators report initiating behaviors that ensure the other individual is kept informed of the situation. They notice elements in the situation that may remove the other team member's attention from the whole situation (e.g., trouble-shooting, programming the computer) and provide the necessary information before seeing signs of a loss.

Results from the analysis of the interviews are being used to suggest some useful research questions and hypotheses. They are also being used to expand understanding of some experimental results.

Measurement Approaches

Initial Trials

One measurement tool used for leaming about team situation awareness is the query technique (see Prince, Salas and Stout, this issue). Perhaps the best known example of this technique is the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1989). As Endsley has described its use, SAGAT is combined with a scenario that is flown by the crewmember in a high fidelity simulator. For our experiment (Prince et al., this issue) SAGAT was modified and paired with a scenario flown on a table-top simulator. Forty one crews (two-person crews) flew a scenario, were stopped three times during the flight, and responded individually to the questions they were asked. These questions had been constructed by subject matter experts to correspond with the important situation information required by the scenario. All pilots had almost identical training and the same number of flight hours. Performance of the uneventful scenario, as measured by landing without incident, was equal for all crews. Despite the equality in the gross measure of performance (a safe landing), the range in scores on the questions was wide (from a low of 81 to a high of 192). A communication content analysis revealed differences in communications in both the pre-flight planning and the flight phase between the lowest scoring crews and the highest scoring crews. High scoring crews did more detailed planning that addressed potential weather problems and did more to familiarize themselves with the course of the flight they would be taking. In flight, these crews were more likely to correct the actions of others and to state their own intended actions. Thus, both communication content analysis and the query technique were able to discriminate between the two extreme groups.

Exploratory Experiments

The manner in which teams share the data that is important for other team members to have can be hypothesized to be a critical element in achieving team situation awareness (Salas et al., 1995). In fact, it might be argued that it is this need for communication which distinguishes team situation assessment from a mere collection of individual situation assessments. This is because no team member has direct access to the assessments and resulting individual situation awareness of the other team members. In some cases, common training, practice, and experiences may help each team member to make an educated guess about another member's individual situation awareness. Improving the accuracy of team members' knowledge about the situation assessments of other team members has been the goal of a number of interventions, including the formation of teams with stable membership, the administration of joint practice, and cross-training. At the same time, however, this approach to achieving a higher level of team situation awareness is obviously limited and can be subject to errors. In lieu of this approach, the only other way for a team to ensure that information important for situation assessment is distributed in the team is to communicate information concisely and accurately among its members.

There are a few studies which have investigated the link between intra-team communications, team situation assessment, and their role in building team situation awareness. An early study by Orasanu (1990) demonstrated that higher performing flight crews (i.e., those who exhibited signs of higher levels of situation awareness) made a significantly greater number of statements related to the situation than did poorly performing crews. Subsequently, Orasanu and Fischer (1991) found that the frequency of situation awareness statements distinguished between good and poor crews in at least one of the two aircraft types they investigated. Following on these results, Jentsch, Bowers, Sellin-Wolters, and Salas (1995) analyzed the frequency of situation assessment communications in crews during a normal period of flight. They found that crews who demonstrated a higher number of statements related to situation assessment *prior* to encountering a

problem situation were likely to be faster at completing a decision making task that required good team situation assessments.

That the correlation between counts of situation assessment, communications, and performance measures may not be positive in all cases was shown by Thornton (1992), who analyzed situation assessment behaviors and performance. In a study designed to assess the effects of automation in the cockpit, Thornton counted the situation assessment behaviors of the crews and compared them to subjective performance ratings made by trained raters. She found a positive relationship between situation assessment statements and errors in flight. Thornton posited that poor crews might have employed a high number of situation assessment behaviors to correct previous mistakes, which would account for the positive correlation between communication frequency and errors. These results suggest that there is still a need to identify the degree and the circumstances required for intra-team communications to contribute positively to situation assessment. They also suggest the need to determine what communication strategies are needed to increase team situation awareness.

Several other behaviors, conceptualized to be associated with team situation awareness, have been shown to be related to performance in laboratory simulations, also. Jentsch, Bowers, Bowen, Nadal, Secrease, and Sellin-Wolters (1995), for example, found that the number of statements directly pertinent to the situation (situation assessment communications), the use of standard phraseology, and the number of commands were related to subjective ratings of performance in decision making tasks. In addition, research by Bowers and his colleagues suggested that crews who performed well in a simulation that required a high level of situation awareness used what is known as closed-loop communications (i.e., question-answer or command-acknowledgment sequences) more frequently than less successful teams (Bowers, Jentsch, Salas, and Braun, 1995).

Finally, an experiment that included both an attempt to measure mental models and situation awareness was conducted. This experiment was designed to investigate the relationship between the mental models of the crewmembers (as measured by Pathfinder; Schvaneveldt, 1990), planning behaviors, and situation awareness (as measured by crew communication strategies; Stout, 1995). Stout (1995) found that in emergency situations, crewmembers whose mental model measurements were more similar to one another gave information that was needed to the other crewmember in advance of that need more frequently than crewmembers whose measurements suggested that their mental models were more dissimilar. The more similar crews were rated higher in their planning behaviors, also.

In sum, we are hopeful that through a variety of approaches we can begin understanding this complex construct. The interviews have provided the crewmember's perspective on teamwork as it relates to situation assessments and the information from these interviews can be useful in interpreting observed behaviors. Experiments where data have been collected from different aspects (e.g., participant responses, communication frequencies, communication content) are also adding to our knowledge about how the team interaction can affect the situation assessments their team tasks require.

Concluding Remarks

The research and the knowledge about training of team situation awareness are in their infancy. We, in research, have just begun to articulate a definition of team situation awareness and to hypothesize what may affect it. There is much work that remains to be done, yet the future is promising. Much of this promise comes from a recognition that situation awareness in a team setting requires consideration and study on its own.

References

Bowers, C., Jentsch, R., Salas, E., & Braun, C. (1995). *Performance differences among aircrews: Analysis of communication patterns.* Unpublished manuscript.

Brannick, M.T., Prince, A., Prince, C., & Salas, E. (in press). The measurement of team process. *Human Factors.*

Endsley, M. R. (1989, November). *Pilot situation awareness: The challenges for the training community.* Paper presented at the Interservice/Industry Training Systems Conference (I/ITSC). Ft. Worth, TX.

Fracker, M. L. (1991). *Measures of situation awareness: Review and future directions* (Tech. Report AL-TR-1991-0128). Wright-Patterson Air Force Base, OH: Air Force Systems Command.

Jentsch, F., Bowers, C., Bowen, S., Nadal, O., Secrease, C., & Sellin-Wolters, S. (1995). Links between aeronautical decision making and crew coordination (Technical Report for the Naval Air Warfare Center Training Systems Division). Orlando, FL: University of Central Florida.

Jentsch, F. G., Sellin-Wolters, S., Bowers, C. A., & Salas, E. (1995). Crew coordination behaviors as predictors of problem detection and decision making times. *Proceedings of the Human Factors Society 39th Annual Meeting (pp. 1350-1354).* Santa Monica, CA: Human Factors and Ergonomics Society.

Orasanu, J. (1990). *Shared mental models and crew decision making* (Tech. Report 46). Princeton, NJ: Princeton University, Cognitive Sciences Laboratory.

Orasanu, J., & Fischer, U. (1991). Information transfer and shared mental models for decision making. In *Proceedings of the Sixth International Symposium on Aviation Psychology* (pp. 272-277). Columbus, OH: The Ohio State University.

Prince, C., & Salas, E. (1993). Training and research for teamwork in the military aircrew. In E. L. Weiner, B. G. Kanki, and R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 337-366). Orlando, FL: Academic Press.

Prince, C., & Stout, R. J. (1995, April). *Situation awareness from the team perspective.* Paper presented at the 10th International Symposium on Aviation Psychology. Columbus, OH.

Robertson, M. M., & Endsley, M. (1994, March). *The role of crew resource management (CRM) in achieving team situational awareness.* Paper presented at the Western European Association for Aviation Psychology 21st Conference, Dublin, Ireland.

Salas, E., Prince, C., Baker, D. P., & Shrestha, L. (1995). Situation awareness in team performance: Implications for measurement and training. *Human Factors, 37,* 123-136.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization.* Norwood, NJ: Ablex.

Shrestha, L. B., Prince, C., Baker, D. P., & Salas, E. (1995). Understanding situation awareness: Concepts, methods, and training. In W. B. Rouse (Ed.), *Human/technology interaction in complex systems* (vol. 7, pp. 45-83). Greenwich, CT: JAI Press.

Stout, R. J. (1995). Planning effects on communication strategies: A shared mental model perspective. *Proceedings of the Human Factors Society 39th Annual Meeting* (pp. 1278-1282).

Stout, R. J., Cannon-Bowers, J. A., & Salas, E. (1994). The role of shared mental models in developing shared situational awareness. In R.D. Gilson, D. J. Garland, and J. M. Koonce (Eds.), *Situational awareness in complex systems* (pp. 297-304). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Thorton, R. C. (1992). *The effects of automation and task difficulty on crew coordination, workload, and performance.* Unpublished doctoral dissertation, Old Dominion University, Norfolk, VA.

Wellens, (1993). Group situation awareness and distributed decision making from military to civilian applications. In N. J. Castellan, Jr. (Ed.), *Individual and group decision making: Current issues* (pp. 267-291). Hillsdale, NJ: Erlbaum.

Situation Awareness: Team Measures, Training Methods

Carolyn Prince, Eduardo Salas and Renée J. Stout

Naval Air Warfare Center Training Systems Division

Situation awareness for aviation crews has been identified as important in flight (Hartel, Smith, and Prince, 1991). Acknowledgement of the significance of crew situation awareness inevitably presents a challenge to define and describe it. This must be done so that training methods for its improvement can be developed.

On one level, team situation awareness has been characterized as individual situation awareness interwoven with teamwork (Salas, Prince, Baker and Shrestha, 1995). This would appear to place it almost beyond our present ability to comprehend fully, since its two major components are so little understood. Situation awareness from the individual standpoint is a complex construct that has not been thoroughly researched and its important elements are not completely known. Research on team processes, or how teams are able to function effectively, is still in progress. However, we believe advancement in the understanding of the concept of team situation awareness can be made by adopting a systematic, theoretically-based research program that begins to examine team situation awareness through its multiple components.

The research on team situation awareness that we and our colleagues are undertaking has the goal of determining the training tools, methods, and strategies that will raise the level of situation awareness that can be achieved by aviation teams. A first step toward this goal is to begin to define the concept. This requires a research program designed to approach the study of team situation awareness from several different avenues; a research program that leads from definition through measurement to training. Since measurement contributes to the understanding and articulation of a concept and is a valuable aid both to providing training feedback and to evaluating training programs (Brannick, Prince, Prince and Salas, in press), it is a significant part of our research.

To start, we developed a framework for team situation awareness based on the synthesis of literature reviewed and a series of critical incident interviews (Salas et al., 1995). In general, we argue that team situation awareness is comprised, first, of individual situation awareness, although it is not a simple sum of each crewmember's situation awareness. It is also made up of team processes, that include specific behaviors and actions of the crews that help in building and maintaining situation awareness. Finally, team situation awareness is affected by the team's context. After developing the framework, our next step was to conduct a series of exploratory experiments to test several measurement methods and tools as possible contributors to team situation awareness training. The first of these methods was the use of a table-top trainer for eliciting and studying situation awareness. This system, consisting of a central processing unit, three monitors, two joysticks and an off-the-shelf software program, had been used previously for team skill training and evaluation (Brannick et al, in press). Given the limited cues that this system can provide, an important question was to determine if it would be possible to discriminate among crews on their team situation awareness. Another question about the system's use to study situation awareness was its applicability to pilots with low experience levels who might benefit most from such a system. Because a scenario with complex mission demands is not relevant for users with a low experience level and also because the system itself cannot present many cues, we needed to determine if a relatively simple scenario on this system could present

enough elements important to team situation awareness to make it useful for training. The first experiment included two different measurement tools, query and communication content analysis. This was done to begin to assess the contribution each might make to team situation awareness training research. By using both tools, we had a unique, but limited, opportunity to compare the information supplied by each.

Method

Subjects

Eighty-two pilots agreed to take part in this exploratory experiment. All had recently completed the requirements for undergraduate flight training, including both basic flight and instrument training. This research was positioned at the point when pilots are moving from the role of student, where their situation assessment is always backed up by the instructor's, to the position of crew member with full responsibility for situation assessment.

Measures

By characterizing team situation awareness as being composed of individual situation awareness and team processes, it necessitated adopting measures that would reveal something about the individual situation awareness of each crewmember and something about the team process behaviors that they were exhibiting in the scenario. For individual situation awareness, we chose to use a query technique modeled on the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1989). The query technique is used in conjunction with a realistic scenario. For this technique, the scenario is interrupted so that its participants can be asked questions about the scenario's situations (e.g., current system status, what this means for the immediate future). The query technique provides an opportunity to test individual knowledge about aspects of the situation. For team process behaviors related to situation awareness, the capability of videotaping the scenarios made communication content analysis possible. An observation form was developed, using the specific team process behaviors that had been identified in previous research as relating to team situation awareness (Prince and Salas, 1993). Specifically, these were: "commented on deviations", "provided information in advance", "verbalized a course of action," "demonstrated awareness of the performance of self and others," "identified problems/potential problems" and "demonstrated awareness of the mission status". To ensure our ability to capture these team process behaviors, we designed scenarios with probes or events to elicit behaviors related to team situation assessment and awareness that could be readily identified. These events were determined through a coordination demands analysis where certain crew behaviors related to situation awareness had been identified as necessary for training for the level of aviator who participated in the experiment. Thus, for example, traffic that was a possible threat to the safety of the subject's airplane was included in the scenario, because recognizing and discussing possible problems with traffic was identified as a training need.

The scenario to support the measurement techniques was developed by three experienced aviators and an aviation team training expert. It was scripted for the air traffic control calls and other communications that the pilots could hear. The scenario's flight started at a regional airport in central Florida and ended at another airport on the Florida coast. Flying time was less than 30 minutes. There was traffic in the area and a threat of summer thunderstorms along the flight path. No aircraft emergencies were included. Using information from a report by Endsley (1989), the aviators who designed the scenario developed three different question sets to be given to the crews for the query technique. Because the system used for the research was not able to present the

questions to the pilots, these questions were written on paper and were given to the pilots by the experimenter. For standardization of the questioning, the questions were presented at three pre-defined points in the scenario. Questions covered the three levels of awareness (perception, understanding, and projection) that are part of Endsley's (1994) definition of individual situation awareness. A group of four experienced aviators who had not been involved in the creation of the scenarios as well as three researchers and two aviators who were involved in the research, reviewed the scenario and confirmed that the questions were relevant to the situation awareness of the team. They also categorized each question as Level I (perception), II (understanding), or III (projection into the future).

Procedure

The crews received training on the table top system, were provided with information on the flight, were given time to plan and brief, and flew the scenario. The pilots were videotaped during their planning session, their brief, and throughout the flight. The scenario was stopped three times at pre-designated points (for all flights), the instruments were blanked and pilots were each given a paper copy of the questions that had been written to pertain to the preceding portion of the scenario. Pilots worked individually on the written answers to the questions. After two minutes, the papers were collected and the scenario was resumed. All pilots were given the same questions, regardless of individual tasking (piloting or navigating). At least one experimenter was with the crews while they were answering questions to ensure that there was no collaboration. After they completed the scenario, the pilots were debriefed on the experiment and were asked for their opinions and reactions to the training device and the scenario.

Data Analysis

Questions used in the query technique were graded according to a pre-determined answer set, except when there had been some important deviations from the plan.
In these cases, the answers were recorded by the researcher. For example, although all crews should have been at a pre-specified altitude at their first stop, some crews were off-altitude. Since they were asked to give their present altitude, the researcher recorded any deviation from the expected altitude. If there were any other non-standard occurrences in the scenarios, these were noted and taken into account. Scores for each individual were calculated for each question set. Individual total scores were a simple sum of the three question set scores. Team total scores were calculated by summing the individual scores of the two team members.

In order to look at the team processes, transcription of the scenarios was required. Because this was an exploratory experiment and communication transcription is a very time consuming effort, our purpose was not to do an analysis of all crews' communications, but to look at sufficient data to suggest the areas of interest for future experiments. The teams of the high scorers and low scorers on the query technique (those at the extremes of the distribution) were selected to be analyzed for their communications. First, the communications of the "extreme" crews were transcribed. Then, one researcher with experience in team research, but with no knowledge of the results of this experiment, subjectively judged the situation awareness classification (high or low) of each of the teams by reading the transcriptions. Next, transcribed behaviors that related to those on the observation form were listed on that form for each of the teams. Completed forms on the teams were then categorized into high and low scorers by a second researcher (who was also unfamiliar with the experiment), based on subjective judgement that took into account the frequency and type of behaviors.

Results

All 41 teams successfully completed the flight. They flew the scenario without incident and landed safely.

Individual scores on the query instrument for all the pilots who participated in the experiment ranged from a low of 81 to a high of 192. Scores were normally distributed. The mean was 146, with a standard deviation of 25. Team scores ranged from 202 to 357, with a mean of 191.2 and a standard deviation of 20. Three teams were more than two standard deviations below the mean, one team was more than two standard deviations above the mean and two were close to two standard deviations above the mean.

Communication content analysis was completed on the three lowest scoring teams and the three highest scoring teams. This analysis demonstrated differences in the crews in the frequency and timing of their identification of problems and potential problems. Higher scoring crews not only identified more potential problems but they were able to do so sooner in the scenario than their lower scoring counterparts. There was also a difference in their awareness of the task performance of themselves and others (evidenced primarily by error correction in the high scoring crews and the failure to correct errors in the low scoring crews). Higher scoring crews also clearly verbalized courses of action, whereas the three lower scoring crews rarely did so. Categorization of the entire transcripts and of the completed forms made by two researchers based on their subjective assessments agreed with the categorization defined by the query technique. That is, each crew who demonstrated behaviors related to good team situation awareness was a crew that scored high on the query technique. Those crews that demonstrated fewer of the behaviors believed to indicate good team situation awareness were crews that had scored low on the query technique.

Discussion

Since all crews successfully completed their flight, landing without mishap, the traditional outcome measure of performance for this uneventful scenario would have suggested that there was no significant difference among the crews. However, the two measures of situation awareness used in this exploratory experiment clearly demonstrated that the crews did have levels of situation awareness that were not the same. That is, both individual crew members' knowledge about situation events as revealed in their answers to questions about those events and behaviors of crews in the scenarios distinguished the highest teams from the lowest. That these two dissimilar measures agreed is important.

In the low scoring teams, for identifying problems and potential problems, alternate fields were identified in the brief but there was no discussion about the fields that were selected, their navigation aids or their available runways. All of the highest scoring teams discussed both the implications of the weather for their flight and clarified with one another whether traffic in the area was a threat. Two of the lowest scoring teams discussed weather as being a problem, but never mentioned the traffic; the third low scoring team discussed the possibility of one of the planes in area being a factor to consider, but never mentioned the weather. The high scoring teams were not error-free, but the pilots corrected one another, indicating that they were aware of the performance of one another. Finally, a clear difference was seen in these six teams in their verbalization of a course of action. In the case of all three high scoring crews, the crewmembers were clear in their intention about actions that they expected to take place, whether it was their own ("I'm going to request..."), the other crewmember's ("Call for a weather update, now") or the crew's ("We're going to divert").

Lessons Learned

The implementation of the query method was easy. The experimenter who was role-playing air traffic control and providing the voices of the other airplane pilots in the crew's area was able to run the simulator and give out and collect the question sets. He was also able to make notes about anything that was unusual in the session, so that would be considered when grading the question sets. Although we found implementation to be rather straightforward, it does require considerable effort beforehand, particularly in the selection of questions to be asked. We also found that when time for asking questions is limited, it is important to ensure that each question will provide useful information. A pilot test of questions is useful because, just as with any test construction, questions must be unambiguous or the results will be meaningless.

Comments made by some of the pilots during the debrief indicated that the query technique may have affected their performance subsequent to the questions. For example, several pilots remarked that because listening to traffic calls was usually handled by their instructors, they did not realize how much more they needed to attend to until after the first question set. This suggests that the query technique has the potential for affecting situation awareness. It also suggests that there is a need to work on developing an on-line assessment tool for team situation awareness which does not require the possible provision of information to the crews.

Communication content analysis is a much slower process than the query technique to arrive at comparisons of teams because it requires transcription of the scenarios and analysis of the results. However, it does provide specific information on how the crews differ in behaviors relating to situation awareness and situation assessment that is not available with the query technique.

Crews did not object to the table-top system and gave as their reason for finding it acceptable its ability to support a realistic scenario. All pilots were able to learn its instruments and could control it with less than fifteen minutes training.

Conclusions

Since communication content analysis and the query technique placed the crews that were in the extremes in the same categories, the results indicate that they may be tapping some of the same construct (or related constructs). Both techniques have strengths and weakness and need to be examined for their unique contributions to our information about team situation awareness.

As stated at the outset, the goal of this research was to contribute to the knowledge about and the development of team situation awareness training. During the debrief session with the pilots who had participated in the research, many of them made comments about the learning potential that they felt the query technique, combined with a simulator scenario, had. Some remarked that the questions made them aware of some of the environmental elements in flight. Other pilots suggested that providing the answers to the questions would have been useful to them. A follow-up to this exploratory work is being planned to extend the query technique to the specific purpose of training.

References

Brannick, M. T., Prince, A., Prince, C., & Salas, E. (in press). The measurement of team process. *Human Factors*.

Endsley, M. R. (1994). Situation awareness in dynamic human decision making: Measurement. In R. D. Gilson, D. J. Garland, and J. M. Koonce (Eds.), *Situational awareness in complex systems* (pp. 79-100). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Hartel, C. E. J., Smith, K., & Prince, C. (1991). *Defining aircrew coordination: Searching mishaps for meaning*. Paper presented at the Sixth International Symposium on Aviation Psychology, Columbus, OH.

Prince, C., & Salas, E. (1993). Training and research for teamwork in the military aircrew. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 337-366). San Diego, CA: Academic Press.

Salas, E., Prince, C., Baker, D. P., & Shrestha, L. (1995). Situation awareness in team performance: Implications for measurement and training. *Human Factors, 37*, 123-136.

Psychophysiological Assessment of SA?

Glenn F. Wilson

Wright-Patterson Air Force Base

Psychophysiology measurement provides an assessment of the relationship between human performance and correlated changes in the operators physiology. Cognitive activity is known to be associated with changes in various physiological systems. These changes are seen as separate from the purely physiological adjustments related to the physical environment such as changes in temperature, the physical demands of the task and G forces. Psychophysiology relates the performance (cognitive) demands of a task to the correspondent changes in the persons physiology. Changes in brain activity in response to the demands of different tasks or different difficulty levels of the same task seem straight forward to understand. The brain is responsible for taking in sensory information, processing that information and initiating responses based upon these processes. Additionally, peripheral systems are also known to exhibit changes that are related to cognitive activity and include eye blinks, heart rate and respiration (For reviews see; Caldwell, et al., 1995; Wilson & Eggemeier, 1991).

Traditionally, psychophysiological measures have been used as metrics in mental workload research but have been used rarely in investigations of situational awareness (SA) (Vidulich, Dominguez, Vogel & McMillan, 1994). Their utility has been discussed. For example, Endsley (1995) did not feel that they would be useful to provide information about an operators state of knowledge. However, there may be settings in which psychophysiological measures can provide information relevant to determining if an operator is or is not aware of certain types of situations and whether or not the operator is actively seeking information. There is only one known study that has actually used psychophysiological measures to investigate SA (Vidulich, Stratton, Crabtree & Wilson, 1994). In this study EEG, eye blinks and heart rate were used to see if these measures could provide information about SA. They found that the EEG theta band increased in the most difficult conditions of their air-to-ground flight simulation task but the eye blink and heart rate measures were not sensitive to the different conditions. They concluded that EEG shows promise as an indirect measure of SA and that the use of these measures should be explored further.

Psychophysiological measures have unique properties that should make them attractive to investigators in the SA field. Some SA measures, such as the query techniques, require that ongoing task performance be stopped while the operator is interrogated about their level of SA. This interference with the primary task performance is problematic and is limited to situations where it is possible to stop the flow of the task. These procedures are not possible during flight for example. On the other hand, psychophysiological measures can be unobtrusively obtained without interfering with performance. A second beneficial feature to the SA researcher is that, in contrast to the discrete nature of some SA measures such as subjective reports, psychophysiological measures are continuously available. The continuous nature of psychophysiological measures can be especially useful in situations where the timing of critical events can not be precisely controlled, or when events of interest are unplanned or uncontrolled. Since the events are unpredictable, the SA associated with these novel or unexpected events would be missed by many of the more standard SA measurement techniques. Because the physiological data are continuously recorded it is possible to go back and assess the situation as it existed when these events occurred and it is also possible to examine the antecedent conditions.

As previously stated, psychophysiological measures have not been used in SA research to any great extent. However, they have been used in numerous mental workload studies. Because of the postulated overlap of mental workload with SA in some situations it is worth examining this data (Endsley, 1995; Taylor, 1989). Several examples will be presented from workload studies that bear upon the study of SA. It is hoped these examples will be of sufficient interest to SA investigators to inspire them to consider using psychophysiological measures in there future work.

The first example is one in which a flight maneuver that requires high levels of SA also produces high levels of mental workload. During a Low Altitude Parachute Extraction (LAPES) a C-130 flew approximately 10 feet above the ground and delivered several tons of material. With the aircrafts rear ramp open, a parachute was deployed in order to extract several tons of material. This maneuver required the pilot to be acutely aware of the situation of the aircraft so that the cargo would be accurately dropped and the aircraft would gain altitude after the drop. With only 10 feet between the aircraft and the ground it is crucial that the pilot be situationally aware. Eye blinks and heart rate were recorded and the data demonstrated that this highly demanding procedure resulted in unique patterns of physiological responding. Eye blinks are typically irregularly spaced with varying eye lid closure durations as seen in the right side of the blink portion of figure 1. The LAPES maneuver produced a quite remarkable pattern of eye blink activity. Prior to the LAPES, an unusual, slow rhythmic pattern of blinking is seen that had remarkably constant closure durations. During the LAPES an inhibition of blinking is seen which is characteristic of very high visually demanding situations. Because blinks block visual input, operators tend to reduce blink rates and shorten the durations of the eye closures during periods of high visual demand. This as a time when the pilot must have very high SA because not having high SA would lead to disaster. In order to maintain the required high levels of SA the pilot exhibits uncharacteristic eye blink patterns during the crucial portions of the maneuver. Heart interbeat intervals, on the other hand, showed a fairly continuous decline that reached the minimum level following the actual LAPES. Heart interbeat intervals are also shown in figure 1 (note that interbeat intervals decrease as heart rates increase). The cardiac system is mediated by slower acting mechanisms and generally is viewed as a measure of the overall level of task involvement. This explains why it is less time locked to the actual LAPES events and why it reaches its minimum response later than the eye blinks which are more closely tied to the demands of the task. This pattern was seen in the same pilot during several LAPES and also in data from other pilots.. These results show that psychophysiological data can be used to determine situations when high SA is required and whether or not the operator is in the proper state.

A second example involved a very low probability event that required high levels of SA. During a study of A7 pilot workload pilots experienced emergency situations (Wilson, Skelly & Purvis, 1989). The study engaged A7 pilots in three situations, flying lead in a four ship formation, flying wing in a four ship formation and flying in a simulator. During one flight there was a bird strike and in a second flight one of the pilots perceived that the lead aircraft was about to turn in front of him. The flight emergencies required very high levels of SA in possibly life threatening situations. In order to correctly react and avoid catastrophe the pilots had to quickly assess the situation, make decisions and react accordingly. Figure 2 shows heart rate and interbeat interval data. The top curve represents the mean heart rate every ten seconds while the bottom curve shows the continuous registration of interbeat intervals during the two minutes surrounding the event. Note the approximately 50% increase in heart rate within 30 seconds following the strike and the return to the pre-bird strike level within about 60 seconds of reaching the peak. The heart rate changes are correlated with the pilots need to acquire awareness of the situation. In the second incident the pilots pre-event heart rate was approximately 15 beats per minute slower than that of the first pilot, however, there still was an approximately 50% increase in his heart rate over a 30 second period. Both of these cases show heart rate increases during the time that the pilot is acquiring SA to a novel and dangerous event and how successful resolution of the problem is associated with a return of heart rate to the preceding level. These dramatic increases in heart rate could be entirely due to the stress of the situation. However, even if this is the case the data are still useful for determining the occurrence of events and evaluating their time course from onset to resolution.

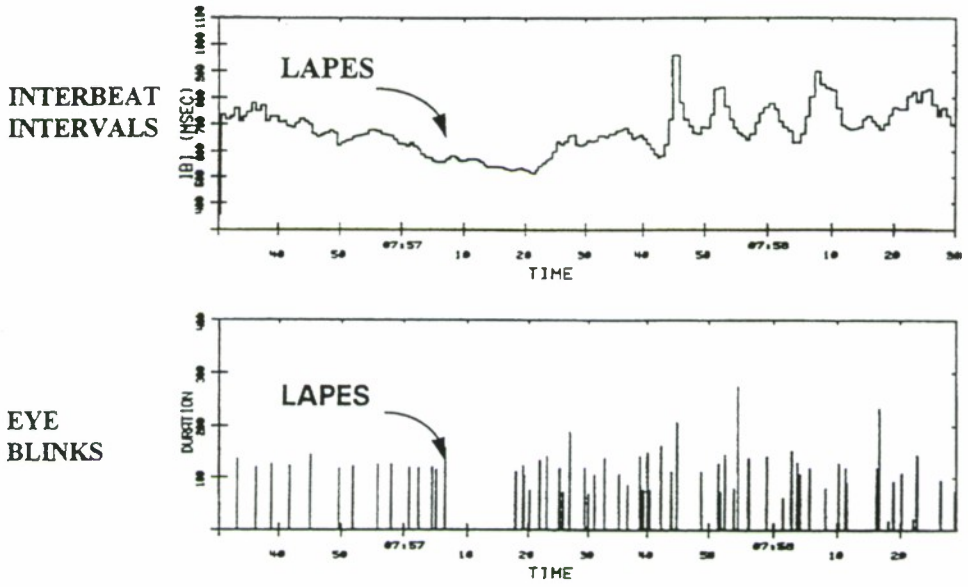


Figure 1. Heart interbeat intervals and eye blink intervals and durations during two minutes of a LAPES maneuver. The arrow indicates the time of load release.

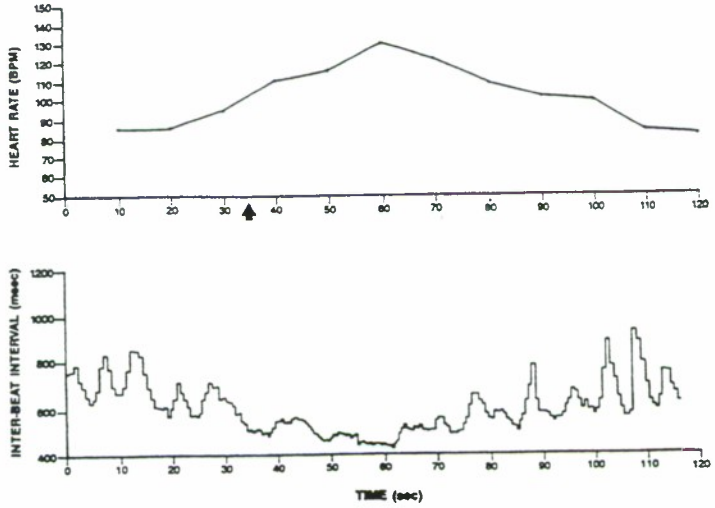


Figure 2. Ten second heart rate means (top) and interbeat intervals during a bird strike (bottom). The arrow indicates the time of the bird strike.

A third flight study demonstrates how psychophysiological measures can show divergence between mental workload and SA. In this study fifteen general aviation pilots flew scenarios under both VFR and IFR conditions (Wilson & Hankins, 1994). The theta band of the EEG demonstrated peaks of power during segments of flight which required higher levels of mental activity compared to segments which required high levels of psycho-motor activity. Relative power of the EEG theta band (3-7 Hz) during the cognitive demanding segments showed higher levels than the segments associated with the three landing segments. Heart rate, on the other hand, increased during the take-off and landing phases of the scenarios and was lower during the more mentally demanding segments. Take off and landing are usually the most dangerous parts of general aviation flight and require high levels of SA because of the large number of factors that must be considered and the potential for disaster. Brain waves (EEG) showed changes related to the mental workload but did not seem to be as related to the pilots need for SA while their heart rates seem to show increases during the segments when it was especially important to have high SA.

A final example is from an experiment designed to measure the mental workload of air traffic controllers in a simulation (Brookings, Wilson & Swain, in press). Task difficulty was manipulated by varying the number of aircraft to be handled in fifteen minutes, by modifying the complexity of the traffic to be handled and in an overload condition which was designed to cause the controllers to lose the picture. These manipulations resulted in significant decreases in performance and increases in subjective estimates of mental workload with the overload condition at the extreme in both cases. Eye blink rate decreased from low to high while the overload condition was associated with the lowest blink rates. Respiration rate increased with task difficulty with the highest rates found during the overload condition. EEG theta band power significantly increased during higher difficulty conditions. Conversely, EEG alpha band activity decreased as the task became more difficult and this band was also sensitive to the nature of the workload manipulation. These data strongly suggest that as the task becomes more difficult and it is harder to maintain good SA there are associated changes in the psychophysiological data. The overload condition which was designed to cause the controllers to lose the picture, which happened in only one of the eight controllers, was associated with physiological response levels that were at the extreme end of a continuum from easy to most difficult. The overload condition did not produce unique data but rather continued the pattern found in the low to high difficulty conditions. One could interpret these results as evidence that psychophysiological measures demonstrate significant changes associated with the maintenance of good SA as the task becomes more difficult.

The above examples were chosen to show that in some circumstances changes in psychophysiological measures parallel the need for SA. While there are other situations in which mental workload and SA requirements are not parallel (Endsley, 1993), the above examples suggest that psychophysiological data may be useful for assessing SA in certain situations. In some cases the lack of psychophysiological changes could be indicative of a lack of good SA because they suggest that the operator is not aware of problems. These metrics may also be used to measure SA when comparing situations having different parameters. This would include different displays or work procedures when selecting the design that will result in the highest SA. This type of test and evaluation would benefit from inclusion of psychophysiological measures.

As with all endeavors which undertake to measure human operator state, a battery of measures should be used. Most systems that humans operate are quite complex and involve several aspects of the humans cognitive capabilities. In this context it is suggested that psychophysiological measures be used as adjunct metrics to performance and subjective measures of SA. Because most metrics can be global or specific it is wise to include complementary measures to provide as complete a picture as possible. This leads to a battery of combined measures that complement one another and are selected to be appropriate for the questions being asked within the constraints of a particular situation. It seems wise and prudent to use as many measures as possible to assess as many aspects of the demands placed on the operator.

Based on current evidence, psychophysiological measures should be considered as tools to study and measure SA. Research to determine the utility of these measures in the SA context and to determine the nature of the information obtained using them is warranted. Hopefully the

examples discussed in this paper will lead workers in the field to consider including psychophysiological measures to determine their place in the tool box of SA metrics. Psychophysiological recording equipment is available that produces high quality data in laboratory, simulator and real world settings. The equipment is small and portable. Thus, it can be worn by system operators at the work place even if the work place is an airplane, automobile or control room.

References

Brookings, J. B., Wilson, G. F., Swain, C. R. (in press). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology.*

Caldwell, J. A., Wilson, G. F., Centiguc, M., Gaillard, A.W. K., Gundel, A., Lagarde, D., Makeig, S., Myhre, G., & Wright, N. A. (1994). *Psychophysiological Assessment Methods.* AGARD Advisor Report, AGARD-AR-324, AGARD, Paris, France.

Endsley, M. R. (1993). Situation awareness and workload: Flip sides of the same coin. In R. S. Jensen and D. Neumeister (Eds.), *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 908-911). Columbus, OH: Ohio State University.

Endsley, M. R. (1993). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37, 32-64.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems, *Human Factors*, 37, 65-84.

Taylor, R. M. (1989). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Situational Awareness in Aerospace Operations* (AGARD-CP-478, pp. 3/1-3/17). Copenhagen, Denmark: NATO-AGARD.

Vidulich, M., Dominguez, C., Vogel, E., & McMillan, G. (1994). Situation awareness: Papers and annotated bibliography. Technical Report AL/CF-TR-1994-0085. Wright-Patterson AFB, OH: Armstrong Laboratory.

Vidulich, M. A., Stratton, M., Crabtree, M., & Wilson, G. (1994). Performance-based and physiological measures of situational awareness. *Aviation, Space and Environmental Medicine,* 65, A7 - A12.

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

Wilson, G. F. & Hankins, T. (1944). EEG and subjective measures of private pilot workload. *In the Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting* (pp. 1322-1325). Santa Monica, CA: Human Factors and Ergonomics Society.

Wilson, G. F., Skelly, J. & Purvis, B. (1989). Reactions to emergency situations in actual and simulated flight. In *Human Behaviour in High Stress Situations in Aerospace Operations* (AGARD-CP-458, pp. 9/1-9/15). The Hague, The Netherlands: NATO-AGARD.

Role of Volitional Effort in the Application of Psychophysiological Measures to Situation Awareness

Evan A. Byrne

Catholic University of America

Introduction

Situation awareness (SA) has been offered as a mediational construct that may augment our understanding of human performance in complex systems (Endsley, 1995a). The acquisition and maintenance of SA, or its collective components, may be essential in automated environments (Gilson, 1995; Sarter & Woods, 1995). Despite lack of a consensus definition of this construct, SA is said to encompass traditional information processing elements, environmental factors, and individual differences. Endsley (1995a), specifies three fundamental components of SA: perception, understanding, and projection of future trends. An alternate view offered by Smith and Hancock (1995), states SA is the ability to direct one's consciousness to perform a task satisfactorily.

Recent emphasis on SA may represent a gradual paradigm shift in human factors in placing greater emphasis on individual differences and motivational factors in the characterization of human-machine interactions (e.g., Andre & Hancock, 1995; Hart 1989). Most definitions of SA suggest it represents information processes beyond detecting a stimulus and initiating a response; and most incorporate some quality of "activity" or "intent" of the operator in the acquisition and maintenance of the base elements of the phenomenon. This transactional concept overtly considers operator volitional characteristics in the determination of performance to a greater extent than does mental workload.

If SA can effectively allocate previously unexplained variance in the prediction of human performance, the development of measurement tools should be of paramount concern. Several methods have been proposed to measure this concept including subjective ratings, explicit performance, and implicit performance measures (see Endsley, 1995b). However, the utility of psychophysiological measures in SA research has not been evaluated (c.f., Crawford et al., 1995); and their potential for evaluating SA as a state of knowledge has been dismissed (Endsley, 1995b). But, if a broader view of SA (e.g., Flach, 1995) is adopted, what role do psychophysiological measures have in this area? As in the study of mental workload and effort, the application of psychophysiological measures to the SA problem may augment existing measurement strategies, and provide unique information.

Psychophysiology and Mental Workload

Mental workload can be viewed as a multidimensional construct capturing the difficulty that a task presents to an individual (Andre & Hancock, 1995). Comprehensive definitions (e.g., Gopher & Donchin, 1986) acknowledge that both individual and task factors contribute to mental workload. However, many empirical studies neglect the individual's contribution to mental workload.

Psychophysiological measures of mental workload have been applied in a variety of environments (see Kramer, 1991). The most widely used measures include electrical activity from the scalp (EEG, ERP), cardiorespiratory activity (HR, HRV, respiration), eye movements and blinks. Despite costs associated with the collection and analysis of these measures, they are considered advantageous because they can provide continuous data. Workload is a dominant construct in human factors psychophysiology. However, past attempts to discover (a) limits in human performance and (b) sensitive and reliable indices of mental workload has led to a narrowing of this concept beyond its original conceptualizations. Although current research is rapidly reversing this trend, the reasons for this narrowing can be ascribed to paradigm bias, workload bias, and conceptual bias.

Paradigm bias

The traditional laboratory study evaluating mental workload systematically varies task load to produce a dose-response curve for a psychophysiological variable. This approach promotes interpretation of physiological changes to changes in task demands. This bias is due to (1) the primary experimental factor of increasing task demand; and (2) the typically short duration of the work periods that drive the subject and thus, decrease the likelihood that volitional strategies can come into play. Recent research illustrates the growing trend toward abandonment of this approach (e.g., Veltman & Gaillard, 1995). However, there is a substantial literature base generated using the short duration "stress-test" approach. The potential problems in generalization beyond the simple task environments and methodology used in this approach should be considered.

Workload bias

Considerable focus for psychophysiological research on mental workload has been on the upper end of the workload continuum. Few studies, by comparison, have been conducted to evaluate the role of psychophysiological measures during underload (c.f., Braby, Harris, & Muir, 1993). The workload bias is gradually being undone with increasing research on the effects of automation on human performance and the use of longer duration tasks. Additional focus on workload transitions (e.g., Ryan, 1994), has also served to draw attention to the lower end of the workload continuum.

Conceptual bias

The interactive (task-operator) nature of mental workload has been emphasized in reviews (e.g., Gopher & Donchin, 1986). However, operational definitions of mental workload are often tied almost exclusively to task demands. Operator contributions are shed from these working definitions or ascribed to the broad term, mental effort. This operational limiting of mental workload to refer to task effects is often a by-product of paradigms that do not allow individual motivational strategies to express themselves or be assessed. As with the other biases, the conceptual bias is slowly being remedied with more complex paradigms and overt recognition of the effects of individual differences in mental effort on workload (e.g., Veltman et al., 1995).

There are clear indications that psychophysiological research on mental workload is returning to more balanced interpretations and investigations of mental workload that include individual factors. This trend parallels the calls in human factors research to move beyond load and start to comprehensively address the nature of work and how operators actively engage in the task (e.g., Andre & Hancock, 1995; Hart, 1989). The impetus may be the inability to adequately predict human performance based on task demands alone.

Psychophysiology and Mental Effort

Calls to re-evaluate the importance of mental effort in the study of mental workload are increasing (e.g., Byrne & Parasuraman, 1995; Gaillard & Wientjes, 1994). In contrast to workload, mental effort can be described as the energy, involvement, and motivation that the individual applies to the task. However, separation of these components is rare and most conceptualizations of effort emphasize how hard subjects must work in contrast to how hard they want to work.

There are two fundamental types of effort, cognitive effort and compensatory effort (Mulder, 1986). Cognitive effort is the type of effort most often associated with resource or capacity theories of information processing (e.g., Kahneman, 1973). It is viewed as a direct consequence (or requisite component) of the performance of a task and not considered under operator control. In contrast, compensatory or motivational effort is considered to be under operator control and is hypothesized to be a greater determinant of performance under conditions of monotony or underload. For some psychophysiological measures a primary question is: are they measuring how hard the operator has to work, or how hard the operator wants to work?

For example, measures of heart rate variability (HRV) have found wide application in aviation research as indices of workload, stress, and effort (e.g., Wilson, 1993). Although HRV is generally recognized to index cognitive effort, it may at times reflect energetic processes (i.e., compensatory effort) depending on the particular task environment. Byrne (1993) examined HRV in 42 subjects ages 18-30 detecting critical events (e.g., loss of transponder for an aircraft target, two aircraft at the same altitude, flight path drift) in a simulated semi-automated air-traffic control (ATC) environment for 42 minutes. Individual differences in the way subjects approached the task produced different profiles of HRV response. Subjects reporting high subjective effort showed significant suppression of HRV from baseline to task performance while low effort subjects showed a linear increase in HRV. Subjects showing initial suppression in HRV also showed faster reaction times to the critical events. Thus, it appears HRV indexed volitional or compensatory effort in this task environment that places subjects in the role of a passive monitor. In subsequent studies, involving a multi-task environment requiring subjects to monitor for infrequent events while performing a compensatory tracking task, we have found only group decreases in HRV in response to task load and no relationship to individual differences in subjective ratings of effort (Byrne et al., 1994); suggesting HRV indexed cognitive effort. The nature of the task requirements between these experiments may be an important determinant: in the monitoring-tracking task, tracking component may continuously engage or "drive" the subjects; in the ATC task, the level of engagement was subject controlled.

The interpretation of psychophysiological measures of workload is difficult when applied to issues of underload. Energetic aspects such as motivation and effort may confound with task aspects. In an underload condition, the task characteristics (short term memory demands, time pressure) which elicit cognitive effort may not predominate and psychophysiological profiles may be more dependent on variations in compensatory effort. Moreover, other energetic factors associated with stress and coping may come into play during underload conditions in contrast to overload conditions. With HRV, under what task conditions do changes reflect (a) how hard subjects *want* to work (i.e., compensatory effort) or (b) how hard subjects <u>have</u> to work (i.e., cognitive effort)? These questions need reconciliation: if decreases in HRV are interpreted as overload, a change in the environment to alleviate this condition is desirable; but, if they reflect appropriate effort invested in the task or engagement, such a change may be counterproductive. It has been recognized that some psychophysiological measures may index task difficulty while others index compensatory effort, and still others index both task difficulty and compensatory effort (Mulder, 1986). Measures of compensatory effort are of direct relevance to research on SA which seems to require "active intent" from the subject.

A Contextual Model for the Duality of Effort

A contextual model is proposed as an aid in classifying past research, structuring new research, and serve as a heuristic illustrating how the interpretation of psychophysiological measures of workload may be strengthened in light of the duality of mental effort. The model accounts for studies using highly structured tasks across the workload continuum where psychophysiological responses are driven by cognitive effort or how hard the subject must work. The model proposes a key factor or dimension affecting interpretation of psychophysiological responses during a task having low response-structure demands and low workload conditions. Under these conditions, compensatory effort, or how hard the subjects want to work may be indexed by psychophysiological measures and may be the dominant factor in performance. This model (see Figure 1), posits two non-orthogonal dimensions. The first labeled "task demands", encompasses the demands of the task per unit time and level of complexity of the task environment. The second labeled "structure", encompasses the temporal response contract between the operator and the task.

According to the model, for a given level of workload, there can be varying amounts of structure placed on the transactional relationship between the task and the operator. At one extreme, with high structure (low flexibility, reactive engagement), the nature and timing of the operator's response is driven by the task environment; and at the other extreme, a task having low structure (high flexibility, proactive engagement), the transactions between the task and the operator are operator-determined. Naturally, as workload increases, the proposed "structure" dimension contracts to accommodate a forced decrease in flexibility in the transactional relationship or manner in which the task can be completed.

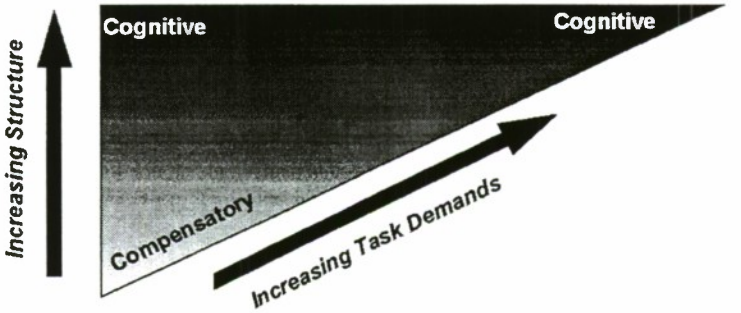


Figure 1. Contextual model for interpreting psychophysiological measures of effort.

This model initiates the task of separating the concepts of effort in psychophysiological measures that have potential to index both compensatory and cognitive effort. Across the top of this model, with highly structured response conditions, psychophysiological changes are thought

to be due to cognitive or "have to work" effort across the workload continuum. However, as the temporal response structure is relaxed, individual contributions, or compensatory "want to work" effort, becomes an increasingly important source of variance. Thus, according to this model, the interpretation of psychophysiological measures under lower workload conditions becomes dependent on the depiction of the response "contract" between the operator and the task.

While not an exhaustive depiction of the factors associated with mental effort and psychophysiological interpretation, this model is valuable as a heuristic. The model begins to reconcile the apparent paradox accompanying some psychophysiological measures that can index both cognitive and volitional effort. It also illustrates the increasing variance contributed by individual motivational factors that needs to be considered when applying these measures to the problem of underload or transitional workload. Models of mental workload that stress the criticality of the operators approach to active management of a task (e.g., Andre & Hancock, 1995; Hart, 1989) are relevant to this model as volitional or compensatory effort may be more important in this regard. That is, a subject who is maintaining some level of volitional effort when working on a low demand task can be viewed as having to go a shorter distance to achieve the optimal state for performance in transient states of higher workload.

Psychophysiology and Situation Awareness

Psychophysiological measures have been given little credence in their ability to assess SA as a state of knowledge (Endsley, 1995b). However, if it can be accepted that the acquisition and maintenance of this state of knowledge is an active process, psychophysiological measures may be especially useful. Moreover, most models of SA (e.g., Endsley, 1995a; Wickens, 1995) consider workload among the component influences that can promote or degrade SA. Clearly, psychophysiological measures of mental workload should have some import in SA research on this basis alone.

Endsley's (1995a) three-level model of SA can be used as a framework to illustrate the potential application of psychophysiological measures to this concept. To achieve level 1 SA, the operator must perceive the relevant system state variables in their environment. Psychophysiological measures such as evoked potentials may be especially useful in this effort, even if they cannot be specifically locked to the relevant state variables. Outside experimental applications, identifying the complete scope of these variables is difficult and they may vary across individuals. However, using psychophysiological measures as a means to evaluate the probability that stimuli are being attended to in the environment is fruitful. Similarly, the use of broader psychophysiological measures such as continuous EEG or HRV may be helpful in ascertaining whether the operator is "engaged" in the environment and thus, might be expected to be perceiving information (e.g., Pope et al., 1995). The comprehension required in level 2 and the prediction required in level 3 do not readily translate to direct psychophysiological investigation. However, these processes are "active" and measures that address volitional effort may prove useful. Psychophysiological measures of workload may suggest when the ability of the operator to efficiently process these latter stages is likely to become degraded. Last, without level 1 occurring, complete SA is not possible, this is the fundamental application for psychophysiology in this model.

There are other ways psychophysiological measures can contribute to our understanding of SA. The ability to achieve and maintain SA in a complex system may be associated with individual temperamental qualities to focus concentration and inhibit or reject distracting information; and psychophysiological measures have been offered as a potential method to aid in screening for these abilities (e.g., Crawford et al., 1995). Besides screening, these measures may have the potential to aid in regulating a task environment to promote the acquisition and maintenance of SA (e.g., Byrne & Parasuraman, 1995; Pope et al., 1995). For example, using the contextual model, if a given environment can be typified by low workload and low structure, than it may be possible to

use psychophysiological measures to identify whether the operator is involved in the active process of updating their mental model; and if not, changes in the response contract may be initiated to promote a more "reactive" task environment and protect the system.

Psychophysiological measures can be used with other available metrics to provide information about the precursors and component elements of SA. Few would condone their use in isolation to this effort. If SA cannot be measured by performance on a task (e.g., Wickens, 1995), then the tools available must provide the ability to evaluate the degree to which an operator is in a state of actively processing information in the environment. Psychophysiological measures have potential in this regard. Indeed, SA may be another expression of the energetic principles of human information processing (e.g., Hockey, Coles, & Gaillard, 1986).

Conclusion

Can psychophysiological measures yield a comprehensive estimate of SA considering the multi-component definitions of this concept? No, and few measures or methodologies proposed to study this concept can. Can psychophysiological measures yield viable estimates of essential components of SA? Yes, as far as SA can be viewed as an active process, requiring some volitional effort on the part of the operator that can be affected by workload. Research efforts to clarify the relationship among cognitive effort, compensatory or volitional effort, mental workload, and task quality will indirectly provide useful data to the understanding of SA. Continued investigation of SA should examine the potential role of psychophysiological measures in furthering knowledge about this phenomenon. If they can explain unique variance or allow inference to be drawn where it otherwise cannot (e.g., providing continuous data) then the effort associated with the acquisition of psychophysiological measures is worthwhile.

Acknowledgements

Preparation of this paper was made possible by research grant NAG-1-1296 from the National Aeronautics and Space Administration, Langley Research Center, Hampton, VA (Alan Pope, technical monitor) awarded to Raja Parasuraman. Views expressed are those of the author and do not necessarily reflect those of the sponsor agency. Address correspondence to Evan A. Byrne at BYRNE@CUA.EDU.

References

Andre, AD, Hancock, PA (1995). Special issue editorial. *The International Journal of Aviation Psychology, 5*, 1-4.

Braby, C.D., Harris, D., & Muir, H.C. (1993). A psychophysiological approach to the assessment of work underload. *Ergonomics, 36*, 1035-1042.

Byrne, E.A. (1993). *A psychophysiological investigation of individual differences affecting performance during a complex monitoring task in adults.* Doctoral Dissertation, University of Maryland -- College Park.

Byrne, E.A., Chun, K.M., Hilburn, B.G., Molloy, R.J., & Parasuraman, R. (1994). Effect of tracking difficulty on secondary task performance, heart rate variability, and subjective perceptions. *Psychophysiology, 31*, S33. (abstract).

Byrne, E.A., & Parasuraman, R. (1995, in press). Psychophysiology and adaptive automation. *Biological Psychology*.

Crawford, H.J., Knebel, T.F., Vendemia, J.M.C., Kaplan, L., & Ratcliff, B. (1995). EEG activation patterns during tracking and decision-making tasks: differences between low and high sustained attention adults. In R.S. Jensen (Ed.), *Proceedings of the Eighth International Symposium on Aviation Psychology (pp. 22-27)*. Columbus: Ohio State University.

Endsley, M.R. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*, 32-64.

Endsley, M.R (1995b). Measurement of situation awareness in dynamic systems. *Human Factors, 37*, 65-84.

Flach, J.M. (1995). Situation awareness: proceed with caution. *Human Factors, 37*, 149-157.

Gaillard, A.W.K., & Wientjes, C.J.E. (1994). Mental load and work stress as two types of energy mobilization. *Work & Stress, 8*, 141-142.

Gilson, R.D. (1995). Special issue preface. *Human Factors, 37*, 3-4.

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept. In, K.R. Boff, L. Kaufman, J.P. Thomas (Eds.), *Handbook of perception and human performance: Volume 2 Cognitive processes and performance, pp 41-1 -- 41-49*. New York: Wiley.

Hart, S.G. (1989). Crew workload-management strategies: a critical factor in system performance. In R.S. Jensen (Ed.), *Proceedings of the Fifth International Symposium on Aviation Psychology (pp. 22-27)*. Columbus: Ohio State University.

Hockey, G.R.J., Coles, M.G.H., & Gaillard, A.W.K. (1986). Energetical issues in research on human information processing. In G.R.J. Hockey, A.W.K. Gaillard, & M.G.H. Coles (Eds.), *Energetics and human information processing (pp. 3-21)*. Dordrecht: Kluwer.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

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

Mulder, G. (1986). The concept and measurement of mental effort. In G.R.J. Hockey, A.W.K. Gaillard, & M.G.H. Coles (Eds.), *Energetics and human information processing (pp. 175-198)*. Dordrecht: Kluwer.

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

Ryan, T.G. (1994). Human factors issues for resolving adverse effects of human work underload and workload transitions in advanced transportation systems. *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*. 784-788.

Sarter, NB & Woods, DD (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors, 37*, 5-19.

Smith, K., & Hancock, P.A. (1995). Situation awareness is adaptive, externally directed consciousness. *Human Factors, 37*, 137-148.

Veltman, J.A., & Gaillard, A.W.K. (1995, in press). Physiological indices of workload in a simulated flight task. *Biological Psychology*.

Wickens, CD (1995). Situation awareness: impact of automation and display technology. Keynote address, NATO AGARD Aerospace Medical Panel Symposium on Situation awareness, Brussels, Belgium, Apr., 1995.

Wilson, G.F. (1993). Air-to-ground training missions: A psychophysiological workload analysis. *Ergonomics, 36*, 1071-1087.

Physiological Measurement Techniques: What the Heart and Eye Can Tell Us About Aspects of Situational Awareness

J. A. Stern[1], L. Wang[1], and D. Schroeder[2]

[1] Washington University, St Louis, Mo.
[2] FAA/CAMI, Oklahoma City,

Introduction

We will not attempt to define Situational Awareness but like Endsley's definition which suggest that in order for SA to occur one must be aware of the current situation, remember relevant events and, what may be most important, make predictions about the future so that strategies can be formulated and responses made. Making predictions about the future involves expectancies. These can be expectancies with respect to events that may or may not occur as well as expectations about the making of simple or complex responses.

We will limit our presentation to one component of Endsley's definition, namely the issue of operators making predictions about the future. We will refer to such predictions about the future as *expectancies* or *subjective probabilities*. Can we use physiological measures to make inferences about the occurrence of such expectancies? We believe the answer to this question is yes. What physiological measures might be considered ? We will single out a few, one dealing with the use of heart rate, the second with oculometric variables, specifically the occurrence and timing of saccadic eye movements, eyeblinks and changes in pupil diameter.

With respect to heart rate there is a reasonable literature that under a variety of conditions the anticipation of an event requiring some sort of action will lead to cardiac deceleration. If we present a warning signal, followed a few seconds later by an "imperative" signal, one requiring a simple manual response, we find cardiac deceleration during the delay period. This deceleration persists over hundreds of trials of task performance.

We will provide an example from one of our studies demonstrating that this cardiac decelerative response occurs in anticipation not only of having to make a motor response but occurs in anticipation of having to acquire information as well as in anticipation of having to "interrogate information" stored in memory.

The task in which we demonstrated such anticipatory deceleration in human subjects is the Sternberg Memory Task. In this task participants are presented with a set of symbols (letters, numbers, colors, random shapes, etc.) and are required to retain that information for a period of time (usually seconds) before being presented with a symbol which is either a member or not a member of the original set. The latter symbol is referred to as the "test" item, the prior set as the "memory set". Participants are required to make a discriminative response ,i.e., the test item is a member of the memory set. In our laboratory we have added another dimension to the task, namely before presentation of the memory set they are informed about the number of items that will be contained in the set that is about to be presented. This period is referred to as the "cue period".

Participants are presented with a Cue, followed a fixed time later by the Memory set, followed a fixed time later by the Test stimulus which is followed a fixed time later by the next Cue stimulus.

We have studied heart rate under conditions where the interval between stimulus presentation was either 6 or 10 seconds and where the memory set was either "small" or "large". Figure 1 depicts the results from such a study, note that two time intervals were used in this study, 6 and 10 seconds between onset of each of the stimulus sets, and the number of items to be committed to memory was either two or six. In the Cue period, where the participant is expecting the presentation of the memory set (with full knowledge about the size of the set) we see initial increases in heart rate over the first few seconds, followed by significant decreases with the lowest heart rate achieved at the point of time where the memory set is expected to be presented. Note that both the accelerative and decelerative effects are more clearly seen when we allow for a 10 second interval between stimulus presentation. Control of some autonomically innervated systems is sluggish.

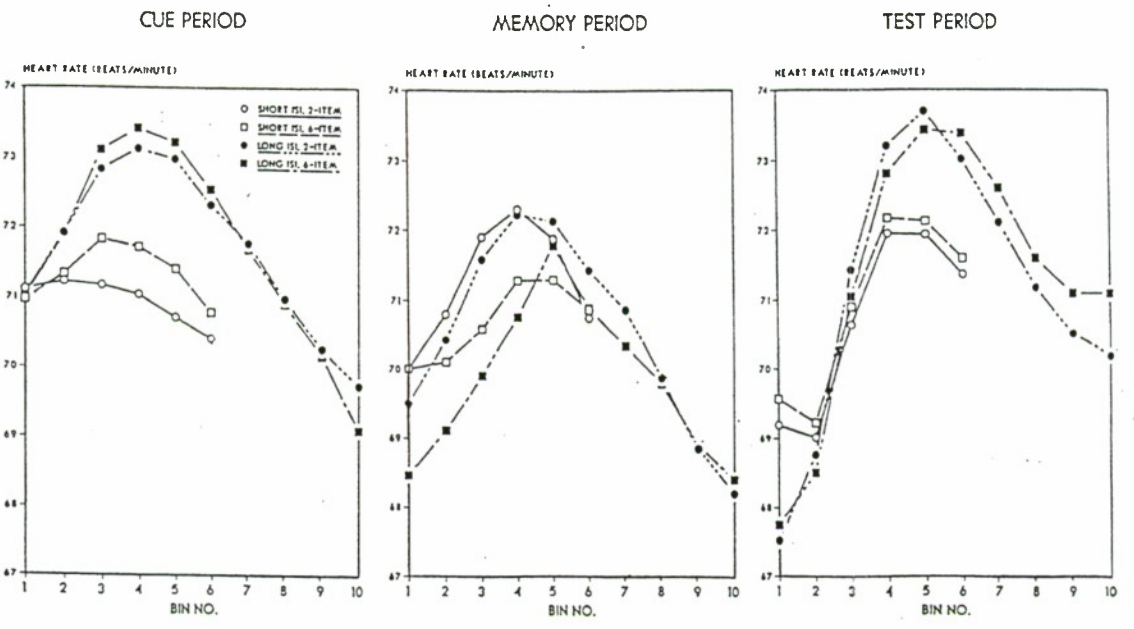


Figure 1

Note also that set size (or expected memory load) has no significant effect on either the accelerative or decelerative component. Expectancy of anything leads to equal amounts of deceleration. Heart rate changes during the information presentation and retention period shows a similar pattern of acceleration and deceleration in anticipation of the test stimulus. Here we do see some differences between large and small set size with some delay (phase shift) in the accelerative component for the larger set size. Presentation of the test stimulus leads to significant heart rate increases following the making of the response and a return to resting levels in anticipation of the next cue stimulus.

We, of course, also recorded the eyeblink in this situation. For those of you not familiar with our work, our working hypothesis (now amply verified) is that "spontaneous" blinks do not occur at random but are time locked to points in time where the participant can momentarily inhibit the taking in of new information or the processing of information. When are subjects most likely to

blink or inhibit blinking while performing the Stemberg memory task? Figure 2 depicts the results.

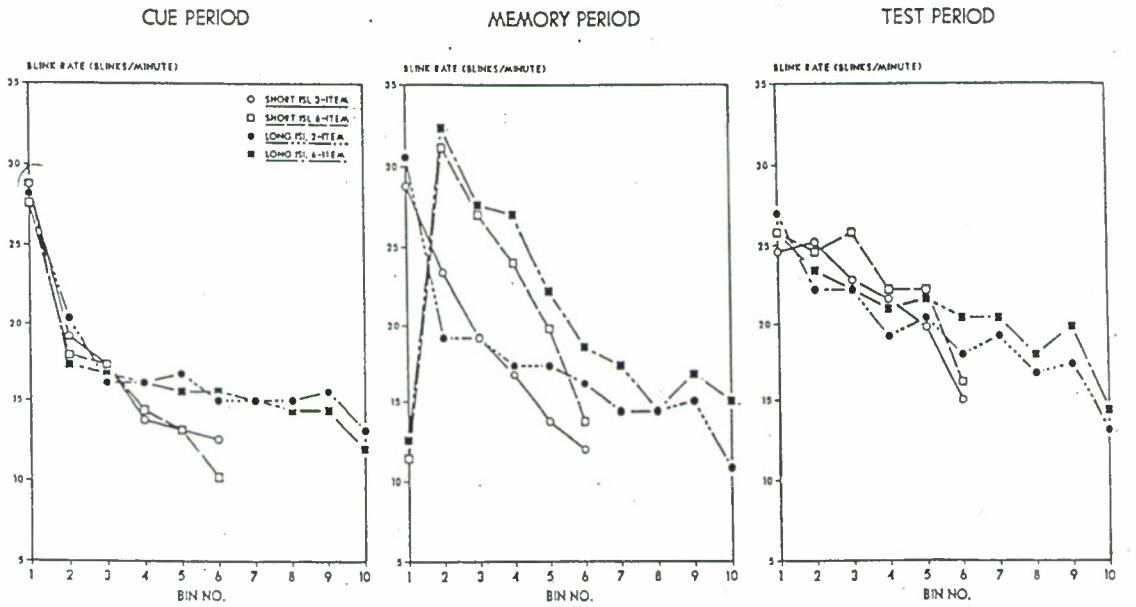


Figure 2

Note, that as was true of the heart rate response, the anticipation of anything leads to blink inhibition. Our data suggest that unlike the cardiac decelerative response the degree of blink inhibition seen with a 6 second stimulus onset asynchrony (SOA) is identical to that observed when SOA is 10 seconds. The CNS control of this system appears to stronger than is true of the heart. During the memory period we see major differences between the small and the large memory set in that blink rate increases to a peak during the first second following stimulus onset for the small set size, while the peak is seen during the second two for the larger set. Thus blinking is inhibited while participants are perceiving and "transferring" information to "working memory". Note also that the speed with which blink rate decelerates is slower for the larger set size. We believe this is associated with "rehearsal" as a procedure for maintaining information in working memory.

Not only is timing of blinks affected by task demands but the nature of the blink is equally affected. One of the measures we have developed is referred to as the 50% window measure. This measure identifies the point in time when the lid is half closed, searches for the point in time where during lid reopening the lid passes that same level and identifies the time interval between these two points as 50% window. Figure 3 depicts the results of this analysis.

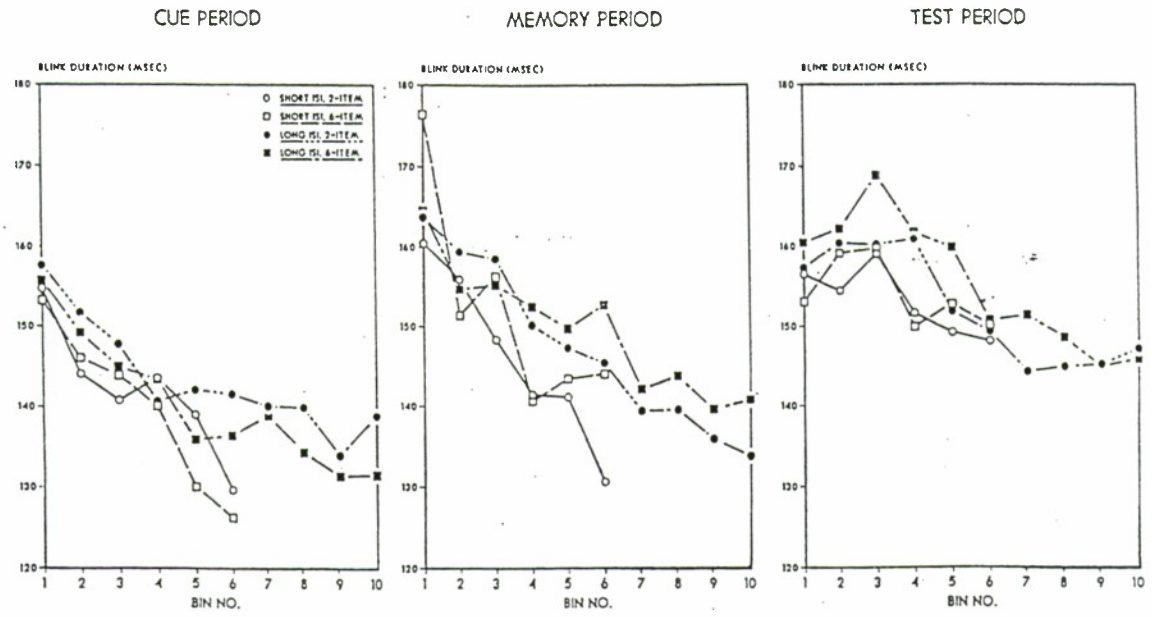


Figure 3

Note a significant decrease in closure duration as a function of expectancy. The closer in time to where we expect an event to occur the shorter the closure duration of the blink. We as well as Glenn Wilson have used this phenomenon to discriminate between levels of "work load" as pilots engage in various flight maneuvers. Stern and Skelly, for example demonstrated that B-52 pilots flying an extended mission in a flight simulator inhibited blinking and utilized more short duration blinks when painted by enemy radar or when in an "attack" mode or searching for a missile that had been shot at them than when flying under more tranquil conditions.

To summarize: *expectancy leads to a decrease in heart rate, inhibition of blinking, and if one cannot inhibit a blink at a point in time close to an imperative event the blink if of shorter duration than normal.*

Let me now turn to recent studies in which we are studying eye movements and pupil diameter changes as well as blinks and electroencephalographic variables, associated with level of *expectancy* .

The model situation we have chosen for these experiments is a vigilance task used by Paul Bakan in the 1950's. In this task subjects are required to view or listen to a series of digits sequentially presented. They are required to respond when a sequence of digits meets a "rule". The "rule" we have selected is, "Respond when three successive digits that are all odd integers and all different in value occur sequentially. Following the presentation of such an array of digits the next digit will always be an even one." Digits are presented at an SOA of 2.5 seconds, and in some of our experiments a sequence of 1440 digits constitutes the vigilance task. It is a vigilance task because the number of events requiring a response is small (60 such events in an hour)when compared to the number of stimuli presented.

The lowest level of expectancy (of having to make a manual response) is during the interval following responding. Here the participant knows that the next stimulus will be one that can be ignored. This is level 1 in our experiment. Level 2 is the expectation that the next digit will be the first of a series of three odd digits. Level three is the expectation that following the first odd digit the next digit will also be an odd one. Level 4 is the expectation that the next digit is a third odd digit and requires a response. These levels of expectancy are graphically presented in figure 4.

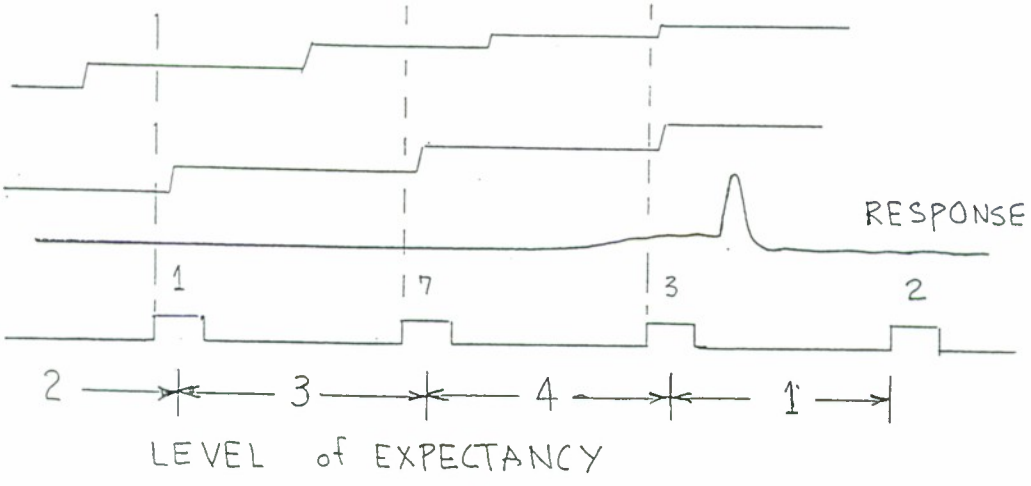


Figure 4

Digits are presented at predictable locations on a CRT. All of them are presented at the same horizontal level and at 4 discrete locations across the screen. The first stimulus appears at the left-most condition, the second one about 7 degree to the right, the third one 7 degrees further to the right, the fourth one another 7 degrees to the right, the fifth one 7 degrees to the left of the last one, etc. The display thus jumps in a regular fashion from left to right and back again for a sequence of 1440 stimulus presentations.

Stimulus presentation is short (300 or 500 ms). With the shortest presentation duration the viewer may have some difficulty abstracting the relevant information if one waits to shift gaze to the predictable location until the stimulus appears at that location. Viewers are thus expected to make *anticipatory* saccades. It was our hope that such anticipatory saccades would occur earlier when expectancy was high than when it was low. With respect to saccades that move the eyes to the target location we hoped that saccade latency might be affected by expectancy and that saccade gain would be similarly affected.

We are in the midst of analyzing this data. I can report on results based on 8 subjects. The results meet our expectations!

Figure 5 depicts anticipatory saccade latency as a function of expectancy level. In this figure as well as the ones that follow we have, for reasons that are not relevant here (namely our interest in evaluating changes as a function of Time-on-Task) only sampled data where a response was required. The results thus are based on a maximum of 60 events per subject. We have also excluded trials on which no response was made, this seldom exceeded 10% of the trials. Anticipatory saccade latency identifies the first saccade that occurs in anticipation of the next number and is measured backward from the time of stimulus presentation. A larger value thus signifies an earlier anticipatory saccade. It was our expectation that the higher the expectancy level the larger the latency value. Those expectations are reflected in the graph. Ranking these latencies for each subjects across the four expectancy values identified demonstrated that for expectancy value 1, six subjects had the lowest rank (shortest latency), one had a rank of two and one a rank of 4. For expectancy value 4, four had a rank of four, two of three one of two and one of one.

The probability of the ranking being attributable to chance was tested with the Friedman two-way analysis of variance by ranks. The results are significant beyond the .0000 level.

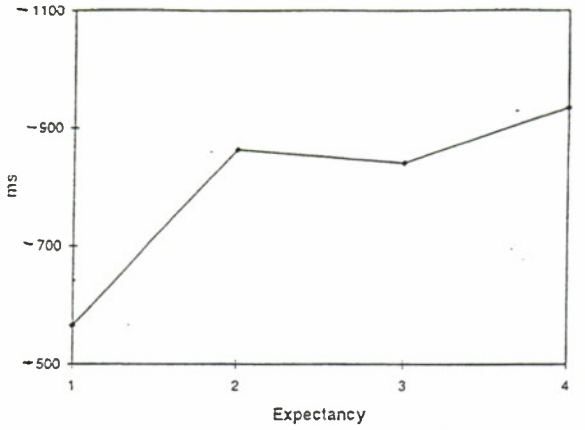


Figure 5. Anticipatory saccade latency (60 chunks).

Similar analyses were conducted for To Target Saccade Latency, To Target Saccade Gain and blink latency following stimulus presentation.

To Target Saccade Latency refers to the timing of the first saccade following stimulus presentation. For this analysis we excluded all trials where the anticipatory saccade accurately brought the eye to the target location, i.e., where there was no to target saccade. Though the differences in saccade latency as a function of expectancy are small (of the order of 7 ms between level 1 and level 4) they again are highly reliable (p.<.0000). Six subjects showed the expected pattern and two the opposite pattern.

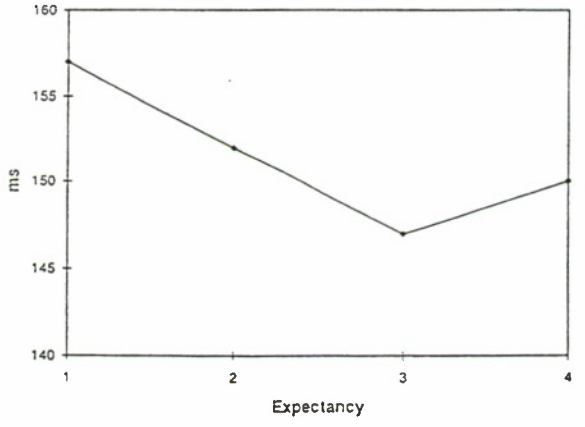


Figure 6. To target saccade latency (60 chunks)

To Target Saccade Gain is a measure of saccade amplitude divided by angular distance of the jump. To the extent that subjects make anticipatory saccades, the greater the accuracy of the

anticipatory saccade the smaller the gain. The results of this analysis are depicted in figure 7. Seven of the eight subjects had the smallest gain for expectancy level 4. Again results are significant ($p<.0000$).

The last measure, blink latency involves the time between stimulus onset and the making of a blink if a blink is made during an interval between stimulus presentation. What we find here is that the requirement to make a manual response produces a significant delay in blinking. Though there appears to be a decrease in blink latency as we shift from level 1 to 2 and 3, we have not tested the reliability of these differences.

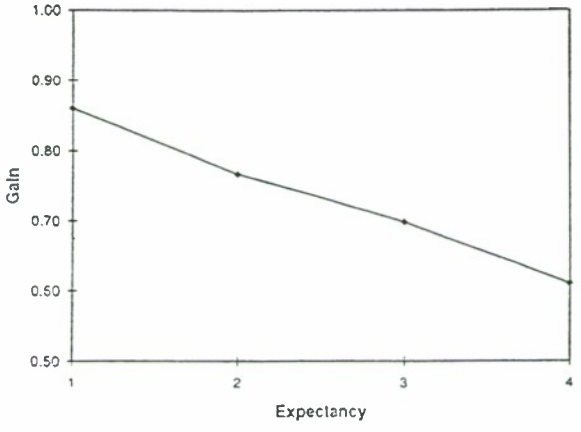


Figure 7. To target saccade gain (60 chunks).

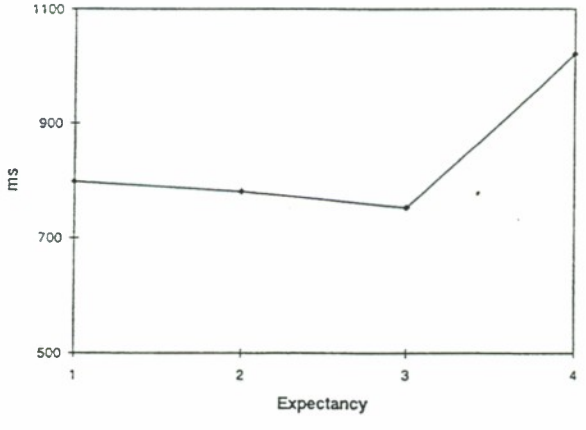


Figure 8. (Blink latency (60 chunks).

To summarize, in tasks where stimulus location and timing are predictable one finds significant effects of expectancy on measures such as anticipatory saccade latency on measures such as anticipatory saccade latency, to target saccade latency and gain as well as blink latency. Other measures will be considered as well, such as anticipatory saccade frequency, and the occurrence of "square wave jerks". Our choice of a one hour task has nothing to do with our interest in SA but is based on our concern with issues of vigilance and /or attention. We also can demonstrate significant Time -on - task effects with many of the above measures but this is neither the time nor place to do so.

The last measure we have utilized in these investigations of expectancy involves changes in pupil diameter as a function of "mental load". Again, this is an area of investigation with which most of you are probably not familiar. Pupil diameter changes are generally associated with alterations in light intensity and the ingestion of central nervous system stimulants and depressants. What is less well known and publicized is that the pupil also dilates as a function of "mental load". The changes are small but reliable. Much of this work was published in the 1960's and 70's by E. Hess, J. Beatty and Kahneman. Our current technology for evaluating oculomotor activity allows for the measurement of pupil diameter and we have started to look at this measure as affected by expectancy. We have not done any quantitative analyses of pupil diameter changes. The following figures are representative of pupillary changes associated with expectancy and the enactment of a motor response. Note that the major change is attributable to the enactment of the motor response. However, one can also see pupil diameter increases associated with expectancy.

What conclusions can we draw from what I have presented? I hope that I have made a positive case for the use of oculometric measures to index aspects of expectancy. I can make a similar case for the use of such measures for the evaluation of vigilance decrements.

Why am I so enthusiastic about the use oculometric measures. Why are they to be preferred over other physiological measures which may do as good a job, if not better, to index aspects of expectancy, vigilance and attention? One of the attractions of oculometric measures is that they can be obtained without the attachment of electrodes. They can be obtained with the use of video recording equipment. There are problems with the technology that are still expensive to implement. Current, commercially available technology does a fine job with respect to spatial resolution (identifying where the eyes are pointed), does an excellent job of measuring pupil diameter, but falls short when it comes to temporal resolution. We have done a reasonable job of developing software to automatically abstract variables of interest such as the occurrence of saccadic eye movements, eye blinks, and changes in pupil diameter, though much remains to be done.

Post-Hoc Assessment of Situation Assessment in Aircraft Accident/Incident Investigations

Barry Strauch

National Transportation Safety Board[1]

Introduction

Aircraft accident investigations have two primary objectives, to determine the cause of an accident and to make recommendations to prevent its recurrence. To meet these objectives investigators often attempt to determine the situation awareness of the crewmembers piloting the accident aircraft, or other persons critical to the cause of the accident.

Situation Awareness

The construct of situation awareness has become increasingly used, particularly in aviation (e.g., Endsley & Bolstad, 1994; Endsley, 1995). As Endsley has (1995) defined it:

> Situation awareness is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future.

As it would apply to piloting an airplane, for example, situation awareness would refer to pilots' recognition and comprehension of the state of the aircraft systems, components and flight path, and the understanding of their respective predicted states.

The importance of situation awareness to aviation safety derives largely from its influence on decision making (e.g., Klein, 1993, Orasanu, 1993). Effective pilot decision making uses accurate situation assessment or situation awareness as the foundation of subsequent decision making. For pilots, as all decision makers, the quality of the decision is dependent upon the accuracy of their awareness or assessment of the situation to which the decision will apply. Thus, even with the numerous comprehensive standard operating procedures, specified phraseology, and other rules and guidance, air transport pilots, air traffic controllers and other participants in the aviation system routinely make decisions that can have profound influence on the safety of flight. Their ability to maintain accurate situation awareness, therefore, is critical to the quality of the decisions that result and to the safety of the aviation system.

Repeatedly, accidents have resulted from flaws in the decision making of the flightcrew members, flaws that largely resulted from inadequate situation awareness. For example, in 1994 (NTSB, 1995a) a warning light illuminated in the cockpit of a regional airliner as the airplane was on final approach, in night, marginal visibility conditions. Although this was a relatively benign

[1]The views expressed herein are those of the author and not necessarily those of the National Transportation Safety Board.

occurrence, the captain misdiagnosed the warning light by interpreting it as an engine failure. Consequently, among other incorrect or inappropriate actions that followed, he failed to assign appropriate responsibility to the first officer, incorrectly decided to break off the approach and attempted a go around, and finally improperly applied engine power. the aircraft crashed about five miles short of the runway.

Although this accident illustrates a rather extreme example of the potential consequences of incorrect situation awareness, the fact remains that aviation safety relies upon accurate situation awareness by all its critical personnel including pilots, air traffic controllers, dispatchers, maintenance personnel, flight attendants, and others. The purpose of this paper is to describe the methods by which aircraft accident investigators assess the situation awareness of critical personnel in circumstances where those deficiencies are believed to have played a part in the accident.

Data

Aircraft accident investigators typically obtain data pertaining to the person or persons considered critical to the cause of the accident, the equipment used, and the environment in which the critical person or persons and the equipment operated. As in empirical investigations, investigators sample the universe of data to collect those necessary to meet the objectives of the investigation. The data that are collected, in addition to meeting standard measures of statistical quality, must be internally consistent, logical and sequentially correct. That is, regardless of the source, data obtained from the various sources in an accident investigation must manifest the application of comparable logic and the description of identical events, in the same sequence, a sequence that ends with the accident itself.

The Person

The majority of aircraft accidents result from an error or series of errors that a person critical to the flight has committed. Irrespective of the person, the data remain the same, data that describe the quality of the person's performance of the critical task, and the state of his or her behavioral and physical health at the time of the accident. The sources of data include primarily medical, personnel and training records and the characterizations of people who were associated with those records and/or who were familiar with the person.

The Equipment

The accident aircraft, as well as other equipment potentially involved in the cause of the accident, e.g., air traffic control consoles, provide considerable data about the events preceding the accident. Items such as control surface positions, instrument readings, non-volatile memory contents, switch and circuit breaker positions, and site damage offer substantial information about the state of the machine before the accident, information that can be critical to understanding the cause of the accident. For example, fire damage that is pre-impact exhibits different smoke patterns than post-impact fire, and determining when the fire initiated is critical to learning the accident's cause. Similarly, aircraft wreckage that is concentrated in a small area results from a different flight path than one that is dispersed, thus describing substantially different types of accident sequences and thus, potentially different causes.

Regardless, the most critical equipment-provided information derives from the two recorders, required on air transport and jet aircraft, that continually record data on the status and operation of

the aircraft. Digital flight data recorders (DFDRs) contain anywhere from 11 to over 100 parameters that measure the status and or position of the airplane's surfaces, engines, systems and pilot controls, as well as its flight path, from the time of the accident back through the preceding 25 hours. Cockpit voice recorders (CVRs) record sounds, conversations, alerts and warnings within the cockpit, including, as well as sounds accompanying changes in aircraft flight status, from the time of the accident through the preceding 30 minutes.

Additional sources of data are supplied by air traffic control facilities that record communications between pilots and air traffic controllers, as well as most communications among controllers themselves. further data are provided by air traffic control facilities that record radar information revealing the precise location, airspeed, and altitude of all aircraft in the airspace.

The Environment

Environmental data include information from weather radar and other devices that regularly measure and record weather parameters in the relevant airspace. These sources can provide information on the direction and velocity of the winds, visibility, temperature and precipitation level at different points in time in the airspace of interest. In addition, dispatch records of information provided to air transport pilots indicate basic information about the flight, including the planned flight path, weight and balance, fuel requirements and other data relevant to the investigation.

Post-Hoc Analysis

Often, within days of the initiation of an accident investigation patterns emerge within the data that suggest to investigators significant issues for further exploration. For example, the absence of pre-existing hardware failure often leads investigators to examine the actions and decisions of a critical actor involved in the accident flight. In that event, the situation awareness of the individual or individuals may be critical to determining the error that caused or contributed to the cause of the accident. As noted, examining the situation awareness follows the collection of data from a variety of sources, and the assurance that the data meet the requisite logical and statistical requirements of accident investigation.

For example, on August 16, 1987, a Northwest Airlines MD-80, flight 255, crashed shortly after departure from Detroit Metropolitan Airport, killing 154 persons. Initial witness reports described flames being emitted from the engines. As a result, much of the early activities focused on collecting data that could corroborate possible powerplant anomalies. However, DFDR data subsequently showed that the flaps and slats had been retracted during takeoff, indicating that without substantial additional airspeed the airfoils were incapable of providing the necessary lift to initiate and sustain the initial climbout. In addition, because the airplane's attitude at rotation was considerably higher than normal, to maintain the desired airspeed with retracted flaps or slats, the airflow into the engines was reduced. The reduced airflow led to compressor stalls within the engines which then produced the flames that the witnesses described.

Further, physical evidence obtained from the aircraft wreckage corroborated the DFDR data on the flap and slat positions. Other data included the calculated climbout performance of an MD-80 that matched the actual performance of the accident aircraft, as recorded on air traffic control radar and on the DFDR, only with the flaps and slats retracted. Finally, information from the CVR revealed that the pilots had failed to check the status of the flaps and slats after their taxi checklist procedures had been interrupted by an air traffic control clearance, further corroborating the findings that the flaps and slats had not been extended during the taxi. In this manner, the data from a variety of sources were consistent in describing the actions of the crew in setting the

configuration of the airplane, the airplane's performance in that configuration, and the effects of that configuration on engine performance.

These efforts established not only the configuration of the airplane at the time of the accident, but more importantly, provided the data necessary to determine the sequence of events leading up to the accident. With this knowledge, the cause of the accident could be determined. More important, by establishing the sequence of events, investigators could then attempt to reconstruct the situation awareness of the crew involved. In this accident, the airspeed the crew selected was appropriate for an airplane with extended flaps and slats. This information, with the information from the CVR, supports the conclusion that the crew believed, almost up o the impact,that the difficulty in the climbout they were experiencing was due to weather factors, as had been discussed before takeoff, and not to an improper aircraft configuration.

Example

On July 2, 1994, a USAir DC-9 crashed near Charlotte, North Carolina, during a severe thunderstorm. Investigators determined (NTSB, 1995b) that the aircraft had traversed an area of intense rain, then encountered a severe microburst and downburst, with a change in wind direction and velocity from a 35-knot headwind to a 26-knot tailwind and a vertical velocity of 30 feet per second. the accident occurred during the go-around, after the pilots had attempted to discontinue the approach. The airplane was destroyed, 37 passengers were killed, and the remaining 20 passengers and crew were injured in the accident. Both pilots, who were experienced in the DC-9 and had unblemished records with the airline, survived. On-site examination of the wreckage provided no evidence for a pre-existing malfunction of the airplane or its components.

Investigators sought to determine the situation awareness of the two pilots, particularly because of the extremity of the weather conditions and the salience of the rain shower the airplane had entered. The assessment of the crew's situation awareness was critical to understanding their attempt to continue flight into such adverse weather, as no rational pilot would deliberately endanger the safety of flight by attempting to traverse severe weather.

The flight sequence had begun in the afternoon in Charlotte, about an hour and a half before the accident, when the crew flew the accident airplane to Columbia, South Carolina. this flight provided them with a first hand encounter with the prevailing weather conditions at Charlotte and those that had been forecast for the time of the accident flight.At that time visual conditions dominated the Charlotte area, but a chance of late afternoon type (convective) thunderstorms was forecast. Before departing Columbia for the return flight to Charlotte, the pilots received both the current and predicted weather information for Charlotte, both of which were essentially unchanged.

The flight between Charlotte and Columbia was about a half hour. Upon nearing Charlotte, the airborne radar on the accident airplane indicated the presence of red or storm cells in the area. Controllers on the ATIS, a continuous tape loop that gives field conditions and other information to pilots, were reporting visual conditions over the area. As late as about seven minutes before the accident air traffic controllers told the crew to expect a visual approach to the field, an indication that weather conditions in Charlotte were good.

As the flight neared Charlotte, the weather over the field had begun to change. The crew observed rain over the airport and, on their airborne radar, continued to monitor the presence of a storm cell near the airport. Just over a minute after being told to expect a clearance for a visual approach, the approach controller informed the crew that they "may get some rain just south of the field," and then to expect an instrument landing system (ILS) approach, an indication that conditions had deteriorated. Although the crew acknowledged this communication, there was no indication , either from their conversations on the CVR or with air traffic control, that they understood its implications. The CVR showed that the pilots were attempting to locate the cell on

their airborne radar, but they did not discuss the deterioration of the Charlotte weather. Rather they had initiated their descent and approach check, a time of considerable activity in the cockpit. As a result, while they addressed the possibility of a go-around, the conversation on the CVR appeared to be in the context of attempting to avoid the storm cell identified on the airborne radar and not a formal review of the missed approach procedures associated with the ILS approach, as would have been required of a crew executing an ILS approach.

In addition, while Charlotte air traffic controllers acknowledged the deteriorating weather, they did not convey all of the critical weather information to the flightcrew and the crew inadvertently did not obtain critical information. For example, tower controllers had observed lightening over the airport, an important indicator of thunderstorm activity, and discussed it among themselves, but did not notify the flightcrew of this. As the rain intensified, controllers updated the ATIS to note the rain over the airport, as required after a change in the conditions. However, the crew missed being alerted to this update because they had changed radio frequencies to communicate with a different controller, the local controller that would issue them landing clearance. Finally, the crew received incomplete, and thus misleading information regarding the location of windshear that had been detected on the airport. Charlotte airport was equipped with a low level windshear alerting system (LLWAS) that detected rapid changes in wind direction and velocity in and around the airport. However, the LLWAS was alerting at all locations around the airport and not in one area exclusively. The crew was aware that they were to traverse the northwest area of the airport and hence, as they later testified, incorrectly believed that the windshear alert did not apply to them. the local controller later told investigators that he was aware of the northeast boundary alert only.

In addition, as all passenger-carrying turbojet aircraft, the accident airplane was equipped with an airborne windshear detection device. This system, as the ground-based LLWAS, detected significant changes in the direction and velocity of the wind the airplane was encountering, and warned the crew when it noted a substantial change in either. However, the system had been designed to inhibit a warning if it detected a windshear when the aircraft flaps were in transition, that is, either being extended or retracted. During the go-around the airborne windshear detection system recognized a windshear encounter, but because the flaps were being retracted to an intermediate position at that time, the system did not signal its detection of a windshear.

Although unaware of much of the weather information that the controllers were privy to, the crew repeatedly discussed the weather conditions during the approach into Charlotte, as the CVR showed. They observed and commented on the heavy rain over the field. They identified the storm cell on their airborne radar and attempted to locate it. They even discussed the potential presence of windshear and were prepared to execute a go-around if the weather conditions so warranted. Finally, they requested the tower controller to provide them with the airport surface winds and the ride conditions experienced by the pilots of aircraft just ahead of them. The controller responded by informing the crew of the direction and velocity of the steady and fairly strong crosswind, and after querying the pilots of the USAir jet just ahead, told them that crew had experienced a "smooth ride."

In summary, the crew had diligently attempted to obtain information regarding the weather. Yet is was clear by both their statements on the CVR and by their apparently routing go-around procedures that, until just prior to impact, they were unaware of the magnitude of the microburst they were encountering. The fact that they had attempted to traverse a severe microburst in itself demonstrates that their situation awareness regarding the weather conditions along the final approach path was deficient and this deficiency led to their decision to continue the approach.

Nevertheless, the evidence also indicates that the crew had obtained, but did not perceive or comprehend, considerable information that would have supported an alternative assessment regarding the severity of the weather. They were given the predicted weather of possible thunderstorm activity. They visually identified heavy rain over the field. They noted the radar portrayal of a storm cell near the field. Yet, despite their obtaining necessary information regarding the severity of the weather conditions, the crew had also been given incomplete or incorrect Charlotte weather information that adversely influenced their situation awareness. Specifically, the absence of an alert from the airborne windshear detection system, the incomplete information on the location of the windshear on the airport surface, and the lack of information regarding both the

lightening and heavy rain over the airport served to convey to the crew an incorrect assessment of the severity of the weather ahead.

However, more significant than the incomplete and inaccurate information they were given, it is likely that the ride report from the crew of the aircraft just ahead was most influential in the crews' situation awareness regarding the weather. As a pilot from the same airline as the accident crew and in command of a turbojet as they were, the reporter was well versed in the kind of weather that would exceed the capabilities of the accident airplane and was knowledgeable on the company's guidance regarding the weather conditions that its pilots could not safely traverse. Further, because he was just ahead of the accident crew and thus, closest to the conditions in the airspace they were about to enter, his report contained the most timely information on the weather. That pilot would not know that the conditions at that time were so dynamic that in the approximately two to three minutes separating the two flights, the conditions in that airspace had deteriorated to the severe level the accident flight encountered.

The incorrect situation awareness of the pilots led them, in addition to continuing an approach beyond the point that it should have been discontinued, to fail to anticipate a microburst of the severity that they then encountered. In fact, the CVR revealed that just before impact, the captain, who was not flying the airplane, commanded the first officer to lower the nose, an action opposite to the guidance airlines provide their flightcrews in a windshear escape maneuver. In response to this call, the first officer did lower the nose and the airplane ceased its climb. The crash then occurred within 15 seconds.

Summary and Conclusion

Assessing situation awareness in aircraft accident investigators requires the collection of evidence from a variety of diverse sources including the aircraft wreckage, the accident site, crew records, recordings of air traffic control, radar, cockpit voice recorder, and flight data recorder, as well as various training and personnel records. Analysis of that evidence, as in the USAir accident in Charlotte, illustrates how the situation awareness of a highly qualified and trained crew can be faulty, and how that deficient situation awareness could cause the crew to make decisions and take actions that proved to be unsafe.

Situation awareness of pilots in a dynamic environment can and should be changeable as more current or more accurate information is obtained. The Charlotte accident demonstrated how susceptible situation awareness can be on information from numerous participants in the airspace. For example, the Charlotte air traffic controllers provided information that adversely influenced the situation awareness of the crew. Yet, air traffic controllers are trained to prioritize their tasks so that information considered "informative" or "advisory" is secondary to that considered critical, i.e., the separation of air traffic. Thus, while the air traffic controllers did not violate their own procedures, this accident demonstrated that their actions directly contributed to the deficient situation awareness of the accident crew. The NTSB's investigation determined that the failure of the Charlotte controllers to provide complete information to the pilots contributed to the accident.

Moreover, in examining the evidence, the USAir pilots of the aircraft just ahead of the accident flight likely inadvertently influenced the deficient situation awareness of the accident flight by providing the information they did on the quality of their flight along the final approach path. Certainly regardless of the ride report, the accident pilots should have been more sensitive to the possibility of an encounter with a severe microburst. Their lack of awareness likely also led to their failure to anticipate and respond immediately to the microburst encounter.

This accident also illustrates the criticality of situation awareness to flight operations. Although maintaining situation awareness requires obtaining and assimilating considerable information in dynamic environments, sale operations also requires that pilots and others operating in the airspace anticipate hazards that may not be apparent, and be prepared for a quick reassessment of the

situation when warranted. The failure of the accident crew in Charlotte to do that is perhaps their most significant error.

References

Endsley, M. R. (1995). Toward a theory of situation assessment. *Human Factors, 37*, 32-64.

Endsley, M. R., & Bolstad, C. A. (1994) . Individual Differences in Pilot situation Awareness. *International Journal of Aviation Psychology, 4,* 241-264.

Klein G. (1993). *Naturalistic Decision Making: Implications for Design.* Crew System Ergonomics Information Analysis Center. Wright-Patterson Air Force Base, Ohio.

National Transportation Safety Board (1995a). Flight into Terrain During Missed Approach, USAir Flight 1016, DC-9-31, N954VJ, Charlotte/Douglas International Airport, Charlotte, North Carolina, July 2, 1994. Author.

National Transportation Safety Board (1995b). Controlled Collision with Terrain, Flagship Airlines, Inc., dba American Eagle, Flight 3379, BAe Jetstream 3201, N918AE, Morrisville, North Carolina, December 13, 1994. Author.

Orasanu, J. M. (1993). Decision making in the cockpit. In E. L. Wiener, R. L. Helmreich, and B. G. Kanki (Eds.). *Cockpit Resource Management. New York: Academic Press.*

Air Traffic Controller Awareness of Operational Error Development

Mark D. Rodgers[1], Richard H. Mogford[2] and Leslye S. Mogford[2]

[1] Federal Aviation Administration
[2] Rigel Associates

Introduction

In the history of the Federal Aviation Administration, no aircraft have collided while under positive control in en route airspace. However, aircraft have violated prescribed separation minima and approached in close proximity. This event can occur as a result of either a pilot deviation, or an operational error (OE). This study analyzed data gathered during the investigation process for OEs. An OE takes place when an air traffic controller allows less than applicable minimum separation criteria between aircraft (or an aircraft and an obstruction). Standards for separation minima are described in the Air Traffic Control (ATC) Handbook (FAA Order 7110.65J, and supplemental instructions). While there is considerable complexity in those standards, at flight levels between 29,000 feet and 45,000 feet, Air Traffic Control Specialists (ATCSs) at en route facilities are required to maintain either 2,000 feet vertical separation or 5 miles horizontal separation between aircraft. At flight levels below 29,000 feet with aircraft under instrument flight rules (IFR), ATCSs are required to maintain either 1000 feet vertical separation or 5 miles horizontal separation. This study focused on data gathered from the Atlanta Air Route Traffic Control Center (ARTCC, also called an en route facility).

Immediately after the detection of an OE, a detailed investigation is conducted in an attempt to fully describe the events associated with the error's occurrence. This includes removing the ATCS(s) from the operating position and obtaining a statement from each of the involved specialists, gathering the relevant data (voice and computer tapes), and reviewing in detail the events associated with the errors occurrence. At the Atlanta ARTCC the Systematic Air Traffic Operations Research Initiative (SATORI) system is used in the investigation process to re-create the error situation in a format much like the one originally displayed to the ATCS (Rodgers & Duke, 1993). SATORI allows for a more accurate determination of the factors involved in the incident. Once the OE has been thoroughly investigated, an OE Final Report is filed. This report, the Final Operational Error/Deviation Report (FAA 7210-3), contains detailed information about each error obtained during the investigation process. It includes information regarding the controller's awareness of the developing error.

Specifically, the 7210-3 requires the Quality Assurance Specialist investigating the OE to provide a yes/no answer to the question: "Was the Employee Aware that an Operational Error was Developing?" Previous research has suggested that controller awareness of error development is related to OE severity (Rodgers and Nye, 1993; Durso, Truitt, Hackworth, Ohrt, Hamic, Crutchfield, and Manning, 1995). Rodgers & Nye (1993) found that controller awareness of error development was associated with less severe errors. Durso, et. al. (1995), in an analysis of data for a one year period not covered in the earlier Rodgers and Nye study, confirmed that controller awareness of error development resulted in significantly less severe errors. Given the relationship of controller awareness to OE severity, further study of the awareness variable is warranted.

There were two reasons for conducting the analyses included in this study. The first purpose for this study was to examine the characteristics of the OEs as a function of whether the ATCS was reported being either aware or unaware of the developing error. A sector is a volume of airspace in which air traffic control services are provided. An area is a group of sectors. Each controller specializes in one area and works only in that group of sectors. It was hypothesized that awareness of error development differed as a function of sector type, number of aircraft, number of people on position, amount of aircraft separation, time of day, and sector complexity. The second purpose of this study was to identify those sectors at Atlanta Center that demonstrated a relationship to the awareness variable. It was hypothesized that certain sectors could be identified or characterized as low awareness sectors. Further, it was hypothesized that sectors identified as low awareness sectors would differ from high awareness sectors in regard to various sector complexity measures.

Method

Operational Error Data

Quality Assurance (QA) personnel at each facility are responsible for gathering data and completing an OE Report in accordance with FAA Order 7210.3K (Facility Operations and Administration). For the purposes of this study, a number of fields from 103 OE reports from the Atlanta ARTCC were coded and entered into a separate data file. Those OEs where more than one sector was involved (13), no final report was available (4), or the error was attributed to an equipment failure (1), were not included in this analysis. This left a sample of 85 OEs, covering a three year period from June 1992 to June 1995.

ATCS Awareness

One of the items contained on the final report requires an assessment of the involved employee's awareness of the developing error. This item has been on the OE final reporting form for the past 14 years. After listening to the associated voice tape, interviewing the involved controller, and reviewing the error with SATORI, QA specialists make a determination as to the controller's awareness. Although SATORI simplifies the formulation of this judgment, most QA specialists find the answer relatively easy to ascertain. Typically, if either the control action to provide separation was not issued in a timely manner, or no control action was initiated, the controller was judged to be unaware of the developing error. However, if the controller actively attempted to provide separation to the involved aircraft, although the control action was either inappropriate or inadequate, the controller was judged to be aware of the developing error.

Airspace Complexity Measures

Two measures of airspace complexity were used in this study. The first, called a facility average complexity estimate, is calculated by facility personnel during the sector validation conducted each year. This assessment involves estimating the sector complexity workload using a formula that weights various ATC functions (FAA Order 7210.46). Functions and their associated weights (in parentheses) include: number of departures (5), number of arrivals (4), number of radar vectored arrivals (2), number of en route aircraft requiring control actions (4), number of en route aircraft not requiring control actions (2), number of emergencies (4), number of special flights (3), and number of required coordination's (1). These 8 functions are evaluated, weighted, and totaled to

derive the sector complexity workload value. The sector complexity workload values from 1995 were used in this study.

The second measure of airspace complexity, hereafter called the 16CF, was obtained using a 16 factor survey developed by Mogford, Murphy, and Guttman (1993). These factors include:

- Amount of climbing or descending traffic.
- Degree of aircraft mix (VFR, IFR, Props, Turboprops, Jets, etc.).
- Number of intersecting flight paths.
- Number of multiple functions the ATCSs must perform (terminal feed, in-trail spacing, etc.).
- Number of required procedures that must be performed.
- Number of military flights.
- Amount of coordination required.
- Extent to which the controller is affected by airline hubbing.
- Extent to which weather related factors affect ATC operations.
- Number of complex aircraft routings.
- Extent to which the controller's work is affected by restricted areas, warning areas, MOAs and their associated activities.
- The size of the sector airspace.
- The requirement for longitudinal sequencing and spacing.
- Adequacy and reliability of radio and radar coverage.
- Amount of radio frequency congestion.
- Average traffic volume.

There are seven areas of specialization at Atlanta Center, with approximately seven sectors per area. Each of the above 16 complexity factors was evaluated for each sector in each area by the Airspace, Policy, and Procedures Specialist assigned to that area of specialization, using a seven point scale.

Results

Of the 85 OEs analyzed, the ATCS charged with the error was judged not aware in 72% (62) of the cases, and aware in 28% (23) of the cases. Chi-square an t-tests of awareness (not aware vs. aware) by sector type (Ultra High, High, and Low), number of aircraft being controlled, number of persons on position at the sector, sector complexity (facility estimation, and 16 CF total), and time of day did not yield significant relationships. Greater separation existed for both the vertical and horizontal dimensions for errors in which the controller was aware. However, as shown in Figure 1, only the horizontal component achieved significance ($t(54,69) = -3.06$, $p \leq .003$).

In an effort to describe the association of controller awareness with specific sector characteristics, for those sectors in which more than one error occurred, two groups of sectors were created. Since the mean number of OEs for all sectors in which any occurred was 1.9 with a S.D. of 2.1, the cut between low and high error sectors was set between 3 and 4. Sectors with 3 or fewer errors were categorized as low error sectors, and those with 4 or more errors were categorized as high error sectors. Furthermore, since no variance was associated with awareness for the single error sectors (i.e., no controllers in single error sectors were aware), those errors were excluded from this analysis.

Sectors were ordered by number of OEs that occurred in each sector (low to high) and plotted as a function of the number of non-aware vs. aware errors (see figure 2). There is an apparent trend in the graph toward less awareness of OEs as the total number of errors in a sector increased. This was demonstrated in a significant overall correlation between total OEs in a sector and the

number of OEs without awareness. This result was true when including all of the sectors with 2 or more errors ($r = .84$, $p \leq .001$), and also when focusing on the high error sectors alone ($r = .86$, $p \leq .01$). Correlation's of total sector errors with errors when there was awareness were not statistically significant in either group.

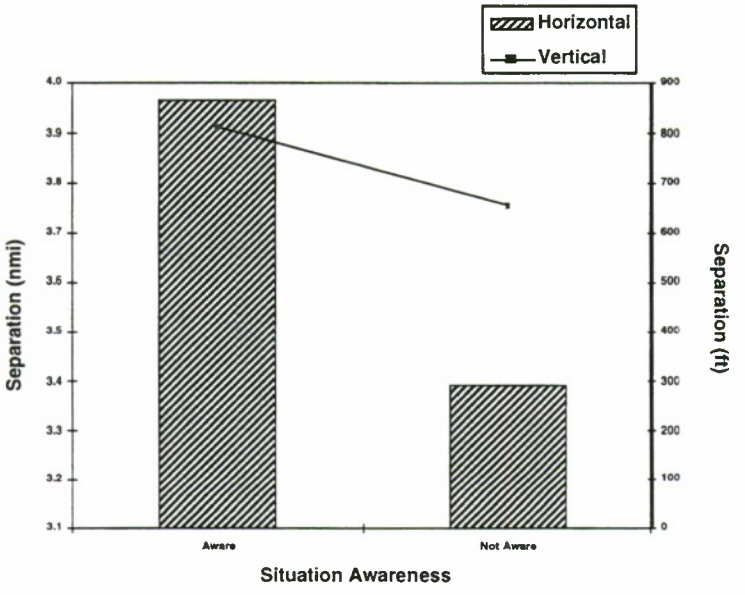


Figure 1. Horizontal and vertical separation

Further analyses were conducted in an effort to describe sector complexity differences as a function of sector error categorization. Sectors were classified as either non-error sectors (0 errors), low error sectors (3 or fewer errors), or high error sectors (4 or more errors). A MANOVA was performed on all complexity variables that exhibited moderately significant ($p \leq .1$) non-parametric (Spearman) correlation's with OE group classification. The MANOVA was significant (Hotellings, $F(26, 58) = 2.12$, $p \leq .009$). Univariate ANOVAs comparing sector classification as a function of both the facility average complexity estimates and the 16 CF Total were statistically significant ($F(2, 42) = 5.45$, $p \leq .008$ and $F(2, 42) = 3.12$, $p \leq .05$, respectively). Two of the individual 16 CF factors, frequency congestion ($F(2, 42) = 6.19$, $p \leq .004$), and weather ($F(2, 42) = 3.75$, $p \leq .03$) also differed significantly as a function of sector error categorization. Tukey's HSD post-hoc test was used to determine which groups contributed to the significant differences. For both of the global complexity measures and the 2 16CF factors the difference between the 0 OE group and the 4 or more OE group was significant at $p < .05$ (see Table 1).

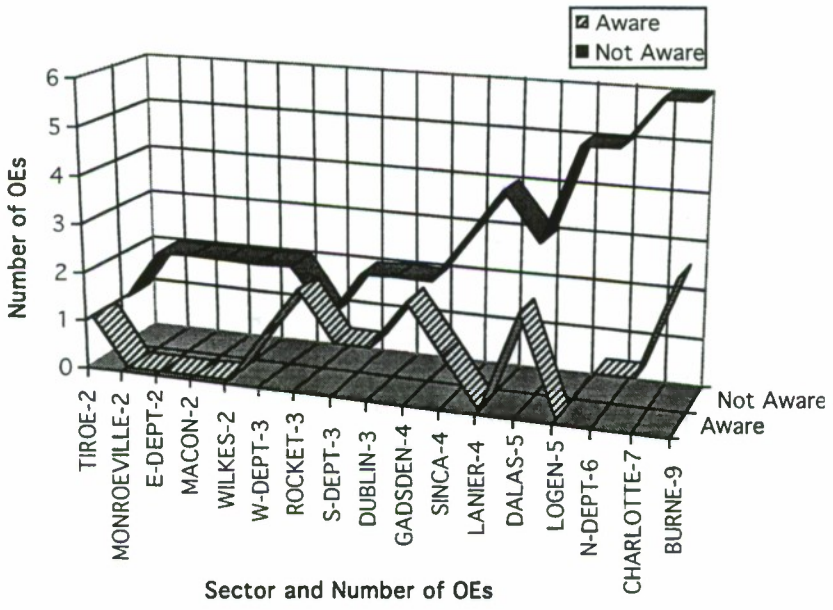


Figure 2. SA for Sectors with Two or More OEs

Table 1. Sector Complexity and Sector Error Rates

Variable	No OEs	3 or Fewer OEs	4 or More OEs	F(2,42)	Prob.
Fclty. Avg. Cmplex. Est.	**269.3**	343.8	**465.0**	5.45	0.008
Total Complexity (16CF)	**70.3**	75.7	**83.0**	3.12	0.054
Frequency Congestion	**4.1**	5.2	**6.1**	6.16	0.004
Weather	**4.7**	5.0	**6.1**	3.75	0.032

Note: The Tukey HSD post-hoc test was conducted on variables identified by the one-way ANOVAs. Bold highlighting indicates a significant difference between groups ($p < .05$).

Discussion

This study found that awareness of error development was significantly related to sector error rates. In general, controllers involved in OEs at sectors in the Atlanta ARTCC that have greater error occurrence tend to demonstrate less awareness of the developing error. With further study, it may be possible to reliably define sector characteristics that distinguish high error / low awareness sectors. It may be shown that low awareness, presumably influenced by sector complexity, leads to higher OE incidence. In support of this hypothesis are the findings that both of the global measures of sector complexity were significantly related to high error rate sectors.

It is interesting to note, for the awareness variable, only the horizontal separation parameter, not vertical separation, was significantly affected. This finding was first suggested by Siddiquee (1974), found by Rodgers and Nye (1993), and confirmed in this study. This may be because altitude, used to provide vertical separation, is reported numerically in the data block. Thus, altitude information is more likely to be salient than is information used to make a judgment about the extent of horizontal separation between aircraft. Horizontal separation judgments are based on visual estimates of distance between two targets on the plan view display and judgments about their relative speed. However, if ATCSs prefer to use horizontal separation, the difference may be based primarily on frequency of usage.

Much work is required to explain how the loss of situation awareness relates to OE occurrence. In an attempt to provide more detailed information about the nature of ATCS awareness of error development, the OE final report was recently modified to allow for a more detailed description of the loss of awareness. Three causal factor categories were added that directly assess the nature of the loss of awareness. The three categories are: failure to detect displayed information, failure to comprehend displayed information, and failure to project the future status of displayed data. These three causal factors directly relate to the Endsley (1995) model of situation awareness. It is hoped that through the use of SATORI and the new OE causal factor categories it will be possible to characterize more accurately the nature of losses of situation awareness during the occurrence of OEs.

References

Durso, F. T., Truitt, T. R., Hackworth, C. A., Ohrt, D. D., Hammic, J. M., and Manning, C. A. (1995). *Factors Characterizing En Route Operational Errors: Do They Tell Us Anything About Situation Awareness?* Presented at the International Conference on Experimental Analysis and Measurement of Situation Awareness, Nov. 2, Daytona Beach, FL.

Endsley, M. R. (1995). Towards a Theory of Situation Awareness. *Human Factors, 37,* 32-64.

Federal Aviation Administration. (1995). Air Traffic Control (*FAA Order 7110.65J*). Washington, D.C.: U.S. Department of Transportation.

Federal Aviation Administration. (1991). Facility Operation and Administration (*FAA Order 7210.3K*). Washington, D.C.: U.S. Department of Transportation.

Federal Aviation Administration. (1984). Establishment and Validation of En Route Sectors (*FAA Order 7210.46*). Washington, D.C.: U.S. Department of Transportation.

Mogford, R. H., Murphy, E. D., and Guttman, J.A. (1993). Using Knowledge Exploration Tools to Study Airspace Complexity in Air Traffic Control, *The International Journal of Aviation Psychology, 4(1),* 29-45.

Rodgers, M. D. and Duke, D. A. (1993). SATORI: Situation Assessment Through Re-creation of Incidents, *Journal of Air Traffic Control,* October-November.

Rodgers, M. D. & Nye, L. G., (1993). Factors associated with the severity of operational errors at air route traffic control centers. In M. D. Rodgers (Ed.). *An Examination of the Operational Error Database.* DOT/FAA/AM-93/22. Washington, D. C., Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine.

Studying Situation Awareness in the Context of Decision-Making Incidents

Gary Klein

Klein Associates Inc.

Introduction

As basic and applied researchers have become more active in studying cognitive tasks, the topic of situation awareness (SA) has become more widely discussed and studied. Experts appear to differ from novices in ability to size a situation up, and we would like to understand more about the situation assessment processes they are using, along with the content of the situation awareness they achieve. This objective is difficult because SA can have so many facets; cataloguing the contents of a person's SA often turns out unsatisfactorily, and comparing the degree of SA held by experts and novices is not always very helpful.

I believe that it is important to study SA in the context of decision-making incidents (both actual and simulated). Otherwise, researchers may run the risk of conducting an open-ended inquiry, since SA could include everything that a person is or could be aware of. There is no convenient stopping rule to guide us about when to terminate the analysis. Another difficulty is that there are no basic "elements" and so the contents of SA will have to be, to some extent, arbitrary. This is especially true of research within a rich context, where the context will affect the way we define the aspects of SA. We can try to avoid these complications by studying SA in an environment with a restricted context, but this will reduce our ability to generalize the findings. For these reasons, it may be useful to study decisions, judgments, and the like, and to examine SA as it contributes to these judgments and decisions. Instead of studying the question of WHAT—what is the content of a person's SA, we can study the question of HOW—how the SA affects action. In doing so, we can identify some of the important aspects of SA—those that impact judgments and decisions. If we try to catalog the contents of SA outside of incidents, outside of a context of action, we risk losing in artificiality what we have gained in precision.

This paper describes two ways of studying incidents: (a) using **retrospective memory** for actual incidents, and (b) using **process-tracing** methods for simulated or ongoing incidents. The Critical Decision method was developed to probe the way people made difficult judgments and decisions involving actual incidents in their professional careers. Process-tracing methods include think-aloud protocols and other ways of gaining insight into the thought processes linked to task performance.

The Critical Decision Method

Klein, Calderwood, and Clinton-Cirocco (1986) developed the Critical Decision method to understand how fireground commanders made difficult decisions. Since we were unable to observe fireground commanders and gather verbal protocols during actual incidents, we settled on retrospective accounts of previous challenging incidents. One of the findings of this study was that

the fireground commanders rarely made decisions about which course of action to pursue. Instead, they were able to use their experience to recognize the typical course of action in a situation. Therefore, most of the information gathered was about how the fireground commanders understood the situation, and how their SA shifted as the conditions changed. Sometimes, the shift was an elaboration of SA as they were able to reduce uncertainty about the details, and sometimes the SA shifted radically as the fireground commanders rejected one interpretation in favor of another.

The Critical Decision method has been further developed (e.g., Klein, Calderwood, and MacGregor, 1989; Klein, 1993). Currently, the method centers around a difficult judgment or decision, one that requires expertise. The objective of the method is to uncover the expertise, particularly the cues and patterns, used to identify the dynamics of the situation. Once the subject-matter expert recalls a demanding event, the interviewer will often conduct several "sweeps" through the incident. The first sweep through the incident is to gain a brief overview of the event, to see if it is suitable and to anticipate the types of questions that will be relevant. The second sweep is a more deliberate recounting of the incident, often putting it on a time scale. The third sweep uses a variety of probes to examine the judgments and decisions that appear to depend most strongly on expertise. These probes include asking about salient cues, relevant knowledge, goals being pursued, overall understanding of the situation, potential courses of action, and hypothetical issues about the effect on the decision if key features of the situation were different. A fourth sweep is sometimes used to probe how a person with much less experience would have sized up the situation at key points, and to learn which cues or patterns an inexperienced decision maker might miss or misinterpret.

In the research conducted by my colleagues and me, the Critical Decision method is our Cognitive Task Analysis strategy of choice. This is because we are able to learn a great deal about the SA and decision making in context. We discover important considerations and factors that are not usually covered in manuals or in global accounts of the task. In a study of anti-air warfare decision making by US Navy officers (Kaempf, Wolf, Thordsen, & Klein, 1992), virtually every one of the 14 incidents studied included a key variable that was context-specific and would typically **not** be included in simulations. These ranged from judgments of the competence of other officers to difficulties of rousing commanding officers from sleep. Kaempf et al. (1992) prepared detailed diagrams of the successive stages of SA for the naval officers, along with the factors that contributed to each change in SA.

The Critical Decision method has been useful in helping us to discover some of the bases of situation awareness in experienced personnel in a number of different domains. Crandall and Getchell-Reiter (1993) described how the method was used to uncover the cues and patterns used by nurses in a Neonatal Intensive Care Unit to diagnose sepsis.[1] Kaempf et al. (1992) described how the method was used to study mental simulation in Navy officers trying to interpret events and also to project a course of action forward in order to evaluate it.[2]

In the course of our research, my colleagues and I have learned that there are situations and tasks for which the Critical Decision method is not well suited. These boundary conditions include tasks which are highly repetitious, tasks which lack clear feedback, and tasks which do not result in dramatic or memorable incidents. When we encounter such tasks, we rely on alternative methods, such as process tracing.

---

[1]This corresponds to Endsley's (in press) Level 1 SA with regard to the cues themselves, and Level 2 SA with regard to diagnosing sepsis.

[2] This corresponds to Endsley's (in press) Level 3 SA, projection of future status.

Process-Tracing Methods

In our work, process-tracing methods usually are employed with simulated incidents, although they can also be used with actual performance. For example, Kaempf, Klinger, and Wolf (1994) studied baggage screeners by observing them at work in airports and questioning them about the judgments they were making in selecting items to scrutinize more closely.

More often, process tracing is used with simulated incidents, and this offers the potential for greater control of the research. The incident can be developed to tap into the variables of greatest interest. Both experts and novices can be studied in the same tasks, to compare their SA. Unobtrusive measures and interventions can be incorporated, e.g., requiring the participants to take actions in order to obtain certain types of information, and testing workload by presenting secondary tasks. The simulation can be halted to allow more detailed questioning. There are many different ways to design simulations and process-tracing strategies.

We have found process tracing to be highly informative about SA as it relates to judgment and decision making in the simulated tasks. We have used the same types of probes as with the Critical Decision method, asking about the cues that were most salient for sizing up the situation, the types of misinterpretations that novices might make, and so forth. For example, by asking participants what alternative actions were possible, we can elicit the features of the situation that precluded these actions. And by asking what would have to happen to make the alternatives more attractive, we could learn more about the interplay between different factors. If we ask whether a certain course of action makes sense, and get the response "It depends," that is a natural opening to explore upon what it depends. Wolf, Hutton, Miller, and Klein (1995) used this type of process tracing to examine the SA of Patriot missile battery officers. Miller and Lim (1993) found process tracing to be very useful in studying the way weaponeers sized up situations prior to developing plans, and were able to design a prototype decision support system based primarily on process-tracing data.

One very powerful use of process tracing is to contrast the SA of experts to novices when faced with the same circumstances. deGroot (1946/1965) was able to make these contrasts using chess positions, uncovering a number of interesting differences between strong and weak players in their assessments of the dynamics of the positions. Charness (1989) did the same using bridge hands. It is worth noting that these simulations involved relatively straightforward stimuli—pictures of chess boards and of bridge hands. Wolf et al. (1995) used simulations built from paper maps and transparent overlays. In a market research study of consumer decision making, we used photographs of supermarket aisles to probe what the consumers were noticing and looking for.

The disadvantages of process-tracing methods include the following: they can sometimes require effort to construct the simulation, including the scenario of the incident and the unfolding of events, particularly if high-fidelity representation of dynamic and subtle cues is needed. A second disadvantage is that simulations are limited to the variables that the researcher already knows about. A third disadvantage is that simulations can be unrealistic, particularly with regard to the ways cues are represented and to stressors such as fatigue, threat, high stakes, and high workload from additional tasks.

The Use of Introspection

Both retrospective and process-tracing methods depend on introspection. Researchers are asking the participants to articulate what they are seeing, noticing, inferring, and interpreting. The dangers of introspection have been clearly presented during the past century. Nisbett and Wilson (1977) have claimed that people are inaccurate about their own reasoning processes. Therefore, researchers must be careful not to uncritically accept self-reports, particularly reports about the

reasons why people performed certain actions or made certain judgments. In fact, we recommend that researchers try to avoid questions about WHY judgments and decisions were made. Instead, the interviews can emphasize the cues and patterns that were noticed, without having to request introspection about the inferential processes themselves.

Despite these cautions, researchers should not feel that they have to abandon self-reports. Howard (1994) has argued that self-reports are important and useful forms of information. He notes that even behavioral measures of performance are susceptible to biases and distortion. The selection of stimulus materials and dependent variables can limit the generalization of findings. Some studies (e.g., Cole, Howard, & Maxwell, 1981; Cole, Lazarick, & Howard, 1987; and Howard, Conway, & Maxwell, 1985) have found that self-report data have **higher** validity than behavioral measures. Clearly, researchers must be careful to understand the limitations of any methods used, and should document methods to allow replication of findings by others. Nevertheless, self-reports can be considered to be useful form of data that may have higher validity and utility for SA researchers than behavioral measures.

Situation Awareness and the Recognition-Primed Decision Model

Klein, Calderwood, and Clinton-Cirocco (1986) have formulated a Recognition-Primed Decision model of how people can make decisions in naturalistic settings without having to compare options. The key is that people use expertise to size up situations and recognize typical courses of action as the first ones considered. Expertise centers around SA. When we distinguish between developing SA and selecting a course of action, it is the former that seems most important.

In our studies, we have identified four aspects of SA that appear to be central in many domains: expectancies about what will happen next, determination of the most relevant cues, identification of plausible goals, and recognition of typical courses of action. When a person understands a situation and sees it as typical in some way, that understanding encompasses expectancies, relevant cues, plausible goals, and typical actions in the situation. These types of knowledge reflect SA. They are not a prior stage that gives rise to SA, nor are they a subsequent stage that is inferred from SA. When a person forms an awareness of the dynamics of a situation, these four aspects enable that awareness to be translated into judgments, decisions, and actions.

Acknowledgements

I would like to thank Beth Candall for her helpful review and recommendations. The preparation of this paper was supported by the Naval Personnel Research & Development Center (Contract N66001-94-C-7034).

References

Charness, N. (1989). Expertise in chess and bridge. In D. Klahr and K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon,* pp. 183-208. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, D. A., Howard, G. S., & Maxwell, S. E. (1981). The effects of mono versus multiple operationalization in construct validation efforts. *Journal of Consulting and Clinical Psychology, 49,* 395-405.

Cole, D. A., Lazarick, D. M., & Howard, G. S. (1987). Construct validity and the relation between depression and social skill. *Journal of Counseling Psychology, 34,* 315-321.

Crandall, B., & Getchell-Reiter, K. (1993). Critical decision method: A technique for eliciting concrete assessment indicators from the "intuition" of NICU nurses. *Advances in Nursing Sciences, 16*(1), 42-51.

deGroot, A. D. (1946/1965). *Thought and choice in chess* (1st ed.). NY: Mouton.

Endsley, M. R. (in press). The role of situation awareness in naturalistic decision making. In C. Zsambok & G. Klein (Eds.), *Naturalistic decision making.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Howard, G. S. (1994). Why do people say nasty things about self-reports? *Journal of Organizational Behavior, 15,* 399-404.

Howard, G. S., Conway, C. G., & Maxwell, S. F. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology, 77,* 187-196.

Kaempf, G., Klinger, D., & Wolf, S. (1994). *Development of decision-centered interventions for airport security checkpoints,* Final Technical Report. Fairborn, OH: Klein Associates Inc. Prepared under contract DTRS-57-93-C-00129 for the U.S. Department of Transportation, Cambridge, MA.

Kaempf, G. L., Wolf, S., Thordsen, M. L., & Klein, G. (1992). *Decisionmaking in the AEGIS combat information center.* Fairborn, OH: Klein Associates Inc. Prepared under contract N66001-90-C-6023 for the Naval Command, Control and Ocean Surveillance Center, San Diego, CA.

Klein, G. (1993). *Naturalistic decision making—Implications for design.* Dayton, OH: CSERIAC.

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fire ground, *Proceedings of the 30th Annual Human Factors Society, 1,* 576-580. Dayton, OH: Human Factors Society.

Klein, G.A., Calderwood, R., & MacGregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(3), 462-472.

Miller, T. E., & Lim, L. S. (1993). *Using knowledge engineering in the development of an expert system to assist targeteers in assessing battle damage and making weapons decisions for hardened-structure targets.* Fairborn, OH: Klein Associates. Prepared under contract DACA39-92-C-0050 for the U.S. Army Engineers CEWES-CT, Vicksburg, MS.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review, 84,* 231-159.

Wolf, S., Hutton, R., Miller, T., & Klein, G. (1995). *Identification of the decision requirments for air defense planning.* Final Report for Litton Data Systems, under PO #TC392277.

Accuracy Estimation in Situation Awareness Research

Richard H. Mogford

NYMA Associates

Introduction

This paper discusses the assessment of situation awareness (SA) accuracy using data recalled or reported by an operator while working with a dynamic system. The domains of interest are aircraft piloting and air traffic control (ATC), although this discussion could apply to a range of occupational areas.

When considering the informational concerns of pilots and air traffic controllers, a typical research issue is the effect of technological or procedural modifications to the work environment on the SA of the operator. A good example is the contemplated change in controller-pilot communication from voice radio to digital data link text messages. Use of a data link system in the cockpit would result in a shift in communication modality from an auditory to a visual display. It has been speculated that this could result in increased pilot "head-down" time, or visual attention to incoming messages, as opposed to "head-up" time, or attention focused out of the window. This shift could be potentially problematic in the approach phase of flight when the pilots are concerned with searching for other aircraft and the runway.

Currently the same radio channel is used by all aircraft sharing a sector of airspace and the resulting "party line" of ATC and aircraft exchanges can be monitored by all those tuned to the frequency. Pilots use this to glean information about weather, the location of other traffic, and emergencies (Midkiff and Hansman, 1992). With the implementation of data link, ATC messages will be directed to specific aircraft and the use of radio, and hence the presence of the party line, will be greatly diminished. Pilots have voiced concern that this will reduce their SA.

On the ATC side, researchers have been interested in how the air traffic controller creates a mental "picture" of the air traffic on the radar display (Whitfield, 1979). Controllers report the picture to be a three-dimensional, geographical representation which assists with the understanding of the ATC situation. It is described as a plan against which aircraft activity is compared and which helps in the detection of conflicts.

Research on the SA of both pilots and controllers is primarily concerned with how much the operator knows about aircraft in a well-defined area. In the case of the aircrew, it is awareness of the location and movement of other aircraft in the immediate vicinity that might pose a collision risk. For the controller, it is the maintenance of the picture of the air traffic situation in order to effectively manage separation. (In both cases, there are other elements, such as equipment status and weather patterns, that are incorporated in SA, but these will not be the focus of this discussion.) Researchers concerned with the SA of pilots or controllers must develop methods to evaluate SA accuracy in order to assess the effects of new technologies or procedures.

Measurement of Situation Awareness

In order to pursue the discussion of accuracy estimation of SA, a brief review of theory and measurement techniques is required. Endsley (1995a) has proposed a useful model of SA and has developed a measurement technique known as the Situation Awareness Global Assessment Technique (SAGAT).

Levels of Measurement

Endsley (1995a) suggested that there are three levels of SA. Level 1 SA consists of the "status, attributes, and dynamics of relevant elements in the environment" (p. 37). A pilot or controller's Level 1 SA could consist of distinct entities such as aircraft, as well as each aircraft's characteristics, such as position, altitude, heading, or speed.

Level 2 SA results from a synthesis of the elements and characteristics from Level 1. It is a comprehension of the current situation, in the context of operator goals. Level 1 data is integrated so that the pilot or controller has an understanding of the air traffic situation and the meaning of aircraft movements. For example, two aircraft traveling in-trail on the same route, with 5 miles of separation would represent a normal and safe traffic flow.

Level 3 SA involves projection of the future status of a Level 2 air traffic situation. For example, an air traffic controller, monitoring two aircraft at the same level, 20 miles apart, on intersecting headings, might determine that a loss of separation will occur, and act accordingly. As Endsley (1995a) observed, comprehensive SA includes detecting relevant facts about the environment, determining their meaning in the immediate situation, and anticipating future developments.

Measurement Techniques

Various methods have been suggested for measuring SA (Endsley, 1995b). Those techniques that seek to sample Level 1 SA may use some type of questionnaire administered after a simulation or actual work session, probe queries designed to tap SA knowledge during task activities, or an SA assessment conducted during a pause in a simulation. SAGAT employs a "freeze technique" that presents questions to an operator during a temporary halt in a simulation run. Typical queries seek recall of SA elements, such as aircraft range, heading, and altitude, although Levels 2 and 3 SA data can also be recorded. Once the SA data are collected, it becomes necessary to devise an accuracy scoring method.

A potential problem emerges when comparing the data reported by an operator to the actual values represented at the time of the simulation interruption. Should a response be counted as correct only if it is exactly the same as the real value? Alternately, is the answer acceptable if it falls within some kind of range? The method adopted for determining accuracy depends upon some basic assumptions regarding SA.

Assessment of Situation Awareness Accuracy

Complete and accurate SA may be the exception rather than the norm. Norman (1983) once commented that mental models "are neither complete nor accurate, but nonetheless they function to guide much human behavior" (p. 46). This could also refer to SA. Errors in SA may occur and lead to undesirable or tragic consequences (Endsley, 1995b). However, a partial and somewhat

inaccurate internal representation of the operator's environment may be sufficient for safe and efficient task completion. In fact, it could be argued that precise maintenance of all possibly relevant SA elements in short- or long-term memory is wasteful of cognitive resources. An operator may prioritize information required for specific tasks and goals and allow less essential data to fade from SA.

Research in air traffic control in which subjects were asked to recall aircraft data suggests a prioritization of information. Means, Mumaw, Roth, Schalger, McWilliams, Gagne, Rice, Rosenthal, and Heon (1988) found that controllers recalled enough information to identify 86% of the aircraft that had flown in a dynamic simulation exercise. Aircraft position was reported with 84% accuracy while altitude and heading were 79% and 80% correct, respectively. Aircraft identifier was recalled with only 24% accuracy. Bisseret (1970) found that highly qualified controllers reported altitude and position information most reliably. Mogford (in review) completed a study in which questions about aircraft altitude, speed, heading, call sign, and position were asked of air traffic control trainees during a pause in a training simulation. The mean accuracy of recall of aircraft information was: position, 86%; heading, 82%; altitude, 73%; identifier, 55%; and speed, 53%.

In the above studies, aircraft speed did not seem to be strongly represented by controllers. Sperandio (1978) observed in his research on workload in ATC that this variable was reported less accurately as workload increased. Although aircraft speed can be important, controllers note that it does not play as critical a role in maintaining separation as altitude.

Given the heavy cognitive demands of ATC, controllers may be selective about the information they maintain in SA. This could reflect their management of limited cognitive resources. It may be sufficient to represent only certain details in SA; further information can be accessed from displays as the need arises. Operators of complex systems may not, therefore, work to retain everything that a task analysis might deem important. They could instead develop a strategy to select information that will improve their performance on high priority tasks. There may then be three kinds of data in such environments: that which must be remembered, that which can be searched for when needed and then forgotten, and that which can be ignored. Only the first type of data needs to be retained in SA.

Endsley (1995a) suggested that some SA of all task-relevant elements must be continuously maintained, but acknowledges that an element's relevance may vary across time. It could be argued that irrelevant SA elements need not be represented. The operator's mental model of the system will determine when such elements should be sampled and brought forward in awareness. Other elements may only need to be maintained to a low degree of precision; a controller or pilot may only have to know if an aircraft is there or not, without recalling all of the details. If this is so, it would then be important to develop methods of determining the quality of Level 1 SA, given that it may be composed of incomplete or inaccurate information about an entity, such as an aircraft of concern to a pilot or controller.

The primary technique for Level 1 SA scoring is to establish tolerance ranges. This assumes, based on the preceding discussion, that insisting on exact SA for all elements is not realistic. A "ballpark" approach is created to evaluate each piece of data. Endsley (1995b) noted that subjects' reports of Level 1 SA elements in a flight simulation experiment were evaluated "using tolerances that had been determined to be operationally relevant" (p. 77). Mogford (in review) used a similar approach in an air traffic control experiment, relying on subject matter experts (SMEs) (in this case ATC instructors) for the creation of tolerance ranges. For example, if there is an aircraft on an ATC radar display at 35,000 feet, a 1000 foot range above and below the actual altitude may be acceptable in terms of an SA representation. However, at a lower altitude, a more limited range may be appropriate, given higher traffic concentrations in this zone.

A second approach is to create a set of rules that would identify the minimum Level 1 SA elements required to uniquely define an aircraft. For example, if two or more letters of the identifier are correct, the altitude is within 1000 feet, and the heading is within 15 degrees, SA is considered to be accurate for that aircraft. It cannot be confused with another target on the display or in the area. This again relies on SME input to help determine the tolerance ranges and rules required. Depending upon the purpose of the SA study, it may be important to focus on SA

elements at the level of attributes of aircraft or to determine if the aircraft themselves have been correctly represented.

Another means to compute SA accuracy is to calculate differences between reported SA element values and actual values. Endsley (1995b) and Marshak, Kuperman, Ramsey, and Wilson (1987) used this approach in SA experiments. However, this technique is only suitable for comparing SA quality between systems or situations and cannot be used for measuring absolute SA accuracy, without the determination of tolerance ranges.

Given that accuracy criteria for Level 1 SA measures need to be determined in many cases, reliable, objective and replicable methods should be developed to define the tolerance limits. In most SA research, these limits are set using a top-down approach. This involves making assumptions about SA requirements (often in consultation with SMEs), and proposing that the operator should be maintaining Level 1 SA information at a certain level of accuracy. However, the research mentioned above suggests that operators may have specific strategies for managing SA that do not involve accurate representation of all elements at all times.

It would be preferable to pursue a bottom-up approach to SA accuracy. What level of representation does a skilled pilot or controller have of aircraft and their attributes? If this question could be answered, SA research would have a yardstick against which to assess operators and systems. If sampling of SA elements indicates decrements in awareness, causes can be sought in training programs, environmental influences, human-computer interface factors, or human performance variables.

A research program of this kind would rely on simulation studies which sample controller or pilot SA at intervals during a set of test runs. A freeze method, such as SAGAT (Endsley, 1995b), would be employed to test selected SA elements at regular intervals. A primary assumption would be that the skilled operators used in the evaluation have, by definition, sufficiently accurate SA to safely and effectively perform their tasks. These workers would be involved in their normal duties, using familiar equipment, and monitoring realistic flight scenarios or air traffic. Therefore, there would be no reason to expect that their SA would be compromised in any way.

With sufficient sampling of SA elements, it would be possible to build distributions of data on such aircraft attributes as location, heading, speed, and altitude. Given this database, future SA studies in the same domain could establish tolerance limits for Level 1 SA data using confidence intervals, or other statistical techniques. It would then be possible to vary from existing conditions and test the effects of manipulating training, environmental conditions, or the user interface. Such an approach would eliminate or at least greatly reduce the need for SME determination of operationally relevant accuracy expectations for SA data.

Conclusions

Measurement of SA is concerned with the assessment of operator knowledge of basic data relevant to the domain of interest (in this case, aircraft), the meaning of the ongoing situation, and projections of future events. This paper proposes a method for objectively setting the accuracy bounds for Level 1 SA in specific environments by sampling skilled operator SA and building a database for comparison to other situations. While this approach has only focused on Level 1 SA, these basic data are the foundation for higher level comprehension. The suggested technique could be extrapolated to study the actual Levels 2 and 3 SA that controllers and pilots maintain during the course of their work.

References

Bisseret, A. (1970). Mémoire opérationelle et structure du travail. *Bulletin de Psychologie,* 24, pp. 280-294.

Endsley, M. R. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors,* 37, pp. 32-64.

Endsley, M. R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors,* 37, pp. 65-84.

Marshak, W. P., Kuperman, G., Ramsey, E. G., and Wilson, D. (1987). Situational Awareness in map displays. *In Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 533-538). Santa Monica, CA: Human Factors Society.

Means, B., Mumaw, R., Roth, C., Schalger, M., McWilliams, E., Gagne, E., Rice, V., Rosenthal, D., & Heon, S. (1988). *ATC training analysis study: Design of the next-generation ATC training system* (OPM Work Order No. 342-036). Richmond, VA: HumRRO International.

Midkiff, A. H., & Hansman, R. J. (1992). Identification of important "partly line" information elements and implications for situational awareness in the datalink environment. In: *Proceedings of the AEROTECH Conference and Exposition.* Anaheim, CA: AEROTECH.

Mogford, R. H. (in review). Mental models and situation awareness in air traffic control. *International Journal of Aviation Psychology.*

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman and S. W. Draper (Eds.), *User Centered System Design* (pp. 31-61). Hillsdale, NJ: Lawrence Erlbaum.

Sperandio, J. (1978). The regulation of working methods as a function of work-load among air traffic controllers. *Ergonomics,* 21, pp. 195-202.

Whitfield, D. (1979). A preliminary study of the air traffic controller's picture. *CATCA Journal,* 11, pp. 19-28.

Factors Characterizing En Rroute Operational Errors: Do They Tell Us Anything About Situation Awareness?

Francis T. Durso[1], Todd R. Truitt[1], Carla A. Hackworth[1], Daryl Ohrt[1], Janis, M. Hamic[1], Jerry M. Crutchfield[1], and Carol A. Manning[2]

[1] University of Oklahoma
[2] F.A.A. Civil Aeromedical Institute

It is the responsibility of the Air Traffic Control Specialist to maintain adequate separation of aircraft, as prescribed by the Air Traffic Control (ATC) Handbook (7110.65J). To perform satisfactorily, controllers must maintain a sufficient level of situation awareness (SA). Poor SA on the part of the controller may result in the loss of separation between two or more aircraft, otherwise known as an operational error. Each time an operational error occurs a lengthy, detailed investigation and report of the incident is completed by the facility concerned.

Operational error reports have previously been analyzed by a number of researchers (e.g., Rodgers & Nye, 1993; Schroeder & Nye, 1993; Redding, 1992). We took a similar approach to analyzing operational errors that occurred during 1993. We analyzed the 412 nonoceanic operational errors reported by air route traffic control centers (specifically en route centers) during calendar year 1993. From these reports, 388 were complete along the variables we considered and formed the database for the current study. An effort was made to determine those factors, if any, that characterized those errors. In particular, we attempted to determine the factors that influenced severity of the operational error. In each case, we used a stepwise multiple regression procedure. To be included in the resulting regression model, a variable had to be significant at the $p < .15$ level.

The primary purpose of the current study was to determine if controllers who differed in their reported awareness of the error also differed in the types of errors they made. We began by trying to predict the severity of the operational error from 1) other characteristics of the error, 2) personnel variables, 3) situation variables, and 4) causal factors. Finally, we explored whether aware controllers and unaware controllers differed in the type of psychological process that went awry during the error. If reported awareness is an unimportant factor, then we should see little difference between aware and unaware controllers.

Predicting Error Severity

Variables characterizing operational errors

To avoid an error in en route airspace, aircraft must be 5 NMnm apart horizontally or 1000 feet (2000 over FL 290) apart vertically. Violations of these minima yield errors of different severity. *Severity*, as used here and elsewhere (Rodgers & Nye, 1993), is the result of a configural rule for combining violations of vertical separation and horizontal separation. Scores range from 0 to 20.

For example, a major error is given 20 points and occurs when aircraft are separated by less than 500 ft vertically *and* less than .5 miles horizontally. Moderate errors have severity scores ranging from 14 to 19. Minor errors have a severity scores of 13 or less. In our 1993 dataset, there were no major errors.

Virtually all of the errors reported were exclusively human (99%), that is did not involve equipment or procedures, involved only one facility (96%), and were not related to procedural deficiencies (98%). The flights involved in the errors we examined were predominantly under radar control (99%) and typically one flight was descending (83%) either through an altitude occupied by a level flight (42%) or through airspace used by a climbing flight (33%).

Sixty-two percent of the controllers were not aware that an operational error was developing, whereas 38% were aware of the error but were unable to rectify it in time. Typically, the errors were ultimately recognized by means of the conflict alert (46%) or were self-identified (40%).

Predicting severity from the set of error cohort variables using stepwise multiple regression, yielded an equation of three predictor variables accounting for 6.8% of the variance in severity: Controllers who were aware that an error was developing made significantly less severe errors (8.93, range 3 - 17) than those who were unaware (10.66, range 3 - 19). The other two variables of the model, each accounted for less than 1% of the variance. Procedural deficiency, an indicator of deficiencies in established procedures, was more likely to be implicated for the more severe errors. Finally, although few incidents involved nonradar flights, those that did yielded significantly more severe errors. Operational errors involving the four nonradar flights (*M*ean severity = 13.25) were more severe than were errors involving radar flights ($M = 9.96$).

Because these variables occurred simultaneously with the error, or could be identified only after the error occurred, they are not of great use in identifying variables that could be changed or modified. However, as statistical predictors, they could be used in considering subsequent analyses. For example, in this report, we take advantage of the large difference between controllers who were aware of the impending problem and those who were not.

Because we were particularly interested in factors that could be indicators of poorer SAsituation awareness, in the following multiple regression analyses, we considered separately the 150 reports for which the controller was aware, and the 238 reports for which the controller was unaware. We did this, in part, because it was impossible to predict *awareness* from variables about the personnel or about the situation. No regression model of personnel variables emerged that allowed us to predict who would be aware of an impending error and who would not. For a situation based model, the type of control procedures involved (e.g., radar, nonradar) and number of persons involved were implicated, but together they accounted for only 1.6% of the variance in awareness. Thus, given the large impact of awareness on severity, and given the difficulty of predicting awareness from other variables existing before the error, we instead analyzed controllers who were aware of the error separately from those who were not. For ease of exposition, we will often call the former group *aware controllers* and the latter group *unaware controllers.*

Personnel variables

Controllers who made operational errors in 1993 were, on the average (median), full performance level controllers (FPLs) for 41 months (0 to 305) who had last been (re)certified 30 months earlier (0 to 293). About half had medical waivers or restrictions (47%). Typically, they had just had a regular day off (RDO) the day before (30%) or two days before (26%) and had returned from a break from controlling traffic 41 minutes before (1 to 395).

Stepwise multiple regressions were used to predict the severity of the error given the personnel variables. For those controllers aware of the impending error, length of time as an FPL was the only contributor to the model, accounting for only 2.8%. The model for unaware controllers accounted for an equally unimpressive 2.4%. Like aware controllers, the longer an unaware controller was an FPL the less severe the error; unlike aware controllers, time since certification was a factor for unaware controllers: The more recently the controller was (re)certified, the less severe the error. This latter finding is somewhat surprising, and perhaps suggests that controllers

who were recently required to go through the recertification process might be somewhat more attentive or more careful. A controller can go through the certification process for several reasons, including switching facilities or areas of specialization. Regardless, the ability to predict the severity of operational errors from personnel variables is disappointingly weak.

Situation variables

Variables characterizing the situation in which the operational error occurred did not suggest that these errors took place in obviously unusual situations. Traffic complexity was average (3, on a scale ranging from 1 to 5) with approximately 8 aircraft (2 to 30) under radar procedures (99%) in positive control airspace (54%). Sectors were usually not combined (74%), nor were positions (56%), no training was in progress (85%), no special procedures were in effect (86%), and national procedures were being followed (99%). More errors (48%) tended to occur near the end of the calendar year (July through November).

Situation variables did not prove good predictors of error severity. For aware controllers, a simple model including only the presence of special procedures emerged from the analysis, accounting for 3.4% of the variance. Errors with special procedures were worse. For unaware controllers, no model emerged from the 11 situation factors that were used to predict severity.

Causal factors

The investigators who filed the operational error report attempted to discern the causal factor(s) responsible for the error. Table 1 lists the causal factors with the percentage of times they were implicated in an error. The most frequently identified causal factor was other inappropriate use of display data (IUDD), which was implicated 44% of the time. This other factor was often something rather uninformative, such as the controller failed to recognize traffic; failed to observe aircraft was 5NM east of J48, and so on. The second most frequent cause, inappropriate use of Mode-C, was implicated 10% of the time.

We began by comparing the proportion of times a factor was implicated for aware and unaware controllers. Only two causal factors were reliably ($p < .05$) different: Aware controllers were more likely to have made an other IUDD error ($\chi^2(1) = 4.96$) than were unaware controllers, although they also made this type of error more than any other. The unaware controllers were more likely than were aware controllers to commit an error involved with readback of altitude ($\chi^2(1) = 6.60$).

Table 1. Percentage of uUnaware and aAware controllers reports implicating each causal factor.

Causal Factors	% Unaware	% Aware
None	2.1	4.7
Computer Entry		
Incorrect input	2.1	0.7
Incorrect update	3.3	0.7
Premature termination of data	1.7	0.0
Other (Explain)	5.0	1.3
Flight Progress Strip		
Not updated	4.2	1.3
Interpreted incorrectly	1.3	6.0

Posted incorrectly	0.4	0.0
Updated incorrectly	1.7	1.3
Premature removal	0.4	0.0
Other (Explain)	0.8	1.3
Radar Display(Misidentification		
Failure to reidentify aircraft when the accepted target identity becomes questionable	0.4	0.0
Overlapping data blocks	2.1	2.7
Other (Explain)	10.0	8.0
Radar Display(Inappropriate Use of Displayed Data		
Mode C	8.3	12.7
BRITE	0.4	2.7
Conflict alert	2.9	1.3
Other (Explain)	39.2	49.3
Communications Error		
Phraseology	4.2	2.7
Transposition	7.1	3.3
Misunderstanding	5.0	2.0
Acknowledgment	8.3	3.3
Other (Explain)	4.6	2.0
Communications Error(Readback		
Altitude	12.5	4.7
Clearance	4.2	4.7
Identification	3.3	0.7
Other (Explain)	0.8	1.3
Coordination(Area of Occurrence		
Inter-position	1.7	0.0
Intra-position	2.5	3.3
Inter-sector	4.6	3.3
Inter-facility	3.3	1.3
Other (Explain)	1.7	0.0
Coordination(Improper Use of Information Exchanged in Coordination		
Aircraft identification	1.3	0.0
Altitudes/Flight level	1.7	0.7
Route of flight	0.8	0.7
APREQS	0.0	0.7
Special instructions	0.4	1.3
Other (Explain)	0.8	0.7
Position Relief Briefing		
Employee did not use position relief checklist	0.8	0.0
Employee being relieved gave incomplete briefing	0.4	0.0
Relieving employee did not make use of pertinent data exchanged at briefing	1.3	0.0
Other (Explain)	1.3	0.7

The difference in the proportion of times a factor was implicated for aware and unaware controllers does not suggest that these factors would be important in predicting the severity of the error. Thus, we again attempted to predict severity, this time from the occurrence of particular causal factors.

Compared to our other attempts to predict severity, the model based on causal factors was better, due in part to the inclusion of several factors. For aware controllers, a six6 variable model

predicted 13% of the variance: Altitude -readback accounted for 3.3% of the variance, interfacility coordination, 3%, and inappropriate use of conflict alert, 2%. Communications transposition, atypical (other) computer entry errors, and overlapping datablocks accounted for the remaining 5%. The unaware controller model was again substantially more complex and had no overlap with the model of the aware controllers. For unaware controllers, nine variables produced a model that accounted for 19% of the variance. From most to least predictive: Altitude readback, incorrect interpretation of flight progress strip (FPS), interposition coordination, special instructions, atypical inappropriate use of displayed data, incorrect computer input, other readback errors, premature removal of FPS, and other area of occurrence coordination errors. Altitude readback errorss accounted for 4% of the variance, incorrect interpretation of FPS accounted for 3.1%; additional variables tended to add 2% or 1% to the model.

Only one factor, altitude readback, was common to both models. Although altitude readback errors occurred significantly more often for unaware controllers, altitude readback errors were prominent in both aware and unaware controller models. For aware controllers, if altitude readback was implicated as a causal factor, severity was 11.6, compared with 8.8 if altitude readback was not implicated. For unaware controllers, severity was 12.6 if the error involved altitude readback and 10.4 otherwise. As Rodgers and Nye (1993) found, altitude readback again proves to be a reliable factor involved in more severe operational errors.

Overall, the regression analyses suggest that if a controller is unaware of the impending error it will impact severity and implicate different factors as the cause of those errors. These analyses also suggest, however, that discovering the precursors to errors in either the personnel or the situation may be difficult. The personnel factors accounted for virtually no variance, and the discovery that more experienced controllers make less severe errors, although comforting, is not illuminating. The situation factor of special-procedures was implicated for aware controllers, and may eventually yield useful interventions, but alone it predicts little. Worse is the ability to use situation factors to predict severity for unaware controllers. None of the 11 situation variables, which included measures of traffic load (e.g., complexity, number of planes), workload (e.g., ongoing training, combined sectors, combined positions), and special situations (e.g., type of procedures, special procedures, type of airspace, ongoing training) were able to illuminate those characteristics of a situation that caused the unaware controller to produce more severe errors.

Content Analysis of Investigators Narrative

The quality assurance specialists who investigate the errors are required to classify each operational error in a number of ways. As part of this classification procedure, they are required to produce a narrative explaining why they believe the error had a human component. This narrative contains much information that cannot be gleaned from other parts of the report. We analyzed these short narratives, looking for statements that implicated psychological factors that might be of use in understanding the controllers SA.

These statements and phrases were then classified along basic cognitive dimensions: attention, perception, memory, and thinking. Statements were classified as attention if the narrative description indicated that the employee was distracted from the task, was or not attentive to the situation, or failed to monitor the situation (e.g., Employee As attention was not focused on Aircraft #1). Perceptual classifications were made if the error was duerelated to information that was occluded, misread, or misheard to a perceptual failure without implicating an attentional failure (e.g., Employee A failed to observe that Aircraft #2 had executed a left turn). Errors were categorized as memorial when poor retrieval or episodic confusions or omissions were implicated as factors. (e.g., Employee A forgot about Aircraft #2), whereas thinking inaccuracies were instances of poor judgment, reasoning, planning, false beliefs, misinterpretations, lack of understanding, or erroneous assumptions (e.g., Employee A misjudged that five miles existed

between Aircraft #1 and Aircraft #2; Employee assumed Aircraft #2 would stay on an easterly heading after deviating around weather). Uninformative narratives were those that provided no insight into the underlying cognitive process and were excluded from analyses; forty-one percent of the aware reports and 24% of the unaware reports were excluded. A chi-square analysis of awareness (i.e., unaware vs. aware) conducted for each of the four psychological classifications revealed a significant dependent relationship. In comparison to aware controllers, unaware controllers made significantly more perceptual, ($\chi^2(13) = 410.181$, $p<.05$ and memory errors ($\chi^2(1) = 6.00$). In contrast, aware controllers committed significantly more thinking errors, $\chi^2(1) = 9.26$. The results are presented in Figure 1. Interestingly, attentional factors were implicated equally often in aware and unaware controllers operational errors. However, memory factors were implicated considerably more often for unaware controllers compared with aware controllers, whereas thinking factors were implicated much more often for aware than unaware controllers. Again, the data seem clearer for aware controllers than for unaware controllers. Aware controllers make a considerable number (56%) of thinking errors, errors of assumption, judgment, and decision making. However, the types of errors committed by unaware controllers are more evenly distributed across the defined psychological categories. Unaware controllers make relatively more memory errors than aware controllers, but in fact the frequency of memory errors is no higher than perceptual or thinking errors. Unaware controllers seem to make all types of errors.

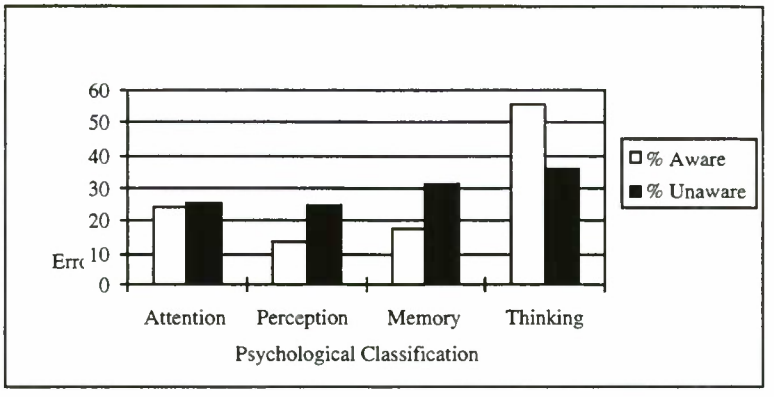


Figure 1. Psychological classification by awareness

Discussion

How might aware controllers differ from unaware controllers? The current work allows some speculation. One possibility is that aware controllers had reasonably good SA, but because of a misjudgment or false assumption, they inadvertently performed one action, when they should have performed another. Given that the only situation variable for aware controllers was the presence of special procedures, it may be that aware controllers incorrectly defaulted to more standard procedures, rather than implementing the special procedures. In a schema view, this would mean that the controller relied on default values in their schema, or default settings in their mental model. This would lead to errors of interpretation and judgment. Although one could speculate in this

way, this type of speculation is premature, due in large part to the lack of any understanding of situation factors influencing the unaware controller.

Some understanding of the difference between aware and unaware controllers comesmay emerge from the analysis of the narratives produced by the quality assurance specialists. A gross differentiation along cognitive dimensions suggested that aware controllers made errors in thinking (i.e., judgments, decision making, and assumptions), whereas the errors made by unaware controllers tended to involve memory. Although aware and unaware controllers were different in the narrative analysis, and although aware controllers tended to make thinking errors, it is again less clear what to make of the data for unaware controllers. It is possible that unaware controllers made more memory and perceptual errors, but it is also possible that unaware controllers made errors across the board, in roughly equal frequencies. Overall, as with some of the other analyses (e.g., situation- based stepwise regression), the characteristic error made by the unaware controller is not readily apparent from the retrospective operational error report. It seems clear that any fine-grained understanding of situation awareness should rely upon on-line measures, rather than retrospective reports.

References

Redding, R. E. (1992). Analysis of operational errors and workload in air traffic control. In *Proceedings of the Human Factors Society 36th Annual Meeting, 2,* 1321-1325. Santa Monica, CA: Human Factors Society.

Rodgers, M. D., & Nye, L. G. (1993). Factors associated with the severity of operational errors at air route traffic control centers. In Rodgers, M. D. (Ed.), *An examination of the operational error database for air route traffic control centers* (DOT/FAA/AM-93/22, pp. 11-25). Washington, D. C.: Office of Aviation Medicine.

Schroeder, D. J., & Nye, L. G. (1993). *An examination of the workload conditions associated with operational errors/deviations at air route traffic control centers* (DOT/FAA/AM-93/22, pp. 1-9). Washington, D. C.: Office of Aviation Medicine.

Acknowledgments

This research was supported by contract # DTFA-02-94-T-80261 from the Federal Aviation Administration to Francis T. Durso. Address correspondence to Frank Durso, Department of Psychology, University of Oklahoma, Norman, OK 73019-0535 or e-mail: fdurso@uoknor.edu.

Situational Awareness at Different Levels of Abstraction: The Distributed Cooperative Problem-Solving Domain of ATCSCC-Airline Operations

C. Elaine McCoy[1], Judith Orasanu[2], Philip J. Smith[3], Amy VanHorn[1], Charles Billings[3], Rebecca Denning[3], Michelle Rodvold[2], and Theresa Gee[1]

1 Ohio University
2 NASA -Ames Research Center
3The Ohio State University

Abstract

Situational awareness is most frequently considered as it pertains to an individual's goals and tasks. Distributed and cooperative problem-solving presupposes that the tasks and goals are not only being considered at the level of the individual's perception, but that the task and goals of the larger task group are being cooperatively addressed. Within the three worlds of flight deck, ATC, and airline operations, individuals shift their efforts to help others achieve success, resulting in a team mentality in which the concern becomes how to help the system work better while optimizing individual goals. The distributed, cooperative problem-solving of ATCSCC-Airline Operations is examined as an illustration of how situational awareness can be maintained at different levels of abstraction within a distributed cooperative, problem-solving domain.

This work has been supported by the FAA Office of the Chief Scientist and Technical Advisor for Human Factors (AAR–1) and NASA Ames Research Center.

Introduction

When moving from examining situational awareness of isolated individuals to studying individuals responsible to a distributed, cooperative problem-solving domain, the awareness of levels of abstraction becomes an important consideration.

In the cooperative problem-solving domain, individuals who have learned the degree of background knowledge and experience of their counterparts develop a sense of trust in these other individuals. Unless a less than satisfactory solution to a problem is provided that conflicts with the assessment of the individual, little additional detailed information is needed or sought. In instances in which the other individual is an unknown or the proposed solution conflicts with the first individual's assessment, then more detailed information is sought relative to the problem. Thus, depending on the circumstances, the necessary level of detail needed to maintain situational awareness will vary.

In the authors' long term study of the aviation system, the relationships of airline operations control (dispatch) and Air Traffic Control System Command Center (ATCSCC) seem to illustrate such shifting levels of abstraction.

ATCSCC-Airline Operations

Airline Operations

Airline operations (dispatch) task characteristics reflect a distributed, cooperative problem-solving environment. Information and data are distributed among a number of individuals. Dispatchers need to coordinate and cooperate with individuals who have differing goals and constraints, including traffic management specialists at ATCSSC and at the Traffic Management Units at the Enroute Centers, flight crew, and other staff within the airline operations center. Airline dispatchers must work cooperatively with other airline and ATC staff who are geographically distributed to accomplish preflight planning and enroute amendments.

Under FAR 121, a Dispatcher and a Captain are jointly responsible "for the preflight planning, delay and dispatch release of a flight." The Dispatcher is also responsible for monitoring the progress of each flight, issuing necessary information for the safety of the flight, and canceling or redispatching a flight if in his/her opinion or in the opinion of the pilot in command, the flight cannot operate or continue to operate safely as planned or released. From an airline management perspective, the Dispatcher is also concerned with factors such as cost, timeliness and passenger comfort.

ATC Coordinators work within the airline operations control centers and are typically experienced Dispatchers who function in a special role as liaisons to ATCSCC and the Enroute Centers.

ATCSCC

ATCSCC is the strategic planning organization for the ATC system, dealing with the airline operations control staff (often through the airline's ATC coordinator) and with the Enroute Centers to plan daily traffic (including replanning flights to deal with weather, airport problems, etc.) ATCSCC has a number of specialist positions for dealing with specific components of this strategic planning, including a position to deal with airline requests for route changes for particular flights.

In the evolution of collaborative airline-ATCSCC communication, certain new procedures have been developed and integrated so as to encourage cooperation. The goal in adopting these procedures has been to improve the efficiency and timeliness of flights, while maintaining or improving safety, thus resulting in lower costs and better service for passengers and cargo delivery. The factors influencing the effectiveness of these new procedures, though, appear to be fairly complex.

Requesting Non–Preferred Routes

The role of situational awareness and the individuals' shifting need for details at differing levels of abstraction in a distributed and cooperative problem solving environment is evident in the evolution of the process by which dispatchers can request non–preferred routes for their flights. McCoy, Smith, Orasanu, Billings, VanHorn, Denning, Rodvold, Gee (1995)

As part of the National Route Program (NRP), many commercial airline flights are assigned a preferred route by ATCSCC although airlines can request alternatives. Traditionally, a somewhat cumbersome procedure requires that requests for non–preferred routes must be submitted to ATCSCC via teletype. The ATCSCC staff member responsible for such requests then contacts the

necessary Enroute Centers by phone to see whether they can accommodate the request. Some requests, or portions of requests, may match a list of non–pref routes that can be automatically approved without contacting the affected Center. If a request for a segment of a route is denied by a Center, that Center may suggest an altemative.

Once all of the affected Centers have been contacted, the ATCSCC staff member informs the ATC Coordinator or Chief Dispatcher at the airline or in some cases an individual Dispatcher who made the request, communicating by phone or teletype regarding its approval, proposed modification or disapproval. The reasons behind a proposed modification or disapproval may or may not be given. Finally, the relevant Dispatcher at the airline must concur with the ATC Coordinator that the approved route is viable.

Current communication by telephone allows for much richer interactions and makes it more likely that personal ties will develop, enhancing cooperation and trust. Explanations can be requested or offered when the need arises for additional detail. One measure of success is financial. One airline reports that it saved $4.3 million in fuel costs in one year: "Last year the non–prefs saved our airline $4.3 million. Our upper management finally came back and said: How can 2 guys in the Dispatch Office save this much money? .... We proved it and they told us to hire another ATC man." McCoy, et al. (1995)

The success of the non-pref route program has been achieved even though the technologies used for this particular program have been rather unsophisticated. The cooperation and communication of the individuals has created a domain of situational awareness that draws from the knowledge and expertise of the individuals who collectively constitute an awareness of parameters that are more global than the parochial concems of the individuals immediate responsibility.

Factors Related to Levels of Abstraction that Help the System Work

Task allocation

Assignment of an ATCSCC staff member to the task of approving or disapproving routes as his or her sole responsibility on a shift is likely to encourage that individual to adopt as a personal goal creating ways to get non–pref routes approved. In addition, because this person is focusing on this one task, he or she is more likely to develop an understanding of the motivations and behaviors of the ATC Coordinators or Dispatchers making requests.

Similarly, assigning ATC Coordinators the task of interacting with ATCSCC makes it more likely that these individuals will develop an understanding of the procedures and constraints that the ATCSCC specialist must deal with. Equally important, because a relatively small number of individuals is involved in direct communications (at ATCSCC and the airlines), the individuals are more likely to develop a stronger interpersonal bond and a sense of shared goals. Trust may be established and maintained that will reduce the need for additional levels of detail.

Distribution of knowledge

To work as an effective task group, certain knowledge must be shared. This question becomes what level of abstraction is needed or desirable. ATC Coordinators who generate non–pref requests are continually in this milieu. Because communications involve discussions of why requests have been rejected, the ATC Coordinators begin to leam what routes are viable as requests. They therefore begin to limit their requests appropriately. One ATC Coordinator commented:

> When we started this, even Central Flow didn't know where all the choke points were. But as we pressed the system and said 'now we want to fly over here', we'd call the Albuquerque Center and they'd say: 'Well, you can't go eastbound over St. John at 4 o'clock in the afternoon'. Well, that was tribal knowledge in the Albuquerque Center. The tribe expanded to include Central Flow; Central Flow expanded the knowledge to the airlines and we began to build better routes. So rather than having to fly a 2000 mile route because it didn't work at one point, we began jogging around and making routes that were smarter.

Distribution of Responsibilities

The distribution of tasks contributes to this successful collaboration as well. Four groups of individuals are directly involved in selecting non–pref routes: staff at the Enroute Centers, the non–pref route specialist on duty at ATCSCC, ATC Coordinators at the airlines and Dispatchers at the airlines. Meteorologists at the airlines and at ATCSCC also provide information.

Since each individual has a different set of primary goals and responsibilities, and makes use of different sources of data, the system provides checks against bad decisions. Local situation assessment comes into play as information is relayed at a higher level of abstraction. Another local situation assessment is occurring with the other party and may conflict. That individual may then interrogate for greater detail. The Dispatcher in charge of a flight, for example, may point out to the ATC Coordinator that the approved non–pref route is questionable in terms of weather. Similarly, the ATC Coordinator may point out that the route proposed by ATCSCC is impossible because of increased fuel requirements. Thus, because tasks, information and workload are distributed (with some redundancy), it is more likely that good solutions will be discussed, and that poor solutions will be detected.

Implications

Situational Awareness is enhanced by understanding the goals of the parties with whom one interacts. The three worlds of flight deck, ATC, and airline operations shift to help others achieve their individual goals. A task-group mentality is developed and the concern expands to how to make the system work better.

In order to work efficiently and effectively as a task group, it is important for various members to understand what others are trying to do, how they are doing it, and how and why they have arrived at particular conclusions. This need for a particular level of abstraction varies.

To study the implications of situational awareness for cooperative problem–solving, a broader conceptual framework needs to be considered. The "situation" must be defined not only by available real–time information, but also by background knowledge held by all participants. Successful cooperation is affected by longer term processes that provide feedback to the system, as well as by immediate interactions. While tasks and information may be distributed, when goals and priorities differ among the participants, there needs to be a shared understanding of the local situations faced by each of the individual participants in order to support cooperation. Interpersonal bonds that develop through communication and experience with other task-group members establishes a differing threshold for the need for detailed knowledge regarding decision-making. As trust and understanding increases so does the level of abstraction at which interactions occur.

Acknowledgements

We would like to express appreciation to Eleana Edens, Larry Cole and Tom McCloy of the FAA, the staff of ATCSCC, the participating Dispatchers, Pilots and Airlines, and the Airline Dispatchers Federation for their assistance.

References

Hayes–Roth, B., and Hayes–Roth, F. (1979). A cognitive model of planning. *Cognitive Science, 3(4),* 275–310.

Hoc, J.M. (1988). *Cognitive psychology of planning.* London: Academic Press.

Layton, C., Smith, P.J., and McCoy, C.E. (1994). Design of a cooperative problem–solving system for enroute flight planning: An empirical evaluation. *Human Factors, 36(1),* 94–119

Lehner, P.E., and Zirk, D.A. (1987). Cognitive factors in user/expert–system interaction. *Human Factors, 29(1),* 97–109.

McCoy, C.E., Smith, P.J., Orasanu, J., Billings, C., VanHorn, A., Denning, R., Rodvold, M. and Gee, T. (In press) Airline dispatch and ATCSCC: A cooperative problem–solving success story with a future. *Proceedings of the Eighth International Symposium on Aviation Psychology,* Columbus, OH: Ohio State University.

McCoy, C.E., KollrossWoleben, J., Smith P.J. (1994) Individual differences in weather situation awareness and assessment. Eds. R.d. Gilson, D.J.Garland, J.M. Koonce. In *situational awareness in complex systems.* (pp. 239-249). Daytona Beach: Embry-Riddle Aeronautical University Press.

Miller, G.A., Galanter, E., and Pribram, K.H. (1960). *Plans and the structure of behavior.* New York: Holts.

Orasanu, J. & Connolly, T. (1993). The reinvention of decision making. In G.A. Klein, J. Orasanu, R. Calderwood, & C.E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 3–20). Norwood, NJ: Ablex Publishers.

Orasanu, J. & Salas, E. (1993). Team decision making in complex environments. In G.A. Klein, J. Orasanu, R. Calderwood, & C.E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 327–345). Norwood, NJ: Ablex Publishers.

Orasanu, J., Wich, M., Fischer, U., Jobe, K., McCoy, C.E., Beatty, R., & Smith, P. (1993). Distributed problem solving by pilots and dispatchers. In R. Jensen (Ed.), *Proceedings of the Seventh International Symposium of Aviation Psychology* (pp. 198–203). Columbus, OH: Ohio State University Press.

Thierauf, R.J. (1988). *User–oriented decision support systems: Accent on problem finding.* Englewood cliffs, NJ: Prentice Hall.

Wilensky, R. (1983). *Planning and understanding: A computational approach to human reasoning.* Reading, MA: Addison–Wesley.

Measurement of Air Traffic Controllers' Situation Awareness and Performance During Simulator Training

Esa M. Rantanen[1] and Joseph S. Butler[2]

[1] The Pennsylvania State University
[2] Embry-Riddle Aeronautical University

Introduction

The concepts of expertise, situation awareness (SA), and performance in air traffic control (ATC) are neither separate or synonymous. Expertise is a general term and of interest in many instances. Air traffic controllers' expertise, on the other hand, is very well represented by the concept of SA. Good SA is essential for safe and efficient performance of the controllers' duties, and expert controllers palpably exhibit better SA than novices. Finally, controllers' performance provides means to assess their SA and expertise through observable actions and errors.

The predominant performance evaluation method today in ATC is the over-the-shoulder (OTS) evaluation by a designated evaluator or instructor. This method has been used since the beginning of organized ATC both in operational and simulated environments. Based on direct observation, the OTS method meets the validity and reliability requirements for performance evaluation for training and proficiency assessment, but it depends heavily on the expertise and skill of the evaluator. The expertise of the observer is critical because the cues on controller performance can be very subtle and may not be sufficiently salient for an observer not familiar with the domain (Olson and Biolsi, 1991). Another prerequisite for efficient usage of this method is a standardized checklist, where all the items to be evaluated must be explicitly defined. The evaluators must also be sufficiently trained to achieve reasonable inter-rater and intra-rater validity. This is particularly important in operational ATC, where there seldom is "just one correct answer" to a problem.

If these prerequisites are met, the OTS method is probably the best currently available performance evaluation technique. Experienced evaluators not only intimately know the appropriate control techniques and "tricks" for the specific requirements of the facility, but also the pitfalls and most common mistakes from personal experience. Thus they may be able to follow the evaluatees' thought processes and detect erroneous planning with very few and subtle cues available to them. The OTS method also captures the "whole" of the task, including the use of aids and equipment, communication and coordination with other controllers and other sectors, and performance on several secondary tasks, which are not part of the controller's primary responsibilities but which must be performed anyway.

The OTS method does, however, have a number of significant disadvantages. First, it is extremely labor intensive, with one-to-one evaluator-evaluatee ratio. Additionally, a human evaluator may not be able to provide sufficiently accurate quantitative data for research purposes, due to the limitations of human observation capabilities. This is the case particularly in observation of simultaneous events. For these reasons there is a need to develop valid and reliable automatic evaluation and performance data collection methods to be used in conjunction with high-fidelity, realistic, ATC simulation.

This paper describes a number of performance measures that are based on automatically collected data and that can be used for assessment of controllers' SA. The approach is decisively different from recent attempts to employ artificial intelligence (AI) techniques to essentially replace

a human instructor. The measures described below are designed to be used on ATC simulators running on personal computers (PC), with limited memory and computing capabilities. Furthermore, merely the data collection is automated, the final responsibility of assessing student controllers' performance remains with a human instructor. Thus, these measures should be viewed as tools that aid the instructor in her task of judging students' performance based on best possible data from simulated exercises.

Definitions

Ericsson & Charness (1994, p. 731, see also Charness, 1988) define expertise as "consistently superior performance on a specified set of representative tasks for the domain that can be administered to any subject". In addition to one particularly important aspect of expertise, i.e., its domain-specificity, this definition highlights the role of performance as a representative of expertise and the requirement of consistency in expert performance, which has implications on the methods of elicitation (naturalistic or laboratory environment, task length, etc.). However, expertise in ATC has a number of unique features which contrast with domains of traditional research on expertise (e.g., chess, music, and physics, see Charness, 1988). The characteristics of ATC environment--time-constrained multiple tasks, complex and dynamic information environment, and teams of operators who need to coordinate their actions to perform the tasks--demand a special form of expertise that not only requires extensive domain-specific knowledge, but also efficient time-sharing skills and problem-solving strategies. These demands imply a need for highly efficient organization of controllers' knowledge structures to facilitate optimum retrieval of information under time pressure (Seamster, Redding, Cannon, Ryder, and Purcell, 1993).

The essence of ATC and the demands placed on controllers are succinctly summarized by the definition of situation awareness as "the perception of elements in the environment within a volume of time and space, their comprehension and meaning, and the projection of their states in the near future" (Endsley, 1994a, p. 31). This definition and another, by Smith and Hancock (1994), which defines SA as adaptive, externally-directed consciousness, establish the relationship between controllers' expertise and their environment and provide guidelines for developing performance measures.

Types of Measurement for Computer-Based Performance Evaluation

Types of suitable measurements can be roughly divided in two main categories: System measures and measures of the controller (Buckley, DeBaryshe, Hitchner, and Kohn, 1983; Hopkin, 1980). System measures are defined in system terms, i.e., capacity, throughput, delays, and channel occupancy times. Although they are greatly influenced by human performance, they are usually insufficient in the measurement of the performance of an individual controller (V. David Hopkin, personal communication, March 1993). Hopkin identifies task performance, human activity, errors, omissions, physiological and biochemical indices, subjective assessment as possible measures of the performance of an individual controller. Task performance measures, which resemble system measures, compare the controller's output to that what is required in the task and encompass broad measures of errors and omissions. Human activity measures passively record what occurs in the task, such as radio transmissions, equipment usage, and communication and coordination with other sectors in terms of times, frequencies, and sequences of the activities. Errors and omissions in ATC domain must also include timing errors. In addition to erroneous

actions, improper timing of correct action must be classified as an error, or much delayed correct action as an omission.

A System Approach

Buckley, et al., (1983) propose a largely system-based task analysis for a basis of an automatic performance evaluation system. Air traffic controllers have a major impact on the fuel consumption of the aircraft operating under their control and their performance in their tasks can be efficiently measured indirectly through the efficiency of the flights conducted within the system. This view is based on the fact that there is an identifiable optimum flight path between two points on earth in terms of time and fuel consumption, and on the assumption that airlines attempt to optimize the use of their aircraft by planning their flights along these parameters. Deviations from the flight plan thus result in less than optimum aircraft performance. It is not always possible to conduct flights along these optimum paths, particularly in the congested airspaces around major airports. Thus, it can be argued that a controller's skill on the task can be measured in terms of the restrictions imposed on individual flights and the magnitude of deviations from their optimum flight paths.

This kind of evaluation system must contain detailed information on the performance characteristics of various different aircraft types. It must also be able to differentiate between the best and less optimal decisions between the types of restrictions for aircraft. For example, it is better to vector a departing aircraft around conflicting traffic than to restrict its climb, and speed restrictions to arriving traffic should take into consideration the aircraft's minimum clean speeds and allow the aircraft to be flown as long as possible in a clean configuration with minimum drag penalty. A good controller should be able to take the aircraft fuel efficiency into consideration when designing traffic flow patterns and issuing restrictions. Other systems measures providing quantitative data in an easily collectible form are a number of aircraft handled in a given time, cumulative delays, number of missed approaches and holdings, number of miles flown by aircraft in excess of their optimum routing, and number of separation violations.

Present-day PC-based ATC simulators are capable of automatic collection of these data. The targets' x,y -coordinates are continuously recorded during a simulation and this data can be used to output aircraft coordinates to a file, e.g., every sweep . This gives sufficient information to graph flight paths of all aircraft in an exercise. Since aircraft performance parameters and traffic flows in a given airspace are known, a three-dimensional best fit line can be developed and compared to the actual flight paths. With enough computational capacity numerous factors affecting aircraft performance, e.g., temperature, altitude, and weight of the plane at each point, can also be taken into consideration. Difference can be measured in feet, both horizontally and vertically, and translated into fuel consumed. While this gives the instructor abundant data to assess student performance in a simulation, the data can also be used to enhance instruction during debriefing sessions. Additional benefits could be achieved it the optimum flight paths are displayed on the radar screen during training and the students would try to match the computer-generated optimum routes with their vectors, significantly enhancing the learning process.

Despite the significant advantages of the system approach such global measures suffer from problems of diagnosticity and sensitivity. They give only the end result of a long string of cognitive processes (Endsley, 1994b) and do not reveal a full picture of a controller's performance and SA. In the definition of the purpose of ATC as the provision of safe, orderly, and efficient flow of air traffic, safety comes first. Although an automated system can detect and tally separation violations, it can not track potential conflicts due to erroneous planning or controller inattention that do not result in separation violations. However, it is imperative that also these types of errors are detected before they result in more serious incidents.

Individual Measures

Human activity measures
Possible individual measures suited for recording by an automated system are human activity measures. The use of available aids and tools can reveal poor performance in the form of underuse, overuse, or misuse. These patterns can be recorded. In ATC, communication patterns are probably one of the best indicators of performance. Unnecessary radio transmissions, repeated "say again" phrases, and frequent revisions to, or reversals of, given clearances indicate poor performance and an incomplete picture of the traffic situation. In contrast, short, concise, and accurate transmissions generally indicate good performance. Words and phrases stated by the controller can be recorded using a computer that is equipped with sound analysis software. The radio transmissions can be grouped by aircraft identifiers and the time in which the transmission was made. Repeated voice patterns can be identified and keywords indicating hesitation and poor SA noted. The voice data would be saved to an audio file and the voice wave patterns analyzed by existing voice software. Because manual analysis of these would be too time-consuming, voice analysis should be highly automated.

Withheld information.
A powerful technique to externalize data acquisition and monitoring behavior during a problem-solving situation is to withhold critical information until the problem-solver specifically requests it (Woods, 1993). The time and order in which missing pieces of information are requested (or not requested) can be used to reconstruct the subject's mental model and problem-solving sequence. The beauty of this technique is that such situations are an intrinsic part of air traffic controllers' jobs, and that the caveat Woods (1993, p. 233) makes about introducing a mismatch between the test behavioral situation and the target situation need not apply in this domain. In experimental conditions withheld information can be included in simulations camouflaged as "pop-up" targets, i.e., aircraft that appear on the frequency and request service but do not have a flight plan. The controller must thus verbally request all pertinent information from the pilot of the aircraft to determine whether the aircraft will cause a potential conflict within her area of responsibility and what would be a most efficient way to handle the pilot's request. If the aircraft is placed in the simulation correctly, controller's inquiries for information and her subsequent actions based on it will provide a clear "snapshot" of her SA at the moment.

The momentary traffic situation can be saved at any given moment during a simulation to a file in a pcx -format. The situation can also be configured to any graphics format, such as gif or jpeg. This snapshot can be later analyzed by the instructor together with a voice transcript as described in the previous section. Subsequent snapshots would show what actually happened and reveal the consequences of the controller's actions.

Recall of information.
Mogford (1994) found significant correlation between controller trainees' ability to recall key aircraft parameters in an interrupted simulated scenario and the subjects' scores in the final simulator exam. These key parameters were aircraft altitude, heading, speed, and position, in that order, which all are relevant in determining whether two aircraft are on a potential collision course. In addition to recall accuracy, grouping of aircraft recalled provides an indication of controller's SA. Efficient grouping of aircraft according to their flight paths and profiles allows an expert controller work more traffic, formulate a better sector plan, and use fewer control actions to maintain the traffic flow and control in the sector (Seamster et al., 1993).

While intuitively attractive and feasible for automatic data collection, assessment of a person's SA through the accuracy of her recall and description of the situation should be approached with caution. Since SA resides in controllers' working memory (WM) it is subject to the capacity and decay limitations of human memory system, and particularly those of WM. Additionally, running memory, which is responsible for updating dynamic information in WM, has even smaller capacity than that of the static WM (Moray, 1986; Yntema, 1963) Thus, subjects' performance cannot be expected to be very good in a recall task, and the quality of data is heavily dependent on the

accuracy criteria used. This, in turn, will result in sensitivity problems. Additionally, the recall technique is very intrusive as it necessitates interruption of the task at some critical moment.

Secondary task measures.
Secondary task measures have a long history in human workload research. The underlying paradigm in these measures is the assumption that secondary task performance is inversely proportional to the primary task resource demands (Wickens, 1992). The fact that makes these measures inherently suitable for elicitation of air traffic controllers' expertise is that a number of secondary tasks are an intrinsic part of the total task, and the most common drawback of the technique, its intrusiveness, can thus be mitigated. Since it is important to evaluate controllers' performance also in relation to other controllers' performance in neighboring sectors, hand-off procedures provide a useful set of secondary tasks to detect controllers' SA and performance in coordinating the traffic with other sectors. Items to be measured include the point of initiation of the hand-off, total time elapsed during the hand-off, and whether the aircraft being handed off is within the confines of the appropriate transition area. On the other hand, also the time elapsed in accepting a hand-off from another sector must be measured. Any omissions or delays in performance of these tasks indicate degraded SA.

When a hand-off is initiated the computer starts counting the seconds until the hand-off is accepted. Because the target's x,y -coordinates are continuously recorded, the computer also takes into consideration the aircraft's position relative to the appropriate transition area. The transition areas are defined in the simulation based on letters of agreement and the aircraft position and altitude can be checked against these criteria. Deviations from the confines of the transition areas and the elapsed time of a hand-off is recorded and used to detect lapses in controllers' SA.

Response time measures.
The expert controller's dynamic picture provides not only an efficient knowledge structure but also facilitates rapid retrieval of information relevant to sector control. Therefore, response times to various events within the sector can be used as a measure for the controller's current SA and level of expertise (Seamster et al., 1993). Level of controllers' SA is probed by placing various queries in simulations of varying workload and complexity and by measuring the controllers' response time to them. Again, these queries can be masked as normal control tasks, such as replying to requests for information from other controllers at other sectors or relaying non-tactical information to pilots. Recording the elapsed time between controller and pilot responses is also possible on the simulator. The elapsed time can be included in the voice output file and extracted later for separate analysis.

Validation and Verification of the Measures

Measures of controller SA through performance indices must, like all measures, be valid and have a scale. While collection of quantitative performance indices is not overly difficult, special attention must be paid to the validity of such measures. Furthermore, controller performance is very dependent on a situation, true to the very definition of SA (Smith and Hancock, 1994). The situational demands must therefore be known as well and controller performance assessed against them. A thorough task analysis is thus the most critical prerequisite for any type of performance assessment, and automated performance evaluation is no exception. Unless the tasks in which a person's performance is to be evaluated are identified, their criticality assessed, and the optimum performance in them defined, performance evaluation will lack a scale, it may not measure the right things, or it may not weigh the performance in the tasks according to the real demands of the job. Air traffic controllers' job is, however, complex and it consists of a vast number of tasks and subtasks. Therefore, the measures described earlier are based on a collection of representative

situations in simulated exercises. This will reduce the otherwise forbidding variability in the operational ATC to a manageable level and allows formulation of optimum solutions to problems with a finite number of known variables.

Since SA, or the underlying mental models, are the cornerstone of a human controllers' performance, identification of these models is critical for the study of performance in complex environments (Hahn, 1988). Mental models are commonly measured through the use of verbal protocols given by the subjects and subjective reports regarding the performance shaping factors and workload. Since controllers have to verbally communicate clearances and instructions to aircraft as well as coordinate with controllers of other sectors, it is essential for them to be able to articulate their goals, intentions, and needs of additional information. This fact can be used directly to elicit information on controllers' mental models and problem-solving strategies via thinking-aloud protocols, retrospective and cued verbal reports, and records of verbal communications via formal communications media (Woods, 1993). Experienced controllers are very well capable of commenting on their own or their peers performance on videotaped sessions or articulating their strategies when solving specific problems on paper and pencil tests.

This appears to be the only practical way to validate and to develop appropriate scales for the measures described earlier. The optimum solutions to simulated problems are thus based on expert controllers' performance and consensus on a best strategy in each situation. Thus, an expert controller "profile" forms the scale on which the trainee's performance is measured. However, since there are several methods to analyze the performance data and these methods and the extent of automation used in the analysis is left to the discretion of the instructor, these scales are flexible and can be modified according to the trainees' progress and individual facilities. Thus, the imstructor will always have the final control over the inferences made from the data.

Conclusion

The multidimensionality of air traffic controllers' tasks, in addition to the sheer number of relevant tasks and subtasks, make comprehensive performance evaluation an extremely complex subject. The complexity of the topic is heightened when the performance indicators are considered. None of these can be used alone, as a comprehensive measure that could be recorded in quantifiable form, and that could give an indication of the controller's SA and level of expertise. Therefore, valid and reliable assessment of air traffic controllers' expertise must combine data from many sources, from automated systems, human observations, and subjective reports. This task is left to the instructor. Automatic performance data collection enhances and expands the traditional OTS method by providing the instructor a "hard copy" of the student's performance in the simulation and alleviating the time constraints and information overload often associated with direct observation. Furthermore, quantitative performance data can be subjected to statistical analyses to reveal underlying patterns in students' performance.

References

Buckley, E. P., DeBaryshe, B. D., Hitchner, N., & Kohn, P. (1983). *Methods and measurements in real-time air traffic control system simulation.* (Report No. DOT/FAA/CT-83/26) . Atlantic City, New Jersey: FAA Technical Center

Charness, N. (1988). Expertise in chess, music, and physics: A cognitive perspective. In L. K. Obler and D. Fein (Eds.), *The Exceptional Brain: Neuropsychology of Talent and Special Abilities*, (pp. 399-426). New York: The Guilford Press.

Endsley, M. R. (1994a). Situation awareness in dynamic human decision making: Theory. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 27-58) Orlando, Florida: University of Central Florida, Center for Applied Human Factors in Aviation.

Endsley, M. R. (1994b). Situation awareness in dynamic human decision making: Measurement. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 79-97) Orlando, Florida: University of Central Florida, Center for Applied Human Factors in Aviation.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*, 724-747.

Hahn, H. A. (1988). Model for measuring complex performance in an aviation environment. *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 875-878).

Hopkin, V. D. (1980). The measurement of the air traffic controller. *Human Factors, 22*(5), 547-560.

Hopkin, V. D. (1994). Situational awareness in air traffic control. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems*, (pp. 171-178). Orlando, Florida: University of Central Florida, Center for Applied Human Factors in Aviation.

Mogford, R. H. (1994). Mental models and situation awareness in air traffic control. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems*, (pp. 171-178). Orlando, Florida: University of Central Florida, Center for Applied Human Factors in Aviation.

Moray, N. (1986). Monitoring behavior and supervisory control. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance: Volume II, Cognitive processes and performance* (pp. 40-1 - 40-51). New York: Wiley Interscience.

Olson, J. R. & Biolsi, K. J. (1991). Techniques for representing expert knowledge. In K. A. Ericsson and J. Smith (Eds.), *Toward a General Theory of Expertise: Prospects and Limits* (pp. 240-264). New York: Cambridge University Press.

Seamster, T. L., Redding, R. E., Cannon, J. R., Ryder, J. M., and Purcell, J. A. (1993). Cognitive task analysis of expertise in air traffic control. *The International Journal of Aviation Psychology, 3*(4), 257-283.

Smith, K., & Hancock, P. A. (1994). Situation awareness is adaptive, externally-directed consciousness. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems*. Orlando, Florida: University of Central Florida, Center for Applied Human Factors in Aviation.

Wickens, C. D. (1992). *Engineering psychology and human performance*. New York, NY: HarperCollins.

Woods, D. D. (1993). Process tracing methods for the study of cognition outside the experimental psychology laboratory. In G. Klein, R. Calderwood, and J. Orasanu (Eds.), *Decision-making in Action: Models and Methods* (pp. 228-251). Norwood, NJ: Ablex.

Yntema, D. B. (1963). Keeping track of several things at once. *Human Factors, 6*, 7-17.

A Simulation Study of Air Traffic Controller Situational Awareness

Randy L. Sollenberger[1] **& Earl S. Stein**[2]

[1] Princeton Economic Research, Inc.
[2] DOT/FAA Technical Center

Introduction

Controllers are required to process vast amounts of information in order to safely and expeditiously conduct air traffic. Since most air traffic control (ATC) information is constantly changing, working memory and situational awareness (SA) are critically important. Human working memory has a limited capacity and is often thought to be a contributing factor to operational errors in the ATC system (Federal Aviation Administration, 1987). One approach to reducing the incidence of errors is to enhance working memory by providing memory aids to controllers. The present research is part of a Federal Aviation Administration (FAA) program to develop memory aids intended to support controller working memory and improve controller SA. This paper describes the development of a technique used to assess controller SA and its application in an ATC simulation study evaluating potential memory aids.

The present research draws its theoretical foundation and methodology from Endsley's (1988, 1989, 1990) SA research. According to Endsley's (1988) model, SA can be described by three hierarchical levels. The first level is the perception of the elements in the environment. The second and third levels go beyond simple perception of the elements and involve the comprehension of their meaning and the projection of their status in the near future. In ATC, it seems reasonable that the relevant elements are aircraft and associated flight data. The controller's memory for this basic information forms the foundation for the second and third levels of SA (Fraker, 1989; Mogford, 1994).

The Controller Memory Enhancement Project began with a field study which interviewed air traffic controllers from facilities nationwide to determine the techniques which are currently being used to support controller memory. Controllers suggested several simple activities that they always use on the job. Also, controllers described a few innovative techniques for using their standard workstation equipment as memory-aiding devices. Studies using college students and a low-fidelity simulator that required participants to enter their control instructions with the keyboard of a personal computer were conducted (Zingale, Gromelski, & Stein, 1992). In later studies, actual air traffic controllers were used with the same low-fidelity simulator (Zingale, Gromelski, Ahmed, & Stein, 1993). Preplanning and flight strip marking were tested as potential memory aids. In general, the memory aids were not useful given new controller participants. The research clearly indicated the need to use professional controllers and more realistic simulation equipment in the future.

In the present study, actual controllers used a new, high-fidelity simulation capability developed in the Human Factors Laboratory at the FAA Technical Center in Atlantic City, New Jersey. The memory aids which were evaluated were specially-designed arrival and departure procedures. Similar to Standard Terminal Arrival Routes (STARs) and Standard Instrument Departures (SIDs), the experimental procedures were preplanned flight instructions used by pilots to navigate through the terminal environment. Potentially, the arrival and departure procedures could improve SA for

air traffic controllers in several ways. First, it was expected that controllers would need to spend less time exchanging communications with pilots. Therefore, more time could be devoted to scanning the radar display, reviewing flight strips, and performing other activities that should increase SA. Second, once the arrival and departure procedures have been learned, they could serve as a well-developed and highly-structured mental model of traffic flow in the sector. The mental model can help to organize aircraft information and reduce the burden on working memory (Endsley, 1990; Rantanen, 1994).

Method

Participants

Sixteen air traffic controllers from Atlantic City TRACON participated in the experiment. All participants were full performance level controllers and had current radar experience. The controllers ranged in age from 24 to 40 years old (mean = 33.13) and ranged in experience from 1 to 20 years of active service (mean = 7.72). Each controller participated in the experiment individually.

Simulation Facility

The experimental apparatus consisted of a state-of-the-art controller workstation with voice communication equipment and ATCoach simulation software (copyright UFA Inc., 1992). The experiment was conducted by a research psychologist and an air traffic control specialist (ATCS) observing the participant in one room. A voice communication link to another room allowed the controller to issue commands to personnel serving as simulation pilots. Two trained simulation pilots provided voice feedback to the controller and directed aircraft movements using simple keyboard commands. They also served as ghost controllers performing the tasks of other controllers in the simulation.

Experimental Design

The main independent variable of the experiment will be referred to as the MEMORY factor. This manipulation required that controllers perform eight different scenarios over 2 days of testing. On one of the days, controllers used their own techniques without any special instructions from the experimenters. On the other day, participants used the specially-designed arrival and departure procedures as memory aids while controlling traffic. The second independent variable will be referred to as the TRAFFIC factor. This manipulation involved constructing four scenarios with a low volume and four scenarios with a high volume of air traffic. Low traffic scenarios consisted of 14 aircraft appearing within 30 minutes and high traffic scenarios consisted of 23 aircraft appearing within 30 minutes. The experimental design can be summarized as a 2 x 2 repeated measures design with the factors MEMORY (no memory aids, memory aids) and TRAFFIC (low, high).

Procedure

When controllers arrived at the laboratory, they were informed of the experimental procedures and response format. On each day of testing, controllers worked one practice scenario followed by

four 30-minute test scenarios with rest periods between sessions. Half of the participants worked scenarios without the memory aids on the first day and with the memory aids on the second day. The other half worked scenarios with the memory aids on the first day and without the memory aids on the second day. A training program was developed to assist controllers in learning the specially-designed arrival and departure procedures in the experiment.

Controller performance was assessed by collecting a large data base of system information every second of the simulation. After the experiment was completed, the raw data for each scenario was condensed and used to compute a list of common ATC performance measures (Buckley, DeBaryshe, Hitchner, & Kohn, 1983). The most important variables for the purpose of this report are the number of separation errors, handoff errors, and controller transmissions. In addition to these objective measures, controllers provided self-ratings of performance and SA after each scenario. Also, an ATCS observing the simulation provided independent performance ratings. These subjective ratings were on a scale from 1 (poor performance or low SA) to 10 (good performance or high SA).

The SA assessment technique consisted of freezing the simulation and having participants answer questions about the current situation without viewing displays or flight progress strips. Each scenario was paused randomly once between 17 and 28 minutes from the start and again between 30 and 35 minutes. The scenario ended immediately after the second pause. Controllers were not informed when the scenario would be frozen. During the first pause, three aircraft were randomly selected from the radar display and controllers responded to a series of questions for each aircraft that was not handed off yet. The questions were presented on a laptop computer, and the format consisted of an aircraft call sign and a question about the aircraft. Controllers were requested to respond with the aircraft's current altitude, speed, and heading and the aircraft's most recently assigned altitude, speed, and heading. During the second pause, eight aircraft were randomly selected from the radar display, and controllers were asked to locate each aircraft that was not handed off yet. The procedure consisted of placing the number associated with each aircraft's call sign in the proper location on a paper map of the radar display.

A special scoring procedure was developed which awarded different points for each question depending upon the accuracy of the response. This scoring system differs from that used by Endsley and her colleagues and was designed to improve SA measurement sensitivity. For questions requiring a numeric response and the aircraft location task, three different point-scoring ranges were defined. If the response was within the closest range of accuracy, three points were awarded. If the response was within the middle range, two points were awarded. If the response was within the outer range, one point was awarded. A response beyond the outer range was not awarded any points. Hit (3 points) or miss (0 points) scoring was used if the participant responded with a fix name or indicated that no assignment was made. Table 1 shows the actual questions and the accuracy levels used to define the different scoring ranges for each question. A percentage score for each of the questions and aircraft location task was calculated by dividing the number of points that was actually obtained by the number of points that could have been obtained and multiplying by 100.

The simulation software recorded the correct answers to the questions in a data base which was compared to the controllers' responses after the experiment was completed. The aircraft location data were manually scored by obtaining a graphics snapshot of the radar display at the time the scenario was frozen. The snapshot was a hard copy that was spatially identical to the map of the radar display given to controllers for locating the aircraft. The snapshot and the map were overlaid, and measurements were taken to determine the discrepancy between the actual aircraft locations and the controllers' placements.

Controller workload was assessed using the Air Traffic Workload Input Technique (ATWIT). ATWIT is an unobtrusive and reliable method for determining participants' workload levels as they control traffic (Stein, 1985). In the present study, a touchscreen was used to present the workload rating scale and to record the controllers' responses. Controllers were instructed to indicate their current workload level by pressing one of the touchscreen buttons labeled from 1 (low workload) to 10 (high workload). The touchscreen was programmed to request a rating every five minutes.

Table 1. Accuracy Levels Defining the Different Scoring Ranges for the SA Questions

Question	3 Points	2 Points	1 Point
1-What is the current altitude of the aircraft?	± 1000 ft	± 2000 ft	± 3000 ft
2-What is the current airspeed of the aircraft?	± 20 kn	± 40 kn	± 60 kn
3-What is the current heading/fix of the aircraft?*	± 10 deg	± 20 deg	± 30 deg
4-What was the most recently assigned altitude for the aircraft?*	± 1000 ft	± 2000 ft	± 3000 ft
5-What was the most recently assigned airspeed for the aircraft?*	± 20 kn	± 40 kn	± 60 kn
6-What was the most recently assigned heading/fix for the aircraft?*	± 10 deg	± 20 deg	± 30 deg
7-Aircraft Location Task	5 nm	10 nm	15 nm
* Hit or miss scoring was used if the participant responded with a fix or indicated that no assignment was made			

Results

Table 2 shows the means for the experimental factors and the results of the analysis of variance (ANOVA) conducted on the data. In terms of ATC performance, the memory aids reduced the number of controller transmissions and handoff errors. However, there was no effect on the number of separation errors and performance ratings. Controllers recalled less information about current altitude and more information about assigned heading when using the memory aids. Also, controllers rated their SA as slightly higher when using the memory aids. The results indicated that the memory aids did not influence controllers' workload ratings. Finally, traffic volume had large effects on performance, SA, and workload. Recall of current aircraft information was better and controllers' SA ratings were higher during low traffic conditions. However, traffic volume had little effect on recall of controllers' assignments.

A correlation analysis was conducted to examine the relationship between ATC performance and controllers' recall of aircraft information. Although, overall, the correlation coefficients were rather low, some of the stronger relationships will be reported in this section. The number of separation errors was inversely related to the recall of assigned heading ($r = -0.34$) and aircraft location ($r = -0.31$). The number of handoff errors was inversely related to the recall of current altitude ($r = -0.35$), current heading ($r = -0.25$), and assigned altitude ($r = -0.35$). Controllers' performance ratings were directly related to the recall of current altitude ($r = 0.25$) and aircraft location ($r = 0.29$). Lastly, the observer's performance ratings were directly related to the recall of aircraft location ($r = 0.29$). These correlations served as indicators but were not definitive due to their small magnitude.

Table 2. Means for the Memory and Traffic Conditions of the Experiment

Dependent Measure	No Memory Aids	Memory Aids	Low Traffic	High Traffic
Number of Transmissions	48.75	vs 39.30*	32.73	vs 55.63*
Number of Separation Errors	0.19	vs 0.25	0.03	vs 0.41*
Number of Handoff Errors	0.09	vs 0.03*	0.08	vs 0.05
Current Altitude	61.49	vs 56.29*	63.52	vs 53.85
Current Speed	57.44	vs 64.05	69.81	vs 49.89*
Current Heading/Fix	61.30	vs 55.03	65.00	vs 50.64*
Assigned Altitude	66.95	vs 74.21	68.98	vs 72.01
Assigned Speed	99.15	vs 97.17	98.33	vs 98.08
Assigned Heading/Fix	63.84	vs 80.19*	72.50	vs 70.51
Aircraft Location Task	69.00	vs 62.29	71.60	vs 59.70*

Controller's Performance Rating	7.77	vs 7.69	8.48	vs 6.97*
Observer's Performance Rating	7.91	vs 8.08	8.92	vs 7.04*
Controller's Workload Rating	3.46	vs 3.48	2.34	vs 4.60*
Controller's SA Rating	7.61	vs 7.98*	8.56	vs 7.03*
* Indicates that means are significantly different ($p < 0.05$)				

Discussion

As expected, the memory aids greatly reduced the number of controller transmissions, but there were few other improvements in ATC performance. The number of handoff errors was slightly reduced, however, there was no change in the number of separation errors. Also, the observer's ratings and the participants' self-ratings indicated that the memory aids did not improve performance.

The results of the information recalled by controllers may explain why the memory aids did not make more improvements to ATC performance. Contrary to expectations, controllers recalled significantly less about current altitude and slightly less about current heading and aircraft location when using the memory aids. Current altitude, heading, and aircraft location are some of the most important sources of information for controllers to be aware of as they perform their duties. Although the results indicated at least slightly better recall of the assigned altitude and heading when using the memory aids, this information is less important and can be easily explained. In the memory aids scenarios, arriving and departing aircraft followed preplanned flight paths and there was usually no need to issue any control instructions. Therefore, it was not difficult for controllers to remember or guess correctly that no assignments were made. In the no memory aids scenarios, altitude and heading assignments were always necessary. It would have been difficult to guess correctly since the assignments depended upon the situation. Also, controllers' SA ratings were slightly higher when using the memory aids, but this may simply be a demand characteristic of the experiment. Since participants were well-aware of the purpose of the study and the experimental conditions of each scenario, their ratings may have been unintentionally influenced by the expectations of the researchers and by controllers' own expectations.

One possible reason why potential effectiveness of the memory aids was reduced may be because controllers were not able to easily remember the new arrival and departure procedures. Several participants' comments during debriefing were consistent with this explanation. Also, controllers' workload ratings suggested that the memory aids required additional mental effort to use. Although the memory aids greatly reduced the number of controller transmissions, workload ratings remained unchanged. Most controllers agree that communications are a major source of workload. However, it is possible that trying to remember the new arrival and departure procedures increased mental workload. The decrease in communications workload and the increase in mental workload may have had a canceling effect that led to no change in controllers' overall workload ratings. Also, the memory aids may have influenced controller planning.

Planning control actions and making transmissions help controllers maintain their SA. The memory aids can be thought of as a form of automation where certain tasks were done for controllers without the need for any action. Although automation is more commonly thought of as a function of computers, simulation pilots performed the duties in the present experiment. There are always potential risks associated with automation in that the human operator may become more passive, less aware, and less able to respond when control actions are required (Hopkin, 1994). Indeed, a few controllers' comments suggested that they may have experienced these incidental effects while working the scenarios.

The results of the present research address the relationship between recall of aircraft information and ATC performance. Although the correlations were not very strong, some types of information were more important to performance than others. Recall of aircraft location emerged as a central

factor and was moderately related to most of the performance measures. Current altitude was also important and, to a lesser extent, heading was also relevant. Recall of controllers' instructions to pilots was generally not important and aircraft speed was not related to any of the performance measures. Previous research has also demonstrated that recall of aircraft altitude and heading are related to ATC performance (Mogford, 1994). In addition, the present study has indicated that aircraft location is important as well.

There were a few potential problems in the present SA technique that should be considered for future research. First, the procedure greatly depended on controllers' memory for aircraft call signs as a retrieval cue for other information. However, many participants stated that it is not necessary for controllers to remember aircraft call signs to perform their jobs. They simply read the call signs from flight strips or the radar display when communications are necessary. The retrieval cues of the present procedure may not have been completely valid and could have underestimated controllers' SA. Flight strips contain aircraft call signs and flight plan data and may be more realistic and potent memory cues. Alternatively, a radar display with only the call signs and target locations visible may be a more natural retrieval cue for other aircraft information.

Another aspect of the present SA technique was that each scenario was paused only twice to assess participants' awareness. It is possible that controllers' SA changed over the course of the scenario depending upon traffic and other conditions. If so, the data collected during the memory probes may not have accurately represented controllers' SA for the entire scenario. This is a methodological weakness that will persist with the use of the freeze and query technique. The performance measurements obtained in the study, however, were based upon overall performance. This may explain the relatively low correlations obtained between recall of aircraft information and ATC performance. The solution to this problem is not easy. Sampling memory more frequently may improve SA measurements, but too many pauses may be disruptive. Although some research has investigated these issues in the aircraft cockpit, more SA work is needed in ATC (Endsley, 1990).

The present study investigated potential controller memory aids using the freeze and query technique and a novel scoring procedure designed to achieve a sensitive measure of SA. Although the memory aids did not improve ATC performance as much as was expected, the study produced several important findings for the assessment of controller SA. Using recall of aircraft information as an indicator of SA, the results showed that SA was better at low traffic levels relative to high traffic levels. Also, recall of aircraft location, altitude and heading were correlated with several different measures of ATC performance. Finally, it was discovered that aircraft call signs were poorly remembered by controllers and should be avoided as retrieval cues when assessing controller SA.

Efforts are underway to upgrade the current ATC system with more modern equipment. Therefore, it is important to develop valid and reliable techniques for assessing controller SA and to employ these methods when evaluating proposed modifications to the system. Since it is likely that controllers will still play a major role in the new ATC system, it will be important to implement changes that will maintain or increase controller SA.

References

Buckley, E. P., DeBaryshe, B. D., Hitchner, N., & Kohn, P. (1983). *Methods and measurements in real-time air traffic control system simulation* (DOT/FAA/CT83/26). Atlantic City, NJ: DOT/FAA Technical Center.

Endsley, M. R. (1988). *A construct and its measurement: The functioning and evaluation of pilot situation awareness* (NOR DOC 88-33). Hawthorne, CA: Northrop Corporation.

Endsley, M. R. (1989). A methodology for the objective measurement of situation awareness. In *Situational Awareness in Aerospace Operations*. Copenhagen, Denmark: NATO-AGARD.

Endsley, M. R. (1990). *Situation awareness in dynamic human decision making: Theory and measurement* (NOR DOC 90-49). Hawthorne, CA: Northrop Corporation.

Federal Aviation Administration (1987). *Profile of operational errors in the national airspace system, calendar year 1986.* Washington, DC.

Fraker, M. L. (1989). A theory of situation assessment: Implications for measuring situation awareness. *Proceedings of the Human Factors Society 52nd Annual Meeting.* Santa Monica, CA: Human Factors Society.

Gromelski, S., Davidson, L., & Stein, E. S. (1992). *Controller memory enhancement: Field facility concepts and techniques* (DOT/FAA/CT-TN92/7). Atlantic City, NJ: DOT/FAA Technical Center.

Hopkin, V. D. (1994). Situation awareness in air traffic control. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.). *Situational Awareness in Complex Systems.* Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Mogford, R. H. (1994). Mental models and situation awareness in air traffic control. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.). *Situational Awareness in Complex Systems.* Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Rantanen, E. (1994). The role of dynamic memory in air traffic controllers' situation awareness. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.). *Situational Awareness in Complex Systems.* Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Stein, E. S. (1985). *Air traffic controller workload: An examination of workload probe* (DOT/FAA/CT-TN84/24). Atlantic City, NJ: DOT/FAA Technical Center.

UFA, Inc. (1992). *ATCoach* [Computer software]. Lexington, MA: UFA, Inc.

Zingale, C., Gromelski, S., & Stein, E. S. (1992). *Preliminary studies of planning and flight strip use as air traffic controller memory aids* (DOT/FAA/CT-TN92/22). Atlantic City, NJ: DOT/FAA Technical Center.

Zingale, C., Gromelski, S., Ahmed, B., & Stein, E. S. (1993). *Influence of individual experience and flight strips on air traffic controller memory/situational awareness* (DOT/FAA/CT-TN93/31). Atlantic City, NJ: DOT/FAA Technical Center.

Construct Validity of Situation Awareness Measurements Related to Display Design

Oscar Olmos, Chia-Chin Liang, and Christopher D. Wickens

University of Illinois at Urbana-Champaign

Abstract

Shortcomings in a pilots hazard awareness are often linked to aircraft accidents or incidents involving penetration into hazardous weather, traffic conflicts, or controlled flight into the surrounding terrain. This lack of situation awareness can be alleviated through the use of cockpit displays that are designed to help support good hazard awareness. As a result, implicit and explicit measurements of how well cockpit displays support this type of awareness are vital to the design of systems that will support a high level of SA in the cockpit. In this study, we review the results of previous studies that we have carried out in our lab as related to display design. From this review a list was compiled of the various measurements that we have adopted to measure SA. The list was presented to commercially-qualified pilots for an evaluation of the construct validity of the various SA metrics. The results of this pilot evaluation are examined in terms of the advantages and disadvantages of the various measurements of situation awareness.

Introduction

Within the field of aviation there has been an increased focus on situation awareness as an essential ingredient to safe and efficient flight operations (Endsley, 1993, 1995; Adams, Tenney, and Pew, 1995; Vidulich, Dominguez, Vogel, and McMillan, 1994). While various definitions of situation awareness exist within the aviation community we propose the following "consensus" definition offered by Wickens (1995) as a foundation for the remaining discussion:

> Situation awareness is the continuous extraction of information about a dynamic system or environment, the integration of this information with previously acquired knowledge to form a coherent mental picture, and the use of that picture in directing further perception of, anticipation of, and attention to future events.

This definition is consistent with the three taxonomies of information processing identified by Endsley (1993, 1995). That is, SA is composed of the perception of elements in the environment, the comprehension of those elements, and their projection in the environment at some future state.

Despite their similarity to one another, Wickens characterization puts forth an important distinction between system awareness and environment or *hazard awareness*, a distinction which becomes crucial within the aviation domain. In particular, system awareness (what is happening with the aircraft?) can be composed of both attitude awareness (e.g., the aircraft is in a steep climb and about to stall) and systems awareness (e.g., what is my fuel level?). While questions regarding system awareness go beyond the focus of this paper, it is important to note that

problems associated with lack of attitude awareness have been well-addressed with the evolution of head-up and helmet-mounted displays (McNaughton, 1985), and deficiencies in systems awareness have also recently come under close scrutiny (Sarter and Woods, 1995).

In contrast to system awareness, hazard awareness is concerned with the pilot's understanding of both present and potential threats that exist outside the aircraft with respect to hazardous weather, conflicting traffic, airspace constraints, etc. Shortcomings in this type of SA are often linked to aircraft incidents and accidents that may involve penetration into hazardous weather, traffic conflicts, or controlled flight into terrain (Wiener, 1977). As a result, there have been several efforts to design cockpit displays that better support a pilot's hazard awareness. The EHSI (electronic horizontal situation display) and the CDTI (cockpit display of traffic information) are two examples of early attempts to support a pilot's SA (Ellis, McGreevy, and Hitchcock, 1984). Although these displays have been well received by pilots, in many cases their evaluation has given insufficient attention to how well they support SA of unanticipated hazards. For example, a pilot may perform well with a given display on tasks of local guidance, i.e., maintaining heading or altitude, regardless of their level of hazard awareness. However, if the pilot needs to make a sudden course change due to conflicting traffic the choice of action will be dependent upon his or her level of hazard awareness which in turn is dependent upon the display being used. As a result, measurements of SA in display design are essential for good pilot performance. The results from these SA measurements would allow designers to focus on those elements within the displays that may either hinder or benefit a pilots hazard awareness.

Measurement of Hazard Awareness

Various measurements exist to assess a pilot's awareness of external hazards with respect to display design (Endsley, 1995). The measures may be subjective in nature, in which case a flight crew member may be asked to evaluate his or her own level or their crewmember's level of hazard awareness. One important limitation of this measure is that the subject in question may not "know what they don't know" (Sarter and Woods, 1994). As a result, objective measures, which can be further classified into implicit and explicit measures, may be a preferable form of measurement. Implicit measures of hazard awareness rely on inferences which can be made based on the influence of prior events on task performance. For example, while evaluating a particular display format, the experimenter may give the pilot a vector into surrounding terrain. If the pilot accepts the vector the experimenter may infer that the pilot lacks adequate terrain awareness and will subsequently judge the display as supporting poor hazard awareness (Mykityshyn, Kuchar, and Hansman, 1994). Unfortunately, such a judgment reveals two potential difficulties in using this type of SA metric. First, implicit measurements are rather narrow in scope (i.e., a single vector in the above example) resulting in only a small assessment of the total situation. Second, inferences made by the experimenter may be inaccurate. Again referring to the above example, the pilot may accept the erroneous vector with the expectation that a second vector away from the hazard will be given at a latter time.

To provide a broader scope of situation awareness, one can employ explicit measures. These measures allow for a direct assessment of the information that is needed and can be employed to assess a broad range of issues. Explicit measures may be either retrospective or concurrent with respect to the experimental mission (Fracker, 1991). If the measure is retrospective, subjects complete the mission and then may be asked to identify various elements of the environment (reconstruct path, identify terrain location, etc.). Such a measure has the advantage of being completely non-intrusive, although long term memory limitations may restrict this measures effectiveness. That is, a poor path reconstruction after a flight may be mistakenly inferred as poor hazard awareness during the flight, when in fact the pilot had simply forgotten the path features since they were no longer relevant. Concurrent explicit measures allow for a more direct

measurement of SA since the probe is inserted at the time of the task rather than after the task has been completed. These measures may take the form of frozen screens or simple questions regarding the mission environment. Obviously, such an approach may be highly intrusive, although Endsley (1995) suggests the level of intrusiveness may not be as high as expected.

Regardless of the measure used to examine the effectiveness of a given display, there are certain criteria that the measure should meet before being employed within an experimental paradigm. That is, the measure should be both reliable and valid. Fracker (1991) and Endsley (1995) have dealt with issues of an SA measures reliability. Our interest here is in the validity of the measure. Validity itself can be measured in several different fashions. For example, *criterion validity* is achieved if a display (or person) measured by a SA test yields or possesses a low SA it will also yield low SA as reflected in some more operational criterion measure of performance (i.e., higher accident likelihood or incident rate). As an example, McMillan (1994) reports that Air Force pilots rated subjectively by their supervisors to have high SA were more likely to show superior (criterion) performance in an air-to-air combat scenario. Our interest in this paper is in a measures *construct validity*, which refers to the degree that a measure can quantify this unobservable psychological attribute. To accomplish this task we employed subject matter experts, namely pilots, to evaluate past measures we have used in our lab related to assessing a pilot's level of hazard awareness. A compilation of past measures was presented to pilot's who have had previous experience with these measures, and the pilot's were subsequently asked to evaluate how well the various measures truly assess a pilot's hazard awareness.

Method

Twenty University of Illinois aviation flight personnel participated in this evaluation. All participants were commercially qualified pilots with flight experience ranging from 200 to 6000 hours and all had taken part in at least one of the studies discussed below. Fourteen of these participants were also certified flight instructors. All participants received the same instructions regarding the purpose of the questionnaire and how to fill them out.

The questionnaire provided a brief explanation of hazard awareness and described the various SA metrics we have used in our lab to measure hazard awareness. Pilot's were asked to evaluate how accurately each of these measures assesses hazard awareness on a six point scale from 1) the measure is not at all an accurate assessment of hazard awareness to 6) the measure is an extremely accurate assessment of hazard awareness. The SA metrics were categorized as follows:

Screen-off out-of-path navigation (2 examples).
On random occasions the map display would blank and subjects would be thrown into an unpredictable bank and pitch angle, which would change their altitude and heading from its current location for 5 sec. Upon reappearance of the display, subjects were to turn and pitch in the appropriate direction as rapidly as possible, to orient to the still-active waypoint (Andre, Wickens, Moorman, and Boschelli, 1991).

At the end of the navigation task the map display would blank and subjects were asked to fly back to the starting point of the mission (Williams, Wickens, and Hutchinson, 1994).

Out-of-view direction pointing (2 examples).
Subjects were asked in which direction they would have to turn to reach the next checkpoint which was not visible in the forward view at the time the question was asked (Aretz, 1991).

At unexpected times during the landing task, an "X" mark, which was shown only on the map would flash for four seconds before all the terrain features on the map, including the mark, disappeared. Only the aircraft and the path were left on the map. The subject's task

was to indicate the direction needed to fly to the "X" mark (Faye and Wickens, 1995; Liang et al., 1995).

Map-FFV comparison (3 examples).
Subjects had to compare information given on the map display with information provided in a simulated forward field of view (FFOV):
Based on information provided in the FFOV, subjects needed to determine if the aircraft's location on the map display was accurately portrayed (Aretz, 1991).

A "lightning bolt" appeared at different locations in the FFOV. Subjects were then asked to determine the location of the lighting bolt on the map display (Harwood and Wickens, 1991). At unexpected time during the trial a single hot air balloon appeared on both the FFOV and on the map display. The subjects were asked to compare the vertical and lateral positions of both balloons and to respond verbally (Liang et al., 1995).

Screen-on out-of-path navigation (1 example).
Subjects were taken off their present path and placed in a nearby part of the world, from there they were to orient the helicopter to the original target object and fly back to it (Harwood and Wickens, 1991).

Intruder detection (1 example).
Subjects were asked to detect any intruder on the map display (Haskell and Wickens, 1993).

Position report (1 example).
Subjects were required to verbally indicate the position, in both ego-centered (o'clock) and world-centered (absolute degrees) terminology, and the height (high, same, low) of the nearest terrain hazard with respect to the aircraft (Wickens et. al, in press; Liang et. al, 1995).

Screen-off questions (2 examples).
At a given point on each flightpath, the simulation stopped, the screen was blacked out, and participants were prompted by a series of multiple choice questions:

On which leg were you flying? What is the heading you would need to fly to the box on the map? Does the obstacle X exist on the map you were flying? Is it above or below your current altitude? From your current position, you now are heading XXX and descending, will you run into an obstacle before contacting the ground? Enter the letter that represents the obstacle. Enter the shape of that obstacle (Wickens et. al., in press).

Where is the runway in relation to you and in which general direction is the aircraft traveling? What was the next turn on the flight path like? Generally, where is the runway from the aircraft? What is your aircraft position relative to the flightpath? In which direction will the next turn take you? Is the peak of the terrain feature located to your (right, left)? Is the peak of the terrain feature located to the (N, S, E, or W) (Wickens and Prevett, 1995)?

Map reconstruction (1 example).
After completing the flying mission, subjects were asked to either draw the map or reconstruct the last path flown (Wickens et al., in press; Aretz, 1991; Liang et al., 1995; Faye and Wickens, 1995).

Results

A t-test pairwise comparison was performed across all the SA metrics. Results from the Tukey's studentised range test, revealed a significant effect only between the screen-off questions and the

map reconstruction SA metric (Alpha=.1). As depicted in figure 1, pilots rated the screen-off questions (Task 7) as being the most accurate assessment of hazard awareness with an average score of 4.95 while the map reconstruction (Task 8) scored the lowest with an average of 2.65.

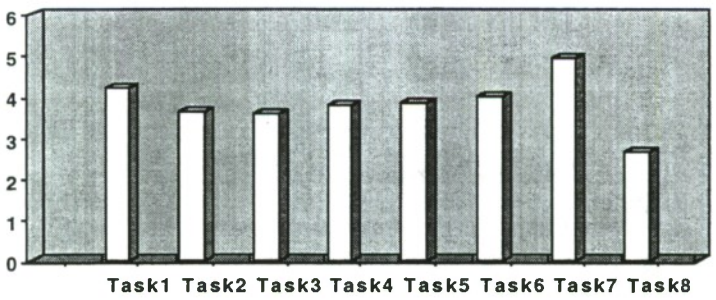


Figure 1

Discussion

The purpose of this paper was to evaluate the construct validity of various SA metrics that we have employed in our lab as related to display design. The results have shown support for the screen off task in which the mission was frozen and several queries were made regarding 1) the aircrafts present and future status and 2) the location of various objects surrounding the aircraft. Such a metric bears a strong resemblance to Endsleys Situation Awareness Global Assessment Technique (SAGAT) which employs similar queries to assess a subjects SA (Endsley, 1995). SAGAT has proven to be a useful measure in evaluating SA, particularly within the aviation domain. We hope the results from the current evaluation will offer further support for such measures. Its advantage in validity is reinforced by observations that it does not greatly disrupt performance of the ongoing task (Endsley, 1995).

The poor construct validity associated with the map reconstruction can be better understood in light of our earlier discussion regarding retrospective measures. That is, these measures may be viewed by pilots as poor measures of hazard awareness due to difficulties associated with failures in long term memory. Results from previous studies conducted within our lab have reported relatively poor results regarding a pilots level of hazard awareness when using this particular SA metric. This despite several attempts to make the reconstruction task easier for subjects (i.e., ask them to reconstruct the last path only not the whole world in which they flew; Liang, Wickens, and Olmos, 1995).

Finally, we should also note that the questionnaire employed was limited in some respects. First, only explicit measures were evaluated within the questionnaire hence it results have a limited generalizability to other SA metrics (e.g., implicit and subjective). Second, it should be noted that the relatively low number of respondents to the questionnaire (N=20) was due both to our need for experienced SMEs (i.e., commercial pilots) and for pilots who have had some experience with the measures we were attempting to evaluate.

Acknowledgments

Research funding was provided by the NASA Ames Res. Ctr., grant NASA NAG 2-308. Vernol Battiste was the technical monitor.

References

Adams, M. J., Tenney, Y. J., and Pew, R. W. (1995). Situation awareness and the cognitive management of complex systems. *Human Factors, 37*(1), 85-104.

Andre, A. D., Wickens, C. D., Moorman, L., and Boschelli, M. M. (1991). Display formatting techniques for improving situation awareness in the aircraft cockpit. *International Journal of Aviation Psychology, 1*(3), 205-218.

Aretz, A. J. (1991). The design of electronic map displays. *Human Factors, 33*(1), 85-101.

Endsley, M. R. (1993). A survey of situation awareness requirements in air-to-air combat fighters. *International Journal of Aviation Psychology, 3*(2), 157-168.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors, 37*(1), 65-84.

Ellis, S. R., McGreevy, M. W., and Hitchcock, R. J. (1984). Influence of a perspective cockpit traffic display format on pilot avoidance maneuvers. *AGARD Conference Proceedings No. 371: Human Factors Considerations in High Performance Aircraft.* (AGARD-CP-371), Neuilly Sur Seine, France, pp. 16-1/16-9.

Faye, E. L., and Wickens, C. D. (1995). *Strategies for display integration in navigational guidance and situation awareness.* (ARL-95-4/NASA-95-1). Savoy, IL: University of Illinois, Institute of Aviation, Aviation Research Lab.

Fracker, M. L. (1991). *Measures of situation awareness: An experimental evaluation.* Final Report AL-TR-1991-0127, Armstrong Research Laboratory, Wright-Patterson Air Force Base, OH.

Harwood, K., and Wickens, C. D. (1991). Frames of reference for helicopter electronic maps: The relevance of spatial cognition and componential analysis. *International Journal of Aviation Psychology, 1*, 5-23.

Haskell, I. D., and 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.

Liang, C.-C., Wickens, C. D., and Olmos, O. (1995). Perspective electronic map evaluation in visual flight. *Proceedings of the 8th International Symposium on Aviation Psychology.* Columbus, Ohio, Dept. of Aviation, Ohio State University.

McMillan, G. R. (1994). *Report of the Armstrong Laboratory Situation Awareness Integration (SAINT) Team.* Interim Technical Report AL/CF-TR-1994-0085. Situation awareness : Papers and annotated bibliography (U). Armstrong Laboratory, Air Force Materiel Command, Wright-Patterson AFB, OH, p. 37-47.

McNaughton, G. (1985). (Ed.) *Aircraft attitude awareness workshop proceedings.* Wright Patterson AFB Flight Dynamics Lab. Final Report.

Mykityshyn, M. G., Kuchar, J. K., and Hansman, R. J. Jr. (1994). Experimental study of electronically based instrument approach plates. *International Journal of Aviation Psychology, 4*(2), 141-161.

Sarter, N. B., and Woods, D. D. (1994). Pilot interaction with cockpit automation II: An experimental study of pilots model and awareness of the flight management system. *International Journal of Aviation Psychology, 4*(1), 1-28.

Sarter, N. B., and 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.

Vidulich, M., Dominguez, C., Vogel, E., and McMillan, G. (1994). *Situation awareness: Papers and annotated bibliography (U)*. Interim Report AL/CF-TR-1994-0085, Armstrong Laboratory, Air Force Materiel Command, Wright -Patterson Air Force Base, OH.

Wickens, C. D. (1995). *Situation awareness: Impact of automation and display technology*. Keynote address, NATO AGARD Aerospace Medical Panel Symposium on Situation Awareness, Brussels, Belgium.

Wickens, C. D., Liang, C.-C., Prevett, T., and Olmos, O. (in press). Egocentric and exocentric displays for terminal area navigation. *International Journal of Aviation Psychology*.

Wickens, C. D., and Prevett, T. (1995). Exploring the dimensions of egocentricity in aircraft navigation displays: Influences on local guidance and global situation awareness. *Journal of Experimental Psychology: Applied*, *1*(2), 110-135.

Wiener, E. L. (1977). Controlled flight into terrain accidents: System-induced error. *Human Factors*, *19*, 171-181.

Williams, H. P., Wickens, C. D., and Hutchinson, S. (1994). Realism and interactivity in navigational training: A comparison of three methods. *Proceedings of the 38th Annual Meeting of the Human Factors and Ergonomics Society Annual Meeting*. Human Factors and Ergonomics Society: Santa Monica, CA.

FliteScript: A Multimedia Test to Index Situational Schemata in Pilots

Alan F. Stokes and Donna Wilt

Florida Institute of Technology

Abstract

FliteScript is an interactive multimedia test instrument designed to quantify the "availability" of situational schemata to individual pilots. The goal of the test is to discriminate skill levels between pilots who may be similar by other measures (for example, flight qualifications or logged flight hours). Unlike many conventional test batteries, FliteScript is not intended to predict any aspect of stick-and-rudder proficiency, but rather to predict flight management and decision-making performance. While the situations used in the FliteScript approach are static, test scores appear to be fairly strongly associated with decision-making proficiency in dynamic contexts. A number of empirical studies are described in which the predictive value of the test is examined. The results support newer perceptual rather than older cognitive models of flight decision making and situation recognition.

Development of FliteScript

Introduction

FliteScript is, in fact, an umbrella term for two related tests, the FliteScript Recognition Test and the FliteScript Recall Test, both of which use radio communication dialog as the priming agent in invoking situational schemata. The tests are intended to index an important facet of expertise in pilots -- domain-specific nondeclarative knowledge -- and, more precisely, to quantify the readiness with which pilots can access situational schemata. This makes it possible to differentiate skill levels between pilots with, for example, similar flight qualifications and logged flight hours. Thus, FliteScript scores place test subjects on a continuum associated empirically with such aspects of performance as cue recognition and decision making (as discussed later). However, FliteScript scores do not purport to provide any absolute or criterion measure of acceptable performance.

Theoretical Basis of the Tests

As described later, the original FliteScript task was devised as a close analog of techniques used in the classic experiments reported by de Groot (1965) and by Chase and Simon (1973). These experiments investigated the nature of expertise using a very formalized domain, that of chess. Chess has long been used by psychologists and computer simulation developers as a standard environment for research on expertise and complex problem solving. (Indeed, so important has

chess become as a yardstick or reference case that it has even been called the "fruit fly" of such research; Hearst, 1978, p. 197, cited in Aanstoos, 1987.) A number of findings in the psychology-of-chess literature have considerable significance for conceptions of aviation expertise and situational awareness. Here we focus on Chase and Simon's demonstration that chess masters are far more capable than novices of accurately replacing chess pieces that have been removed from a board *if those pieces were previously arranged in a coherent game position.* If, however, the pieces are placed on the board in random fashion, chess novices and masters are equally poor at reconstructing the configurations. Moreover, as Calderwood, Klein, and Crandall (1988) have demonstrated, the quality of moves made by chess masters is not impaired by time pressure, whereas novices' moves are very significantly worsened.

The results of these experiments provide little support for the common view that chess masters (or experts in many other domains) must necessarily have inherently better basic cognitive skills: prodigious memories, faster reasoning, superior analytical skills, and the like. Rather, these results are more consistent with the view that expertise in chess is a function of the size and quality of the repertoire of game states internalized by chess players over their years of playing. In this view chess masters "pattern match" board layouts with templates stored in long-term memory (LTM), and are thus able to reconstruct board positions without overloading working memory with individual piece positions. Note that this key element of the decision-making process is in essence perceptual rather than analytical. Time pressure has very little effect on this efficient, perceptually based strategy, insofar as slow cognitive processes in working memory are largely bypassed. In other words, the experimental data favor a "top-down" knowledge representational model rather than a "bottom-up" information-processing model of experts' performance. One feature of the chess masters' performance is that the moment a board position is recognized and categorized, the player appears to entertain only a few potential high-value moves, which are analysed no more than perhaps two or three moves into the future. This contrasts with the supposition that chess masters use their (assumed) superior mental capacities to analyse a rather large number of pieces and to mentally play out future moves to significantly greater depths than novices.

The cognitive capabilities of experienced pilots may be analyzed in a similar fashion. On the one hand, it could be that these individuals, who are often highly selected personnel, have superior basic information-processing capabilities which presumably help them to exhaustively analyse situations and review all alternatives before selecting the best course of action (the information-processing or utility maximization model). Alternatively, by analogy with chess masters and novices, highly proficient flyers may differ little from low-time pilots in basic information-processing skills. Rather, they could be seen as accessing a large repertoire of well-organized situational schemata -- aerial game states, as it were -- which, once recognized and categorized, confine the consideration of action alternatives to a very few high-value "moves." (By high-value moves we mean, of course, options in which the probability of success is high). This is the knowledge representational model.

Stokes and Kite (1994) have argued that experienced pilots routinely function by using this form of perceptual-cognitive strategy in flight management. That is, this strategy is thought to be the preferred or "default" one. Slow computational and inferential processes in working memory (of the type tested by traditional cognitive test batteries) are resorted to only when when such a "pattern-matching" strategy cannot be utilized -- that is, when an appropriate schema is not evoked by the environmental cues (for example, because of unique circumstances or lack of appropriate training).

It is within this theoretical framework, then, that FliteScript was developed with the prime object of providing some means of indexing pilots' ability to access appropriate situational schemata. FliteScript is intended to serve as a test instrument that is sensitive to domain-specific knowledge representations in LTM without being aircraft- or equipment-specific. Moreover, the knowledge representations of interest are operational and procedural in nature, in contrast to the declarative "textbook" knowledge that may already be tested using standard FAA and similar tests. As a test of top-down processes, FliteScript is intended to complement test instruments such as the SPARTANS battery (Stokes, Banich, Elledge, and Ke, 1988) or the

COGSCREEN battery (Horst and Kay, 1991), which are primarily intended to index bottom-up cognitive processes in working or short-term memory (STM).

FliteScript Recall Task

As Barnett (1989) and Stokes, Belger, and Zhang (1990) have described, the Recall Task adopts the Chase and Simon experimental paradigm, but instead of chess pieces uses radio transmissions between pilots and Air Traffic Control (ATC). Participants are presented with a sequence of transmissions (e.g., giving or responding to taxiing instructions, calling for assistance, or discussing a weather problem). Afterwards, they must reconstruct from memory as much of the sequence as possible. However, by analogy with the two conditions of the chess experiments, half the sequences form coherent exchanges pertaining to a real situation, while the other half consist of various radio calls randomly juxtaposed. The ability to reconstruct a sequence is presumably influenced by the extent to which the sequence has evoked a situational schema or script in the first place, and, of course, by the quality of a subject's verbatim recall. The latter variable, however, is controlled out by the use of the random sequences.

FliteScript Recognition Task

The FliteScript Situation Recognition Task is intended to further differentiate between pilots who are able to construct an accurate mental representation of operations within the airspace and those who cannot. As in the Recall Task, participants listen to radio exchanges between ATC and pilots. Rather than reconstruct the radio sequence from memory, however, they identify the situation directly by specifying which of four diagrams best depicts the situation underlying the radio exchanges. The diagrams typically use approach plates, L-charts, or other terrain representations upon which aircraft symbols are superimposed. It is hypothesized that the speed and accuracy of the responses will be influenced by the extent to which a radio sequence invokes a situational schema. The two tasks, the Recall and the Recognition Task, both attempt to tap into knowledge representations in LTM; however, responses in the former are verbally mediated whereas in the latter spatial processes can be expected to predominate.

Empirical tests of FliteScript

A series of experiments on pilot decision making conducted at the University of Illinois produced results that were more consistent with the pattern recognition model of expertise discussed by Chase and Simon than with an information-processing model (Barnett, 1989; Wickens, Stokes, Barnett, and Davis, 1987). In particular, domain-independent information-processing measures were found to be rather poor predictors of pilot performance (an interesting finding in itself, given the ubiquity of such tests in pilot selection and screening procedures). Subsequent experiments using early versions of FliteScript showed that prediction of performance was significantly improved by tests based on knowledge representations in LTM rather than on basic information-processing (Barnett, 1989; Stokes et al., 1990; Stokes, Kemper, and Marsh, 1991).

Stokes et al. (1990) demonstrated that in the FliteScript Recall Task, high-time pilots showed a large improvement in scores from the random to the coherent condition. More important, as hypothesized, the size of this improvement was significantly greater than that observed in lower-time pilots. Stokes et al. (1991) repeated this experiment with similar results and also demonstrated that Recall Task results were moderately strong predictors of flight management decision-making quality.

In the case of the FliteScript Recognition Task, higher-time pilots scored significantly better than lower-time pilots, but Recognition Task scores were markedly superior to flight hours as a predictor of pilot flight management performance. It is important to note that Recognition Task scores were not merely reflecting spatial ability. The SPARTANS cognitive test battery contains a number of spatial tests, and none correlated to a significant degree with FliteScript Recognition scores. Moreover, results from a stepwise multiple regression showed the orthogonal nature of Recognition Task scores and scores from spatial tasks. Recognition task scores showed a surprisingly strong correlation with the quality of a pilot's flight management decision making as determined using a desktop flight simulator ($r=0.6517$, $p<0.001$). Additionally, plots of the relationship showed two distinct clusters, with higher-time pilots in one cluster and lower-time pilots in the other. Recognition score was also a significant predictor of the number of relevant problem cues detected by pilots in various problem scenarios. In a stepwise multiple regression containing both information-processing and knowledge representational variables, Recognition Task scores (on the LTM side) and spatial variables (on the STM side) were the best predictors of pilot flight management decision making (Stokes et al., 1991).

Finally, a study carried out by FAA contractors attempted to predict performance in a Boeing 737 flight simulator using scores from the COGSCREEN, WOMBAT, and the multimedia version of the general-aviation-adapted FliteScript tests (Hyland, 1993). The first two tests are conventional domain-independent batteries made up of various cognitive 'skills' tasks (tracking, spatial abilities, divided attention, etc.). Of the three tests used, only FliteScript attempted to index top-down domain-specific processes. The experimental results demonstrated that FliteScript scores separated Boeing 727 pilots out along a considerable continuum, and that they showed no significant correlation with WOMBAT or COGSCREEN scores -- an encouraging beginning. However, the performance criterion in this study was proficiency in *maneuvering* in a 727 flight simulator. This particular, very restricted, index of flight performance was not predicted by the multimedia FliteScript, as indeed it should not have been. Unfortunately, the study's designers did not include more appropriate performance criteria such as flight management decision making or situational awareness measures, and the opportunity to explore the knowledge-representational approach with airline pilots was lost.

Conclusion

Relatively simple static domain-specific tests designed around a chess analogy appear to be capable of predicting some important aspects of performance in dynamic flight contexts. Such tests seem capable of differentiating between otherwise similar pilots (as determined by conventional measures of qualification and experience). Perhaps more dynamic versions of such tests would be able to improve on this performance. Certainly, more development work and validation research is required, and one of the present authors (Wilt) is conducting just such a program of study at present.

In conducting such evaluations of domain-specific tests of knowledge representation, at least two points are worth reiterating. First, there is the evident need to include appropriate performance criteria in the experimental design. Measures primarily sensitive to stick-and-rudder prowess or the conduct of procedures are not adequate. Moreover, team performance factors (generating the 'team mental model') will also have to be accounted for where crew proficiency is, explicitly or implicity, the performance criterion. It is, for example, unclear (and a subject of current research efforts) whether FliteScript-like scores may be combined to predict crew performance. Second, scenarios in FliteScript-like tests need to be specialized to one subdomain (for example, general aviation, Part 121, or naval aviation), or made selectable such that specific subtests are available for different subdomains. By analogy, when testing chess expertise one would not expect particular success with checkers or Monopoly scenarios:

there can be no generic FliteScript. Finally, the apparent success of simple perception-based tests in predicting flight decision-making proficiency is consistent with the claims made in the literature on natural (i.e., nonexperimental) decision making (e.g., by Klein, Orasanu, Calderwood, and Zsambok, 1993). These claims seriously question the *general* applicability and usefulness of standard analytic/cognitive models of decision making in understanding many fields of expertise, including aviation.

References

Aanstoos, C. M. (1987). A critique of the computational model of thought: The contribution of Merleau-Ponty. *Journal of Phenomenological Psychology, 18,* pp. 187-200.

Barnett, B. (1989). *Modeling Information Processing Components and Structural Knowledge Representations in Pilot Judgment.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Calderwood, R., G. A. Klein, B. W. Crandall. (1988). Time pressure, skill and move quality in chess. *American Journal of Psychology, 101,* pp. 481-493.

Chase, W., H. Simon. (1973). Perception in chess. *Cognitive Psychology, 4,* pp. 55-81.

de Groot, A. (1965). *Thought and Choice in Chess.* The Hague: Mouton.

Horst, R. L., G. G. Kay. (1991). COGSCREEN: Personal computer-based tests of cognitive function for occupational medical certification. *Proceedings of the Sixth International Symposium on Aviation Psychology* (pp. 734-739). Columbus, OH: Ohio State University.

Hyland, D. T. (1993). Experimental evaluation of aging and pilot performance. *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 389-393). Columbus, OH: Ohio State University.

Klein, G. A., J. Orasanu, R. Calderwood, C. Zsambok. (1993). *Decision Making in Action: Models and Methods.* Nrwo(1991). Flight management training and research using a microcomputer flight decision simulator. In R. Sadlowe (Ed.). *PC-Based Instrument Flight Simulation: A First Collection of Papers* (pp. 25-32). New York: American Society of Mechanical Engineers.

Stokes, A. F., M. T. Banich, V. Elledge, Y. Ke. (1988). *Cognitive Function Evaluation in the Medical Certification of Airmen* (Technical Report ARL-88-4/FAA-88-2). Savoy, IL: University of Illinois Aviation Research Laboratory.

Stokes, A. F., A. Belger, K. Zhang. (1990). *Investigation of Factors Comprising a Model of Pilot Decision Making: Part II. Anxiety and Cognitive Strategies in Expert and Novice Aviators* (Technical Report ARL-90-8/SCEEE-90-2). Savoy, IL: University of Illinois Aviation Research Laboratory.

Stokes, A. F., K. Kite. (1994). *Flight Stress: Stress, Fatigue, and Performance in Aviation.* Aldershot, U.K: Ashgate.

Stokes, A. F., K. L. Kemper, R. Marsh. (1991). *Time-Stressed Flight Decision Making: A Study of Expert and Novice Aviators* (Technical Report ARL-91-10/INEL-9-1). Savoy, IL: University of Illinois Aviation Research Laboratory.

Wickens, C. D., A. Stokes, B. Barnett, T. Davis, Jr. (1987). *A Componential Analysis of Pilot Decision Making* (Technical Report ARL-87-4/SCEEE-87-1). Savoy, IL: University of Illinois Aviation Research Laboratory.

Modeling and Measuring Situation Awareness for Target-Identification Performance

Eileen B. Entin, Daniel Serfaty, and Elliot E. Entin

ALPHATECH, Inc.

Abstract

This paper discusses the modeling of situation awareness (SA) in the context of target-identification performance. We present an operational model of SA that is comprised of knowledge of the critical elements of the situation and the values of these elements as a function of time. Based on this model we discuss the research framework we are using to identify the critical elements of SA for this domain and to investigate the development of SA over time.

Introduction

Performance in complex military systems is often based on the integration of human observations and judgments with automated information. In target-identification tasks, the operator's judgment is based on a combination of visual, contextual, and automated information. The level of system performance that can be achieved is a function of the quality of the various sources of information and the operator's ability to fuse effectively information from those sources. Research in aided target identification has shown that performance is affected by intrinsic factors including the quality of the visual information, the time available to make a decision, the quality of automated information, and the nature of the human-machine interface (Entin, Entin, and Serfaty, 1995).

A critical advantage of human operators over automated target recognition (ATR) systems is their ability to make use of *situational information* in identifying targets. Humans are efficient and robust information processors in that they are able to use what they *expect* to see or hear to interpret incomplete and uncertain incoming information. On the other hand, humans can commit serious errors when they make decisions based on their expectations, as has been tragically illustrated by incidents in which soldiers fired at friendly units in locations where they had expected the enemy to be present.

Aviation records suggest that many pilot-caused errors result from a lack of situation awareness (SA). Nagel (1988) notes that breakdowns in SA are one of the most serious problems in aviation operations. Hartel, Smith, and Prince (1991) report that in a Navy and Marine analysis of mishaps, lack of SA was the most frequently cited causal factor. Thornton, Kaempf, Zeller, and McAnulty (1991) found that lack of relevant and timely information was related to Army tactical helicopter crews' poor performance in navigation and threat evasion. Both observational and experimental data support the intuitive assumption that higher levels of SA lead to better task performance, but the quantitative nature of this relationship has not been established.

This research addresses SA in the context of aided target-identification performance. The key premise is that a heightened level of SA will increase an operator's ability to integrate target-related information from various sources and thereby lead to higher levels of system performance. The

key objectives of this work are to demonstrate empirically the relationship between level of SA and aided target-identification performance, and to recommend principles for the display of SA information that optimize overall performance. To achieve these objectives, we must formulate an operational definition of SA in the target-identification domain and devise a method to measure quantitatively an operator's level of SA. In this paper we present an evolving research framework for modeling and measuring SA in the context of target-identification tasks.

Background and Motivation

What is Situation Awareness?

Although SA is an appealing and widely-used term, there is no universally accepted definition of the term, and no standard methodology for defining the elements of SA or assessing an individual's level of SA in the context of a particular task domain. Wellens (1993) suggests that in a military context, SA can be "roughly conceived of as an individual's internal model of the world at any point in time." The most widely referenced definition is one proposed by Endsley (1988) who defines SA as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future."

As a psychological construct SA has been discussed as both process and product. Like Endsley (1995a), we view SA as a state of knowledge captured at a particular moment in time, and situation assessment as the process of acquiring or maintaining SA. Clearly an individual's level of SA can change over time, and is dependent upon the amount and quality of information that is available–which is independent of the particular individual–and the individual's ability to perceive and comprehend that information in a timely fashion–which is a function of the individual's prior knowledge about what elements are important, how they are relevant, and how they change over time (Adams, Tenney, and Pew, 1995). Endsley's definition embraces both of these aspects of SA, with "perception" implying information availability and acquisition, and "comprehension and projection" suggesting the individual's ability to use the information.

The assessment of an individual's level of SA has been an equally elusive problem. (See Endsley, 1995b, and Adams et al., 1995 for a general discussion of the measurement of SA.) Because SA is a vaguely-defined concept and the elements of SA are not easy to identify or to quantify, it has sometimes been assessed in terms of task performance. However, as Endsley (1995a) points out, superior performance can result in spite of poor SA, and likewise high levels of SA do not always result in superior performance. In order for the concept to have meaning, SA must be defined and measured *independently* of performance. For example, in the target-identification domain, we cannot infer an operator's levels of SA by measuring aspects of his or her target-identification performance. Rather, we want to identify and measure those elements in the situation which are *predictive* of performance. A major goal of our work is to identify the critical elements of the situation for aided target-identification performance.

Although there is no agreed-upon definition for SA, researchers do agree that it must be defined in the context of a particular task. Moreover, for the concept to have meaning, one must be able to specify the elements that comprise SA, with particular sets of elements being relevant for particular system states. Furthermore, although the elements of the situation may change dynamically over time, only some of the changes will be large or severe enough to cause a change in the situation from the point of view of the system operator (Pew, 1994). For some elements, there may be certain ranges in which knowledge of the precise value of the element is not critical for SA, and other ranges in which the precise value is critical.

Effect of Situational Factors on Target-Identification Performance: Empirical Evidence

Target-identification performance is typically measured in terms of the hit rate (the proportion of targets correctly identified) and the false-alarm rate (the proportion of non-targets incorrectly identified as targets). The relationship between the hit and false-alarm rates is captured by receiver operator characteristic (ROC) curves based on signal detection theory (Green & Swets, 1974). In research investigating intrinsic system factors that could affect aided target-identification performance Entin and MacMillan (1993) found that as the quality of intrinsic information decreases, operators tend to modify their hit rate, but strive to maintain a relatively constant false-alarm rate. Results of a pilot study investigating how differences in operators' perceptions of the external situation affects target-identification performance provided a motivation for incorporating SA.

The pilot experiment established two levels of a situational factor, density of enemy units. The situational context information was presented to subjects off-line in the form of maps and a written intelligence briefing. The subjects' task in the experiment was to select the target objects in scenes comprised of targets and non-targets and to rate their confidence in their decisions. For each scene that they viewed, subjects made their target selections and confidence ratings twice: once without an ATR available and once with an ATR available. Each subject performed the target identification task under both high- and low-density conditions. In the experiment trials, the scene composition and ATR accuracy were the same in both situational contexts. The only difference was the off-line information that the subjects were told about the density of targets.

Figure 1 shows the ROC curves representing the subjects' performance under the high- and low-density situations. If the subjects' decisions were not influenced by situational context, one would expect that their performance would be the same in both situations, since the actual density of enemy units was the same in the high- and low-density situations. As the ROC curves show, however, in the low-density situation, subjects operated at a lower point on the curve (and were more confident in their decisions) than they did in the high-density situation. In the high-density condition the subjects shifted their operating point, raising their false-alarm rate significantly in both the visual ($p<.001$) and the visual+ATR ($p<.005$) conditions. *Yet the only difference between the two conditions was the prior contextual information they were given.* The introduction of prior information about the situation changed subjects' decision strategies on individual targets.

The experiment data provide empirical evidence that situational context strongly affects target-identification decisions. Whereas variations in intrinsic variables such as the quality of visual or automated information had no appreciable effect on the false-alarm rate, variations in situational context had a significant effect on the false-alarm rate. The sharp differentiation in false-alarm rates induced by the context information provides strong evidence that the situation in which individuals make their decisions can affect their willingness to tolerate false alarms. Given that in actuality there was no difference in the relative density of targets in the two situations, these laboratory results confirm outcomes that have occurred in real-world incidents, where soldiers fired on the basis of the expectation of an enemy presence. The results emphasize the importance of high levels of SA for high-level performance in complex systems, and motivate our efforts to quantify level of SA.

In this pilot experiment, we selected a particular element in the situation, the density of enemy units, that we thought was likely to affect target-identification performance. We then described the putative situation to subjects. Having provided information about the situation, we assumed that subjects had SA about the density of enemy units. Even in the absence of actual differences in the situation, we demonstrated that an individual's understanding of the situationcan affect target-identification performance. However, we did not investigate whether different levels of SA about the same external situation would affect performance. That is, we did not demonstrate that the subjects' level of awareness about the enemy, or the degree of accuracy or preciseness of their information about the density of enemy units, affected their target-identification performance. In order to investigate that relationship, we need an operational definition of SA in the target-

identification domain and a way to measure an individual's level of SA, and that is one focus of our current work.

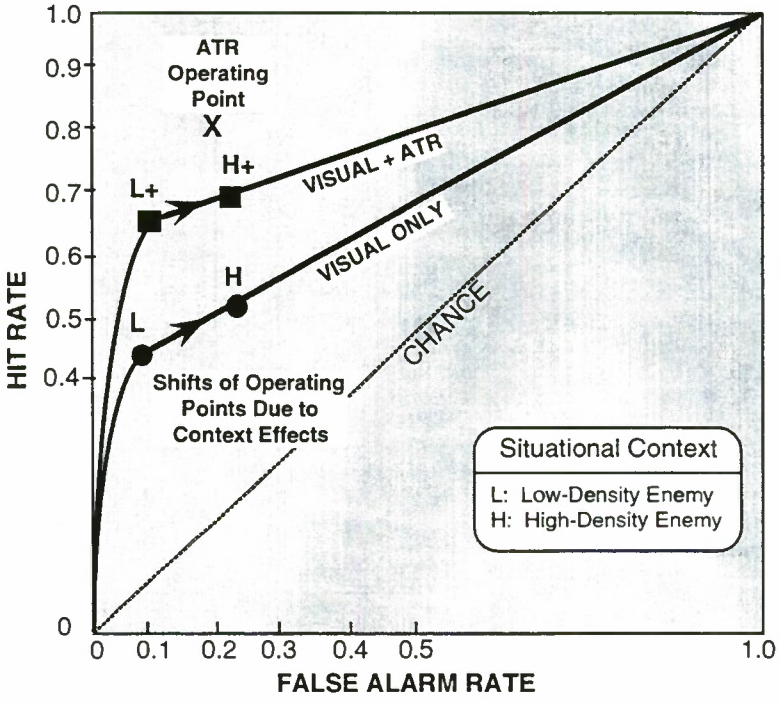


Figure 1. ROC Curves for High- and Low-Density Situations with and without ATR

Modeling Framework

Motivated by the general premise that a heightened level of SA will improve an operator's target-identification performance, and by the empirical evidence showing a relationship between elements of the situation and task performance, we propose a cognition-based, process-oriented research framework for establishing the elements of SA in the context of target-identification performance, assessing an individual's level of SA, and investigating quantitatively the relationship between SA and task performance.

Figure 2 shows the cognitive process model that provides the supporting framework for our work. The figure depicts a two-stage process by which target-identification performance is carried out. The first process embodies the evolution of SA. Situation assessment is a dynamic process, with input to the process coming from the individual decisionmaker's background knowledge and experience, and from global and local elements in the situation. The global elements, encompassing such factors as the geopolitical situation, establish the general situation and determine the local factors that will comprise the critical elements of SA. The global factors are the 'givens' that interact with prior knowledge and experience to provide an individual with an initial

mental model of the situation which, in turn, helps determine the information that is critical for SA in this particular situation. The output of situation assessment is a level of SA, which can be measured at any point in time.

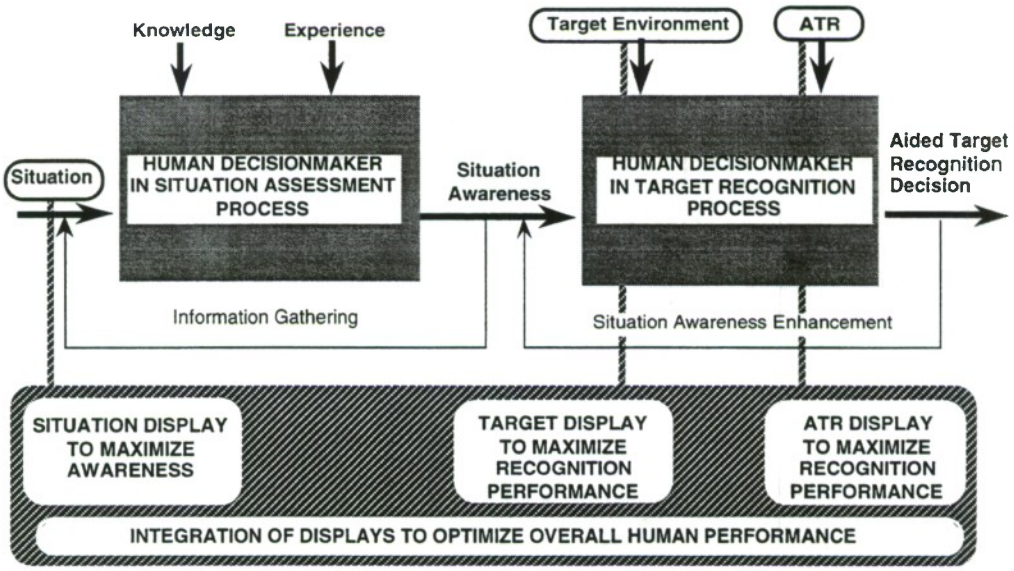


Figure 2. Process Model of Dynamic Situation Assessment and Decisionmaking Supporting Integrated Display Development

The second process in Fig. 2 concerns the target identification process. The decision making process takes as its input the individual's level of SA (measured as the output of the first process), visual information about the target environment, and information provided by an ATR. The output is a final identification of objects in the target environment based on the three input sources. Note that the output of this process becomes an input to the situation assessment process, and can affect the individual's subsequent level of SA.

As shown in Fig. 2, the cognitive process model provides the supporting structure for the development of displays to maximize performance. In previous work (Entin and MacMillan, 1993; Entin, Entin, and Serfaty, 1995) we have investigated unaided and aided target-recognition performance and we have developed displays to maximize target-recognition performance. Our current work is focused on incorporating SA into the framework. The following section presents a conceptual discussion of the model of SA that we propose. The model provides a foundation for the development of displays to maximize situation awareness and optimize overall human performance.

Modeling Situation Awareness

An individual's level of SA is dynamic, evolving over time as the situation changes. Since SA is domain dependent, what information is perceived and how it is interpreted is a function of the individual's task objectives. We propose that from an operational point of view SA is comprised of knowledge of four aspects of the situation. These four aspects incorporate the integrated processes of perception, comprehension, and projection over space and time proposed by Endsley (1988):

- What are the critical variables in the situation now? (comprehension)
- What are the current values (or states) of these variables? (perception)
- What will be the critical variables in the future? (projection)
- What will their values be? (projection)

These four aspects are interdependent. Knowledge of the critical variables and of the interrelationships among variables directs information gathering (perception), which, in turn influences both comprehension and projection. Projection can direct perception and comprehension in the future.

The attention to the critical elements of the situation (the first and third aspects) suggests that not all the situational elements are of equal importance. For example, air-to-air combat fighters view information about enemy aircraft as more important than information about friendly aircraft (Endsley, 1993). Furthermore, the same informational element may not be equally important over time. In target identification, knowledge of the location of an enemy aircraft may become more important over time as the pilot's own aircraft approaches the enemy aircraft. The task of an individual who is trying to maintain a high level of SA is not to represent the whole state at any time, but only the critical elements, the ones that will affect mission performance.

We represent the relationship between SA and performance by a *performance sensitivity model*:

$$[1] \quad D_p = f\,[e_1\ Ds_1 + e_2 Ds_2 \ldots + e_n Ds_n,\ t]$$

where:

D_p= decrement in optimal mission performance due to "less-than-perfect" SA

f = functional relationship

s_i = element of the situation

Ds_i= decrement in accuracy of estimate of value of s_i

e_i = sensitivity coefficient or criticality factor for s_i

t = time

This model represents the relationship between elements of the situation and mission performance in terms of the extent to which a decrement in the accuracy of an estimate of particular elements of the situation is related to a decrement in mission performance. The e_i values reflect the sensitivity (or criticality) of each element for performance. The magnitudes of these sensitivity coefficients are determined by the mission itself, not by the particular individual who is acting in the situation. They are dynamic in that their level of criticality may change over the course of the mission (for example, the weather may be especially critical during a particular phase of a mission). Knowledge of the critical elements is derived from an individual's past experience and his or her mental model of the situation. Studies of decision making expertise in complex task domains (MacMillan, Entin, and Serfaty, 1993) indicate that experts tend to agree on which elements are critical and which ones are less so.

The second and fourth aspects of SA concern the perception of the current values of situational elements and projection of their future values. Accuracy in estimating the values of the situational

elements (Ds_i) is dependent upon the operator's ability to perceive or infer the current values of the elements. Clearly the rate of change of the values of the situational elements is dependent upon the element (e.g., the position of an aircraft will change more rapidly than that of a tank) and the particular scenario as it unfolds (e.g., on some days the weather conditions will change more rapidly than on other days). The rate and extent of change in the values of the situational elements is also determined by actions taken by the individuals involved in that mission.

We represent the relationship between previous and current values of the elements by a *dynamic situation model*:

$$[2] \quad S(t) = g[S(t-1),\ d(t-1),\ e(t)]$$

where:

S(t) represents the values of the situational elements at the current time
g is a function representing the individual's dynamic mental model of the situation
S (t-1) represents the values of the situational elements at the previous point in time
d (t-1) represents actions taken at the previous time (e.g., a change in heading)
e (t) reflects the process uncertainty distribution (e.g., random flux in wind speed)

An individual can obtain the values for the elements of the situation through two mechanisms: direct observation and estimation. For those variables that are observable (for example, airspeed or cloud cover), the individual can perceive the information directly. But the values of some elements may not be directly available. For those elements that cannot be observed, the individual must reconstruct or estimate their value based on a mental model of the situation, current observations of other, related elements, and (if they are known) previous values of those elements. We can think of the function g as in part embodying an individual's mental model of how, in the absence of external forces, the elements are related to one another and how they change over time. For example, if the location of an enemy aircraft cannot be observed at a particular time, an individual can infer its location based on his knowledge of how fast that type of aircraft travels, atmospheric conditions such as wind speed, and the location of that aircraft at time t-1. This knowledge is embodied in the individual's dynamic mental model of the situation.

In reality it is not possible for an individual to have perfect SA at any time. In part, for information that is directly observable, the shortfall may be attributable to observation error. The individual may, for example, misread the airspeed or misunderstand the location of an enemy unit. The magnitude of the observation error will be related to the quality of the information that is available and the individual's observation skill. In part the shortfall may be due to faults in the individual's mental model of the situation. For example, in estimating the location of an enemy unit that is not directly observable, an individual may use a faulty mental model of the enemy's scheme of maneuvers. For situational elements whose values must be inferred from previous values, Equation 2 indicates that accuracy in estimating the current values will depend on the accuracy of the previous estimates of the values of those elements. In part the shortfall in SA may be due to information that is not available and cannot be reliably inferred from other information.

The third and fourth aspects of SA involve projection: what will be the critical variables in the future and what will their values be? Knowledge of these aspects must be based on the individual's dynamic model of how the situation will evolve in the future. Just as the individual's comprehension of the current situation directs his search for current situational information, his mental model of the current situation will direct his projection of the future situation, both in terms of what the critical elements will be and what their values will be.

Thus, an individual's level of SA depends upon his or her knowledge of the critical elements and the degree to which he or she is able to correctly perceive or infer the values of the critical elements of the situation over time. An individual who has accurate estimates of the sensitivity coefficients, who perceives the available information accurately, and who has an accurate dynamic model of the situation (that is, an individual who accurately reconstructs or infers unobservable information at the current time and uses that model to predict future values of critical elements) will

have a high level of SA. Equation 1 indicates that an individual whose estimates of the values of critical elements are perfectly accurate, but whose estimates of the values of noncritical elements are highly inaccurate will suffer little decrement in performance, whereas an individual whose estimates of all elements are only slightly inaccurate may actually suffer a greater decrement in performance.

To test the relationships described in Equations 1 and 2 we must measure SA as it develops over time. A typical attack helicopter mission is organized into five phases: planning, preparation, ingress, battle position, and egress. We can use these five phases as a way of operationalizing time. As shown in Fig. 3, to see how SA develops over time, we can measure SA at each phase of the mission. To gain additional information about how projection contributes to both the formation of SA and to performance, we can measure an individual's *estimated* level of SA for Phase 4, the phase in which the performance task occurs, at each of the three preceding phases. In Fig. 3 we represent SA at each phase by $SA_{(i)}$, the estimates of SA for Phase 4 at each of the preceding phases by $SA_{e4(i)}$.and performance at Phase 4 by P_4. If projection is indeed an important aspect of SA, then we would expect a positive relationship between the estimates of SA for Phase 4 at the previous phases of the mission and the actual level of SA at Phase 4.

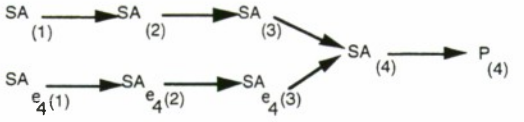


Figure 3. Measurement Plan for Assessing Current and Projected SA

Details on the operational implementation and testing of this measurement approach will be described in a forthcoming paper. In the next section we describe our method for identifying the elements of the situation and the values of the sensitivity coefficients.

Identifying the Critical Elements of the Situation

Using the conceptual approach we have described, our first goal is to identify the critical elements of SA for attack helicopters performing target-identification missions. A three-step process is used to identify the critical elements of SA for target-identification performance. The first step is the identification of a candidate set of elements that may affect target-identification performance at various times (or stages) of a mission. Working with a subject matter expert in the attack helicopter domain, we have developed a scenario that establishes the global elements of a mission (e.g., world situation, mission objectives). Using these global elements we have identified an initial candidate set of local elements of SA for this type of scenario. The candidate set of elements is presented in Table 1.

The second stage in this procedure is to obtain estimates of the criticality of these elements at the various stages the mission–in other words, estimates of the e_i values as a function of time. To accomplish this we will conduct structured, scenario-based interviews with experienced attack helicopter pilots. The pilots will be asked to rate the criticality of the elements enumerated in Table 1. They will also be asked to specify other elements that they believe are critical to target-identification performance. Assessments of the criticality of various elements will be made in each phase of the mission, so we can capture the criticality of the elements as a function of time. In

addition to describing and evaluating the elements, the pilots will be asked to explain how that information is used, and how frequently it is updated. Their responses will provide descriptive information about how experts use projection as an aspect of SA.

In the third stage we will integrate the criticality ratings obtained from the experienced pilots and use the integrated data to derive a set of criticality weights (sensitivity coefficients) as a function of time. Based on the interviews we will also have information about how each element of information is used, how the elements are interrelated, and how accurate an estimate of the value of the element (s_i) is needed, as a function of time. For example, in the ingress stage of a mission, it may be sufficient to know the approximate location of enemy units, whereas in the battle position stage, a more precise determination may be needed. This information is needed in Equation 1, in order to prescribe meaningful units of change.

Table 1. Elements of the Situation for Attack Helicopter Pilots

Task Organization	Service Support
Current Situation:	Commands/Signals
Friendly: number, type location of units	Aircraft Status
Enemy: number, type location of units	Crew Status
Weather: temperature, visibility, cloud types	Armament System: Status/Operations
Mission	Sighting Subsystems: Status/Operations
Execution:	Aircraft Survivability Equipment:
Concept of Operations:	Status/Warnings
Instructions/changes	Communications System: Status/Operations
Scheme of Maneuver	Fuel System: Status/Operations
Friendly Art/CAS	Navigation System: Status/Operations
Enemy ADA Support	Crew Briefing
Coordinating Instructions:	Changes to Planning Information
Action on Contact	
Critical Times	
Report Requirements	
Flight Coordination	
Special Mission Equipment	

Once the critical elements of SA have been identified, we will develop a scenario-based methodology for assessing an individual's knowledge of the criticality and value of each element at various stages in a mission and empirically relating those measures to aided target-identification performance. By comparing the individual's reports of the criticality of individual elements to the values obtained from experts and estimates of the values of the elements to their actual values, we can evaluate the extent to which knowing which elements are critical and knowing what the values of the elements are contribute to high levels of SA. Comparison of an individual's projections of the importance and value of elements in the future to the weighting and value he or she actually gives them at that future time will contribute to our understanding of how projection contributes to high levels of SA and to what extent the comprehension of the current situation and projection of the future SA contribute to performance.

Conclusions

Under performance levels currently achievable with target-recognition algorithms, the human operator plays an essential role in screening ATR detections and rejecting false alarms (Kuperman,

Bryant, and Clark, 1991). The observed tendency of operators to raise their false-alarm rate in the expectation of a high density of enemy units, combined with the unacceptably high false-alarm rates generated by currently available ATRs, makes it is particularly important to design displays that accurately portray situational information.

In order to develop displays to enhance SA, we must first identify the key situational variables that affect performance in this domain. To demonstrate that the displays are effective in improving SA, we must operationalize SA in a quantifiable manner. To design displays that enhance SA it is important to develop theory-based, unambiguous, objective, and quantifiable metrics that can detect performance improvement. We believe that the key to meaningful understanding of the relationship between SA and target-identification performance rests on the identification of the key elements of SA and the understanding of how they are used in the comprehension of the current situation and in the prediction of future events.

Acknowledgments

This work is sponsored by the Army Research Laboratory, Human Research and Engineering Directorate under the supervision of Dr. James Walrath and Ms. Jennifer Swoboda. We thank Mr. Joseph Zeller for his guidance and support in developing the scenario, enumerating the key elements of the situation, and collecting the criterion data.

References

Adams, M., Tenney, E., and Pew, R. (1995). Situation awareness and the cognitive management of complex systems. *Journal of the Human Factors and Ergonomics Society, 37* (1), 85-104.

Endsley, M. (1988). Situation awareness global assessment technique (SAGAT). In *Proceedings of the National Aerospace and Electronics Conference* (pp. 789-795). New York: IEEE.

Endsley, M. (1993) A survey of situation awareness requirements in air-to-air combat fighters. *International Journal of Aviation Psychology, 3*, 157-168.

Endsley, M. (1995a) Toward a theory of situation awareness in dynamic systems. *Journal of the Human Factors and Ergonomics Society, 37* (1), pp. 32-64.

Endsley, M. (1995b). Measurement of situation awareness in dynamic systems. *Journal of the Human Factors and Ergonomics Society, 37* (1), pp. 65-84.

Entin, E.B. and MacMillan, J. (1993). *Human image processing in unaided target detection* (TR-600). Burlington, MA: ALPHATECH, Inc.

Entin, E. B., Entin, E. E., and Serfaty, D. (1995). *Human-computer interfaces for machine-aided target acquisition* (TR-697). Burlington, MA: ALPHATECH, Inc.

Green, D. and Swets, J. A. (1974). *Signal detection theory and psychophysics*. Huntington, NY: Krieger Publishing Co.

Hartel, C., Smith, K., and Prince, C. (1991). Defining aircrew coordination: searching mishaps for meaning. Paper presented at the Sixth International Symposium on Aviation Psychology, Columbus: The Ohio State University.

Kuperman, G.G., Bryant, M. L., and Clark, L. G. (1991). *Man-machine interfaces for automatic target recognition systems*. Wright-Patterson AFB, OH: Human Engineering Division, Armstrong Laboratory and Mission Avionics Division, Wright Laboratory.

MacMillan, J., Entin, E. B., and Serfaty, D. (1993). Evaluating expertise in a complex domain–Measures based on theory. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* (pp 1152-1155). Santa Monica: Human Factors and Ergonomics Society.

Nagel, D. C. (1988). Human error in aviation operations. In E. L. Weiner and D. Nagel (Eds.), *Human factors in aviation*. San Diego, CA: Academic Press.

Pew, R. (1994). An introduction to the concept of situation awareness. In R. D. Gilson, D. J. Garland, and J. M. Koonce. *Situational Awareness in Complex Systems*. Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Thornton, R. C., Kaempf, G. L., Zeller, J. L. and McAnulty, D. M. (1991). *An evaluation of crew coordination and performance during a simulated UH-60 helicopter mission.* U. S. Army Research Institute Aviation Research and Development Activity, Fort Rucker, AL.

Wellens, A. R. (1993). Group situation awareness and distributed decision making: From military to civilian applications. In N. J. Castellan, Jr. (Ed.), *Individual and group decision making*. Hillsdale, NJ: Erlbaum.

Analyzing Situation Awareness During Wayfinding in a Driving Simulator

Jack Beusmans[1], Vlada Aginsky[1], Catherine Harris[2], Ronald Rensink[2]

[1] Cambridge Basic Research.
[2] Boston University.

Abstract

Learning a route through an unfamiliar area requires an ongoing awareness of one's position in the world. We investigated how subjects established this "situation awareness" in a driving simulator. After learning a route, subjects' visual and spatial abilities were tested by having them follow the route in a world with altered landmarks. We found that subjects used one of two different ways to orient themselves. One group of subjects relied almost exclusively on visual scene recognition, being aware of their position only at decision points along the route. The other group, in contrast, used a more spatial representation of their environment, being aware of their position between decision points as well.

Introduction

Although situation awareness (SA) has been studied most extensively in the context of aviation, it is relevant to other kinds of tasks as well. In particular, it is relevant to the more mundane task of driving, which requires SA in the literal sense of being aware of where and how one is situated within the world. Driving is both a source of great convenience and great danger in our lives (in 1993 in the US, over 7 million vehicles were involved in accidents, causing 2.6 million personal injuries and 36,000 fatalities). Consequently, much effort has been directed towards trying to understand the "human factors" component in vehicle accidents. Measures of basic visual performance have turned out to be only weakly predictive of accident rates (Hills, 1980). Instead, what is predictive are measures of cognitive abilities related to and subserving SA, such as being able to divide attention between multiple targets (Owsley et al., 1991; Ball and Rebok, 1994). As such, SA would appear to be a key factor in driving safety.

Given the importance of SA, it has been suggested that recent attempts to improve driving safety by adding "intelligence" to the car without duly considering the human driver may be counterproductive (Owens et al., 1993). As in the case of aviation, there is great concern that intelligent devices (collision warning systems, automatic cruise control, etc.) may decrease SA and so increase drivers' risk. The same issue has been raised for head-up displays, which superimpose visual information on the driver's forward view. If this information is similar to the actual scene—as in some experimental navigational aids—there is a real possibility that SA could be lost.

A better understanding of how drivers maintain SA is needed if intelligent devices are to be designed and used appropriately. The goal of our study was to learn about one particular aspect of SA: how drivers remain oriented within their environment, that is, how they establish a sense of being at a certain place in the world. As such, we view SA as describing the quality of the interaction between an actor and its environment for a particular task (Flach, 1995).

In our experiment, subjects learned to drive a simple route through a virtual world in a driving simulator. As soon as subjects had learned the route (and so reached a definite level of competence and presumably SA), we assessed their spatial knowledge and visual memory of scenes along the route. We used "ex-situ" (out-of-world) tests of spatial and visual abilities as well as more direct "in-situ" tests. The ex-situ tests, of course, "miss the phenomenon [i.e., SA]" (Sarter and Woods, 1991). However, they can reveal mechanisms underlying SA and also aid in interpreting the results of in-situ tests. For example, we used ex-situ tests to categorize subjects according to their visual and spatial knowledge, and then used this categorization to account for their behavior in a subsequent in-situ test.

Materials and Methods

Driving Simulator and Virtual World

The driving simulator consisted of the front two-thirds of a Nissan 240SX convertible. Steering wheel torque was generated by an AC motor attached to the steering column, generating a peak torque of 5.6 Nm and a sustained torque of 2.8 Nm, corresponding to the lower end of the range of torques that occur in normal driving. Audio feedback was in the form of low-frequency engine noise, with frequency proportional to driving speed. An Indigo^2 Extreme workstation (Silicon Graphics, Inc.) updated both car and world models, and rendered the virtual world, which was projected onto a wall 3.5m in front of the driver (image 60 deg wide, 40 deg high). Average frame rate was 12 frames/sec.

The virtual world consisted of a road system with about 50 intersections laid out on a green, textured ground plane of size 350 by 630 meters (Figure 1a). There were 24 rectangular buildings mainly along the route subjects had to learn. Half the buildings were "wide" (28m wide, 15m deep, and 12m high) and half were "tall" (10m wide, 10m deep, and 16m high). Of the twelve wide (or tall) buildings, half were blue and half were red. Each road section and each intersection had its own unique configuration of buildings (e.g., Figure 1b). There were no other cars or road users in the world.

Learning Phase

During the learning phase, subjects had to learn a 1,770m long route. Subjects could control their own speed and direction. They were led along the route by verbal directions from the experimenter. Instructions consisted of the phrases "take the next right" or "take the next left" and did not contain any landmark information. Subjects repeated the drive until they could follow the route correctly without any help from the experimenter. As learning progressed, the experimenter offered instructions only for the turns which the subjects had not yet memorized. Subjects indicated which turns they knew by using their direction signals before they turned.

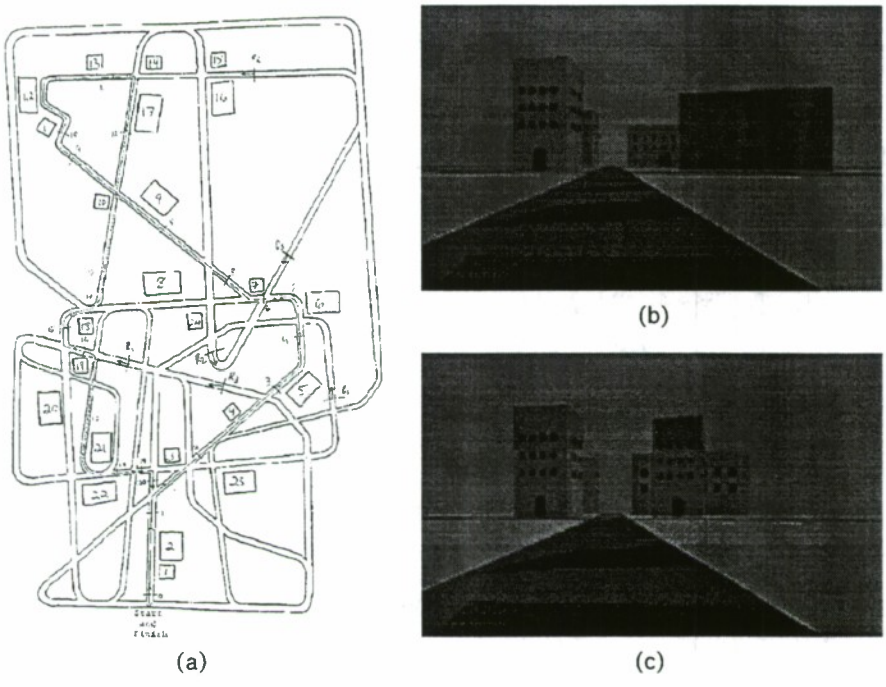


Figure 1. (a) Map of the world showing buildings and the route subjects had to learn. (b) Scene 11 as it appeared during learning. (c) Scene 11 as it appeared during the "in-situ" test.

Test Phase

Following the learning phase, subjects were given two ex-situ tests and one in-situ test (order of tests was: ex-situ test B, in-situ test, ex-situ test A).

Ex-situ Test A: Sketch Maps
Subjects were given a blank sheet of paper (11 by 17 inches) and asked to draw a sketch map of the route.

Ex-situ Test B: Visual Scene Recognition
Subjects viewed two sets of 24 static views or "snapshots" of the world (Figure 1b). Each set contained identical snapshots, composed of 21 views of scenes along the route, and 3 views of areas that subjects had never visited. In the *ordered* set, snapshots were in the order encountered along the route. In the *randomized* set, snapshots were placed in random order.

For each snapshot, subjects had to decide as quickly as possible whether they should turn right, left, or follow the road. Subjects were told to guess if they did not recognize a scene. Reaction times of all responses were recorded (resolution 14 ms). Subjects also rated the familiarity of the scene on a scale of 0.0 (completely unfamiliar) to 1.0 (very familiar).

The 21 route snapshots were divided into three classes of 7 snapshots each, according to the decision subjects had to make: (i) *no choice* road sections, where the visible road offered no

choice; (ii) *passive intersections*, where there was a choice, but the route followed went straight ahead; and (iii) *active intersections,* where subjects had to decide to turn left or right.

In-situ Test: Detecting Building Changes
This test followed ex-situ test B and was performed in the driving simulator. Subjects drove the route they had previously learned, but now 11 of the 24 buildings were changed in some way. Subjects were not told in advance what these changes could be. While driving along the route, subjects had to verbally indicate any differences they noticed. The experimenter recorded what the subjects said and how they were driving.

The 11 target buildings changed in either color (red or blue), shape (tall and thin or short and wide), or both color and shape. Figures 1b and c illustrate a change in building shape at an active intersection (which was noticed by 11 of the 16 subjects). Buildings could also change their location (cross to the other side of the street). Subjects, however, did not generally think of these "location changes" as the change in location of an identifiable building; rather, they interpreted it as the simultaneous disappearance of an old building and appearance of a new one (especially if the color and shape differed as well).

Subjects

Sixteen subjects participated as paid volunteers (10 men and 6 women; ages 19 through 25). All subjects were naive as to the purpose of the experiment.

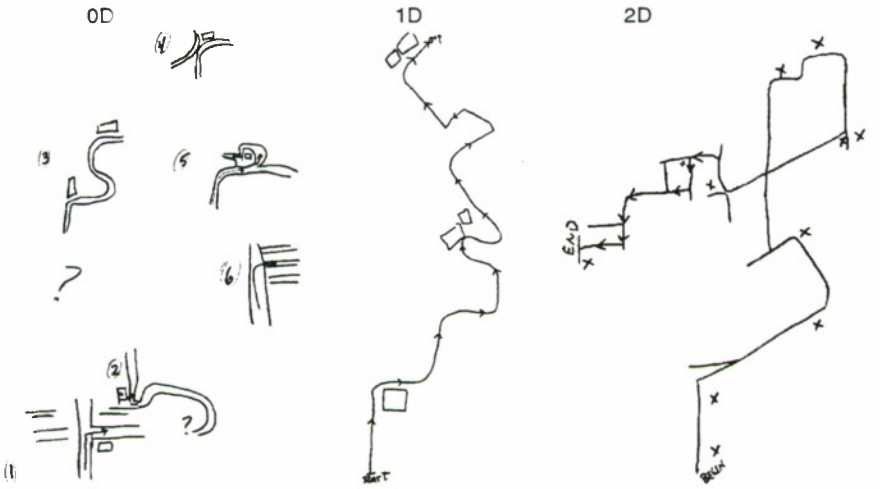


Figure 2. Representative samples of the three types of sketch maps.

Results and Discussion

Subjects learned the route with an average of 7.7 repetitions. The minimum number required was 6 (5 subjects) and the maximum was 10 (2 subjects). The time taken for one traversal of the route varied between 2-3 minutes. Only one subject realized that the start and finish of the route were at the same location.

Ex-situ Test A: Sketch Maps

Sketch maps obtained from subjects rarely reflected the correct (Euclidean) metric relationships among identifiable locations in the virtual world, even for distinct locations that directly followed each other. Turns and bends were typically drawn as right-angle turns even when they were not so in the virtual world. This may in part be due to the limited field of view in the driving simulator, which made it hard to judge sharp bends and turns; however, similar distortions have also been found in sketch maps of subjects who had learned a real-world space (Tversky, 1981). Areas with many turns or curves were enlarged at the expense of long straight road segments.

Sketch maps generally preserved the linear relationships between road segments and also depicted distinct locations in the world where subjects had developed a *sense or awareness of being in a particular place*. In most sketch maps, it was easy to recognize the 8 distinct places located along the route. Based on how these places were connected, three types of sketch maps could be distinguished (Figure 2):

0D Connection (Unconnected). Isolated places, with some local spatial structure. Places sometimes include information for their recognition. (3 subjects)

1D Connection. Places that had been encountered sequentially are explicitly connected in sequence, but there is little global structure. (8 subjects)

2D Connection. Places that had been encountered sequentially are connected sequentially; some of the places not encountered sequentially are connected spatially. (5 subjects)

Ex-situ Test B: Visual Scene Recognition

Averaged over all subjects, there were only small effects of presentation order and intersection type on scene recognition. In the ordered presentation, direction responses were 75% correct for passive intersections and 66% correct for active intersections. Performance was slightly—but not significantly—worse during the random presentation for active intersections (57% correct). Familiarity ratings for active and passive intersections did not differ, nor was there a significant difference between ordered and random presentations. Reaction times (RTs) for the direction responses were surprisingly long and varied considerably from subject to subject. The mean (±SEM) for the median RTs for ordered representations was 3.0 (±0.3) sec, and for random presentations 3.3 (±0.3) sec.

However, grouping the subjects according to their sketch map type uncovered an interesting pattern: for the 2D connection group, the mean percentage correct direction responses dropped from 81% in the ordered presentation to 56% in the random presentation; familiarity ratings dropped from 0.73 to 0.59; and RTs increased from 3.3 to 4.3 seconds. All subjects in the 2D group showed this drop in performance. In contrast, neither the 0D nor the 1D connection group showed this decrease in performance during the random presentation; in fact, direction responses improved slightly for the 1D connection group.

In-situ Test: Detecting Building Changes

First it was verified that subjects could still follow the route in the original world (all subjects could). Next, subjects followed the route through a world in which some buildings had been changed. Any navigation errors could be interpreted as lapses in SA due to these changes; indeed, subjects who got lost tended to notice fewer building changes (3.25 vs. 5.6; the difference was almost statistically significant). Building changes were noticed far more often at active intersections (77%), where subjects had to turn left or right, than at passive intersections or straight road sections, where they could simply follow the road (26%). This difference was independent of the type of building change.

Four subjects (two each from the 0D and 1D connection groups) missed a total of 7 turns. Apparently, the wayfinding actions of these subjects were "triggered" by visual scenes; if a scene was not recognized because of a building change, they would miss the turn. Interestingly, none of the subjects in the 2D connection group made any navigation errors.

Conclusions

All subjects in our study had reached approximately the same level of competence, that is, they could all follow the route. Thus, the differences in performance encountered on the various tests are unlikely to reflect different stages in spatial learning; rather, they would appear to reflect differences in handling the wayfinding problem itself. We found that subjects could be divided into three groups based on the structure of their sketch maps (0D, 1D, and 2D connection types). Only the 2D group showed a significant effect of presentation order in the scene recognition test; and it was the only group for which building changes did not cause navigation errors.

The consistent differences in performance in these groups point towards two strategies in wayfinding, one visually dominated and the other spatially dominated. These different strategies have implications for the kind and extent of situation awareness subjects develop.
The visual strategy relies on the visual recognition of active intersections along the route (e.g., "turn right at the red building"). If a particular intersection is not recognized (due to a change in one of the buildings, say) the turn will be missed. Subjects using this strategy apparently have little SA between active intersections.

The spatial strategy relies on a mental map incorporating aspects of the environment's spatial structure. Although subjects still recognize scenes and landmarks visually, they do not use this recognition to guide their navigation. Their ability to orient themselves via a mental map would explain why they performed better during the ordered presentation of the snapshots than during the randomized presentation. These subjects apparently have SA not only at active intersections, but everywhere along the route.

Our description of these two wayfinding strategies is of course rather crude and simplistic, but it does capture the extremes of the range of possibilities. It is also too simplistic to rigidly assign each subject to either one or the other strategy type—subjects may use different strategies at different parts of the route, and might switch strategies depending on the exact details of the task. Thus, the above interpretation of our results in terms of SA is rather tentative and should only be considered as a working hypothesis.

In any event, it is interesting that subjects with nearly identical levels of wayfinding performance have such different levels—and perhaps even types—of situation awareness. When evaluating navigational aids and head-up displays, it may be important to take these different ways of maintaining SA into account.

References

Ball, K., Rebok, G. (1994). Evaluating the driving ability of older adults. *J. of Applied Gerontol.,* 41, pp 20-38.

Flach, J. M. (1995). Situation awareness: proceed with caution. *Human Factors,* 37, pp 149-157.

Hills, B. L. (1980). Vision, visibility, and perception in driving. *Perception,* 9, pp 183-216.

Owens, D. A., Helmers, G., Sivak, M. (1993). Intelligent Vehicle Highway Systems: a call for user-centred design. *Ergonomics,* 36, pp 363-369.

Owsley, C., Ball, K., Sloane, M.E., Roenker, D. L., Bruni, J. R. (1991). Visual/cognitive correlates of vehicle accidents in older drivers. *Psychology and Aging,* 6, pp 403-415.

Sarter, N. B., Woods, D. D. (1991). Situation awareness: a critical but ill-defined phenomenon. *Int. J. Aviation Psychol.,* 1, pp 45-57.

Tversky, B. (1981). Distortions in memory for maps. *Cogn. Psychol.,* 13, pp 407-433.

Evaluation of "RLMS" Automotive Rear Lighting

David L. Cameron

Embry-Riddle Aeronautical University

Introduction

In the lives of most people, driving an automotive vehicle is the activity in which the most important safety concerns are associated with situation awareness. Among the many situation awareness demands of safe driving, a following driver must be aware of all vehicles ahead, must be aware of any intended lateral movement of any vehicle ahead, and must be aware of braking by any vehicle ahead. The rear lights of automotive vehicles are the primary means by which following drivers are made aware of these factors. In the late 1960s, weaknesses in then conventional all-red rear lighting were recognized as occasionally contributing significantly to the occurrence of a rear-end collision. This led to the addition of the high-mounted, center brake signal light in the mid 1980s. The high-mounted brake light has proven to be effective in preventing some rear-end collisions; its implementation, however, should not be assumed to have perfected rear lighting. Rear-end collision is proving to be a persistent safety problem. According to statistics released by the National Highway Traffic Safety Administration, nearly 1.5 million rear-end collisions occurred in the United States in 1991 (General Estimates System 1991). An unusually high failure rate in the high-mounted light has been noted in a recent survey of local traffic (Cameron, 1995b). It is thought that delayed or incorrect perception of rear lights is not a significant contributing cause in most rear-end collisions, so most such accidents would probably not be preventable by any change of rear lighting. *Some* rear-end collisions, however, probably could be prevented by improvement of rear lighting. Continuing improvement of rear lighting should be aggressively pursued. Better use of color might provide substantial potential for improvement.

An innovative, new color-specific approach to automotive rear lighting has recently been described (Cameron, 1992, 1995a). The new approach has been referred to as the "Red Light Means Stop (or the RLMS) approach". The objective of the RLMS approach is to make display of red light itself an effective, specific signal of braking. In the RLMS approach, red colored light, *and only red colored light*, is displayed at the rear of an automotive vehicle during braking. Nonred tail lights (also called the "presence lights" or the "running lights", illuminated by actuation of the headlight switch) and nonred rear turn signal lights are *extinguished* as soon as the vehicle is braked and the red brake signal lights come on. To preserve the signal of an intended turn during braking, the red brake signal light flashes on the side of an actuated turn signal switch. As soon as braking ends, the red brake signal lights are extinguished and the nonred rear lights (as appropriate) are reilluminated.

There are several straightforward methods by which RLMS rear lighting could be implemented. Existing rear lighting systems with amber rear turn signal lights can be converted to RLMS function by simply incorporating conventional relay switches into the wiring to the rear lamps. Rather than extinguishing tail light lamps and rear turn signal light lamps, another slightly more complex conversion method uses liquid crystal light shutters to change the color of light from amber to red during braking (Cameron, 1992).

Prior Consideration of Color in Rear Lighting

Thirty years ago a report that was sharply critical of red tail lights aroused interest in color as a potentially important parameter in automotive rear lights (Allen, 1964). A rear lighting method referred to as the "tri-light approach" was suggested and tested in the late 1960s and early 1970s. In the tri-light approach, two or even three different colors of light (to wit: amber tail lights and rear turn signal lights with red brake lights, or blue-green tail lights and amber rear turn signal lights with red brake lights) can be simultaneously displayed at the rear of a single vehicle. The simultaneous display of differently colored rear lights would unnecessarily complicate the perceptive task of a following driver (display of tail lights serves no useful purpose while brake lights are on) and could work against safety. The possibility of deleterious effects with multi-color (sometimes referred to as "Christmas tree") displays has been noted (Wickens, 1992, Pp. 523-524). Several reaction time (RT) studies included testing of rear lights as suggested in the tri-light approach (Mortimer, 1969 and 1970; Rockwell and Safford, 1966). Most of the testing was done at night and strongly emphasized ideal circumstances, with unimpaired test subjects having a clear view of the entire rear end of a single test vehicle. Subjects were given only a simple (or "Donders Type A"; Kantowitz, Roediger, and Elmes, 1991) reaction task, i.e., to press a given switch upon the onset of brake signal lights. Constantly illuminated red, amber, or blue-green tail lights simply provided an irrelevant background. When tested in this manner, the color of the constantly illuminated tail lights did not appear to be a significant factor in the RT performance of the test subjects. This led many safety professionals in the late 1970s to abandon the concept of color-specific rear lighting.

The testing of the tri-light approach, however, did not accomplish a thorough evaluation of color specificity in automotive rear lighting. The results of tri-light testing are not applicable to color-coding as suggested in the RLMS approach. Many rear-end collisions occur under adverse driving circumstances that were insufficiently considered in the prior research, or when a following driver is confronted by a choice reaction task rather than just a simple reaction task. As pointed out previously, numerous alternative approaches to rear lighting perform about equally well under ideal circumstances (Case, Hubert, Lyman, O'Brien, and Patterson, 1969). When a following driver has a clear view of the entire rear end of a single vehicle ahead, and the following driver is not himself impaired by fatigue, intoxication, or any other factor, then RT is determined more by inherent limitations in human physiology than by the type of rear lighting system in use on the vehicle ahead. Differences between alternative rear lighting systems are most clearly manifested, and most likely to be significant in accident prevention, when a following driver's view of vehicles ahead is impaired, when the traffic density is high and vehicle movements complex, or when the following driver himself is somehow mentally impaired. It could have been predicted that, when tested under ideal circumstances, any differences between conventional all-red rear lighting and tri-light rear lighting would be relatively minimal.

Human Factors Support for Color-Coding in "Displays"

A substantial amount of research by human factors scientists has been directed toward specifying the effect of color in various types of displays used to elicit appropriate human responses. The principles established by these studies apply directly to reaction tasks in driving as the through-the-windshield view of a driver essentially comprises a dynamic, often highly complex, three-dimensional visual display (Stokes, Wickens, and Kite, 1990). The rear lights on vehicles ahead are simply stimuli with variable, changing positions in the display. In a review of prior literature on color-coding for visual displays, Christ (1975) suggested that color-coding is more effective than coding by size, shape, or brightness. Color is perceived with relatively minimal higher

cognition, making it especially useful to mentally impaired persons. It has been noted that "differences in color are processed more or less automatically" (Wickens, 1992, p. 101). Persons with normal color vision, comprising approximately 92% of all persons, can recognize nine distinct colors on an absolute basis, with red being possibly the most easily recognized color (Stokes, *et. al.*, 1990, p. 79). It has been observed in some studies that color may not improve reaction performance with simple displays, where performance may be optimal even without color (Stokes, *et. al.*, 1990, p. 72). Simple displays, with a limited set of test lights being displayed on the rear of a single vehicle ahead, were utilized in a large majority of all prior testing of rear lighting. Many rear-end collisions occur, however, when a following driver is confronted by a complex through-the-windshield display. Color has been shown to very significantly improve reaction performance when displays are complex, especially when color is used as a redundant code for a given response (Kopala, 1979; Stokes, *et. al.*, 1990, Pp. 72-73). Consistent long term association of given color with a given situation or circumstance may form a population stereotype with respect to the color. Population stereotypes are important to the effectiveness of color in displays (Stokes, *et. al.*, 1990; Wickens, 1992). When the color of a stimulus is consistent with such a population stereotype, the response to the stimulus can be significantly enhanced by use of the color; however, when the color of a stimulus is contradictory toward such a stereotype, the correct response to the stimulus can be inhibited by use of the color. Very strong population stereotypes associate red with "stop", yellow with "caution", and green with "go" or "safe".

Red colored automotive tail lights give a false "stop" signal to following drivers who must cognitively override the signal to prevent an erroneous stop response. Under ideal circumstances, this does not usually cause difficulty in real-world driving, but under adverse circumstances it may. The most deleterious effect of red tail lights may be an occassional "boy-who-cried-wolf" effect: the constant display at night of red light for which a stop response is not needed may at times delay the stop response to red brake lights. The high-mounted light decreases the probability of such delay, but does not eliminate it. Distinguishing an illuminated high-mounted light from red tail lights in a complex through-the-windshield display requires higher cognition significantly beyond that needed for recognition of the color red.

Initial Testing of RLMS Color-Specific Rear Lighting

A preliminary RT study of RLMS rear lighting has been reported (Cameron, 1995a). In this study, test lights on the rear of a single stationary test car were displayed in three formats: as by conventional rear lighting systems, as by an RLMS rear lighting system, and as by a mixed population of vehicles, some with conventional rear lighting and some with RLMS rear lighting. Subjects were given a choice ("Donders Type B") reaction task. At night, even though testing was conducted under ideal circumstances in which differences would likely be minimal, both incidence of identification error and apparent identification cognition time were noticeably decreased when subjects were advised to regard all red lights as brake signal lights and lights were displayed in the RLMS format. In daylight, with strong sunlight falling onto the rear of the test car, RT values for the different formats were about the same. It was suggested that reflected sunlight, especially that reflected as colored light by the rear lenses, caused a high level of "noise" from which lamp illumination had to be distinguished, and that this was an important factor in the delay of identification cognition. The detrimental effects of reflection and glare have been previously noted (Stokes, *et. al.*, 1990, p. 76). The noise problem with respect to automotive rear lights displayed in daylight might be lessened by adapting rear lenses to reflect sunlight primarily as white light rather than as colored light. Color would then distinguish the signal from the reflected light.

Recommendations for Further Testing of RLMS Rear Lighting

Further testing of color-coded automotive rear lighting, especially as suggested in the RLMS aproach, should be performed in a timely manner. Strong general support for the likely safety value of RLMS rear lights, under at least some circumstances applicable to real-world driving, is already provided by prior human factors research. The results of that research provide considerable guidance for the design of specific tests for RLMS rear lights. It is recommended that the following factors be considered in the design of such tests:

1. RT testing under simple ideal circumstances must be included, possibly in an on-the-street format such as has been previously used (Rockwell and Safford, 1966). It is likely, however, that under such circumstances differences between RLMS test lights and conventional test lights will be minimal. To equalize the performance potential of test subjects prior to actual testing, subjects tested with RLMS lights should be given some training with those lights to compensate for long experience with conventional rear lights.

2. To reveal differences that may appear to be insignificant under ideal circumstances, testing under adverse circumstances should be done. Since this would necessarily involve hazardous driving situations, testing must be done in a simulation format rather than on-the-street. Simulated situations should include complex through-the-windshield displays with several leading vehicles, times when a driver would have to make difficult maneuvers through traffic, and subjects mentally impaired by fatigue or intoxication.

Considering the strong support for color-coding provided by previous human factors research, it is likely that specific testing of RLMS rear lighting will confirm its safety superiority under at least some adverse circumstances applicable to driving in the real-world. If that confirmation is obtained, then implementation of RLMS rear lighting in the real-world is probably indicated. It has been noted that principles confirmed by human factors research are often not fully utilized in the design of devices and systems for use in the real-world (Meister, 1971). RLMS rear lighting may represent a special opportunity to actually apply scientific research to the lives of "real people". Since automotive safety is a human issue rather than an esoteric scientific one, it is not needed that a change in rear lighting be likely to prevent large numbers of accidents. Any level of safety improvement, even if applicable to only a limited set of driving circumstances, is sufficient to warrant timely change.

References

Allen, M. J. (1964) Misuse of red light on automobiles. *American Journal of Optometry Archives of the American Academy of Optometry*, 41, pp 695-699.

Cameron, D. L. (1992) An innovative solution to continuing misuse of red light on automobiles. *Optometry and Vision Science*, 69, pp 702-704.

Cameron, D. L. (1995a) Color-specificity to enhance identification of rear lights. *Perceptual and Motor Skills*, 80, pp 755-769.

Cameron, D. L. (1995b) Deterioration of the high-mounted brake light. *Perceptual and Motor Skills*, 81, p 418.

Case, H. W., Hubert, S. F., Lyman, J. H., O'Brien, P., and Patterson, O. E. (1969) *Selection of Vehicle Rear Lighting Systems*. Univer. of California at Los Angeles,

School of Engineering and Applied Science, Institute of Transportation and Traffic Engineering, Report No. 70-9.
General Estimates System 1991. Washington, D. C.: National Highway Traffic Safety Administration.
Kantowitz, B. H., Roediger, H. L. III, and Elmes, D. G. (1991) *Experimental Psychology.* (4th ed.) St. Paul, MN: West Publ.
Kopala. C. J. (1979) The use of color-coded symbols in a highly dense situation display. *Proceedings of the Human Factors Society, 23rd Annual Meeting*, pp 397-401.
Meister, D. (1971) *Human Factors: Theory and Practice.* New York, NY: Wiley- Interscience.
Mortimer, R. G. (1969) *Research in Automotive Rear Lighting and Signaling Systems.* General Motors Technical Center, General Motors Engineering Publ. 3303.
Mortimer, R. G. (1970) *Automotive Rear Lighting and Signaling Research.* Univer. of Michigan, Highway Safety Research Institute, Report No. HuF-5.
Rockwell, T. H. and Safford, R. R. *Comparative Evaluation of an Amber-Red Taillight System and the Conventional System Under Night Driving Conditions.* Ohio State Univer., Systems Research Group, Report EES 272-1.
Stokes, A., Wickens, C., and Kite, K. (1990) *Display Technology: Human Factors Concepts.* Warrendale, PA: Society of Automotive Engineers, Inc.
Wickens, C. D. (1992) *Engineering Psychology and Human Performance.* (2nd ed.) New York: Harper-Collins.

Comparing Explicit and Implicit Measures of Situation Awareness

Leo Gugerty and William Tirre

Brooks Air Force Base

Adams, Tenney, and Pew (1995) describe situation awareness (SA) as a mental model of a dynamic situation that has two elements: (1) explicit focus - active knowledge in working memory, and (2) implicit focus - less active knowledge that is relevant to the current situation, but more accessible than irrelevant long-term-memory knowledge. The *first goal* of the research described here was to develop measures of SA that can assess the knowledge used in dynamic (real-time) tasks, including both the knowledge in explicit and implicit focus. We used driving as a task domain and a PC-based driving simulator as an experimental "vehicle". In particular, we focused on knowledge needed for the driving subtask of maneuvering, that is, how drivers track the locations of the of the vehicles around them.

Other important aspects of real-time cognitive tasks are that the operators must allocate attention among multiple subtasks, and operators' cognitive capacities are often overloaded. Therefore, the *second goal* of this research was to investigate how drivers' knowledge of traffic car locations is affected by changes in working-memory load. In addition, we also investigated how drivers allocate attention among changing objects when their working-memory capacity is overloaded.

Explicit and Implicit Measures of Situation Awareness

A number of researchers have used measures of explicit real-time knowledge such as blanking the simulator screen and asking the subject to recall information about the scenario (Fracker and Davis, 1991; Endsley, 1995b). However, recall-based measures alone may provide an incomplete picture of SA, because many real-time tasks involve well-practiced, automatic processes that may register information in explicit, working memory only fleetingly, if at all. As an extreme example of this phenomenon, one can sometimes drive a car while being engrossed in a conversation and immediately thereafter have no recollection of the past few minutes of driving. Given this criticism, we sought to develop situation-awareness measures that did not depend on explicit recall. Such measures required making inferences from subjects' performance. For example, how often a driver avoids hitting cars in the blindspot while maneuvering could be a useful implicit (performance-based) measure of SA.

Many researchers claim that research supports a distinction between explicit and implicit memory systems or processes (e.g., Roediger & McDermott, 1993). However, recent research suggests that some dissociations between explicit and implicit tests may be due to experimental artifacts or insensitive explicit or implicit measures, and that the growth of explicit knowledge during learning may closely parallel knowledge tapped by implicit tests (Perruchet & Amorin, 1992; Buchner, Funke & Berry, 1995). Applied to the example of driving without recollection, these findings suggest that during the unrecollected period of driving, people have at least brief periods of explicit awareness of key driving knowledge, and that this knowledge may be revealed

by appropriate tests (cf., Endsley, 1995a). Given these arguments, we expected that our subjects' recall and performance-based knowledge would be positively correlated.

In the following, we describe the driving task and the situation-awareness measures used. Then we present two experiments that were conducted to evaluate the SA measures.

Driving Task

The driving task was performed on a PC-based driving simulator. The simulator showed 3-dimensional animated driving scenes in a 6.1" by 4.5" window on the computer screen. The subject saw the front view from the driver's perspective and also the rearview, left-sideview, and right-sideview mirrors. All scenes showed traffic on a three-lane divided highway, with all cars moving in the same direction.

Subjects watched animated scenes lasting from 18 to 35 seconds, and were instructed to imagine that their simulated car was on autopilot. At the end of each scene, subjects' knowledge of the locations of the traffic vehicles was probed using one or both of two methods. In the *recall probes*, the moving scene disappeared and subjects indicated the locations of the traffic cars at the end of the scene on a bird's-eye view of the road, using the mouse. The bird's-eye view showed the road 17 car lengths ahead of the driver and 9 lengths behind, and the driver's car (in the correct lane). After subjects finished recalling the car locations for a scene, they received feedback indicating the correct final car locations for that scene.

In the *performance probes*, subjects could make driving responses while viewing the moving scenes; that is, they could override the autopilot. On some trials, an incident would occur that required a driving response, for example, a car would move into the driver's lane ahead of the driver while moving slowly enough that it would hit the driver. Subjects could make four responses to avoid hazards such as this: accelerate, decelerate, move to the lane on the left, or move to the lane on the right. They indicated these responses with the up, down, left, and right arrow keys, respectively. Subjects could usually make only a single arrow-key press on each trial. When they did this, the moving scene usually ended. After the scene ended, the subjects received textual feedback concerning the correctness of their response.

The hazards were designed so that it was very clear to the subject that a driving response was required. During the part of hazard trials before the hazard occurred, and during all of the non-hazard (catch) trials, it was clear that a driving response was not required. On each hazard trial, a response interval was defined as starting when the hazard car moved into the driver's lane, close to the driver, and ending just before the hazard car would hit the driver (at the end of the trial). A response interval was also defined at the end of each catch trial; this interval was equal in length to the average of the hazard-trial response intervals. For both hazard and catch trials, if subjects responded during the response interval, the scene stopped and they received feedback. If subjects responded before the response interval, the computer beeped, the moving scene continued, and subjects had to be prepared to respond later in the scene, if necessary.

Situation Awareness and Global Performance

Explicit (Recall) Measures of Situation Awareness

To evaluate the "goodness" of subjects' recall data, first, a computer algorithm matched the cars recalled by subjects with the actual locations of cars at the end of each trial, and also identified nonrecalled cars and false alarms. Once the matching was done for a trial, the *percentage of cars recalled* and the *average location error for recalled cars* (based on Euclidean error distances) were calculated. We also combined both of these variables into a composite measure (*weighted-average*

location error for all cars), which assigned a high error distance for nonrecalled cars and weighted errors in recalling distant cars as less important than errors to nearby cars. In this and the accompanying article, the term "average location error" will refer to this composite measure.

Implicit (Performance) Measures of Situation Awareness

We agree with Endsley (1995a) that the concept of SA is best seen as encompassing perceptual and comprehension processes, but not decision-making and response execution processes. The major difficulty in developing performance measures of SA is creating measures that reflect peoples' perceptual and comprehension processes more than they reflect decision and response-execution process, even though all of these processes must contribute to any measure of real-time performance. Using Endsley's (1995b) terms, we wanted to develop *imbedded task measures*, which reflect particular aspects of SA. In contrast to imbedded task measures, *global performance measures* reflect a more even mix of situation-awareness and decision/action processes. Global measures, which are described later, will also be useful as criterion variables to assess how both explicit and implicit SA affect overall driving performance.

The first imbedded-task measure we developed was *hazard detection*, or sensitivity at detecting hazards. We used the A' nonparametric, signal-detection measure of sensitivity (Grier, 1971). On each signal (hazard) trial, the response interval began when a car entered the driver's lane on a trajectory that would hit the driver and ended when it was too late for the driver to avoid the oncoming car. Following the procedure of Watson and Nichols (1976) for measuring sensitivity and bias with continuous signal-detection tasks, we defined catch-trial response intervals that were equal in duration to those on signal trials. A hit was defined as any arrow-key response, even an incorrect response, during the response interval of a signal trial. A false alarm was any arrow-key response during the response interval of a catch trial. For all trials, responses before the response interval, which were infrequent, were ignored in this analysis.

When subjects responded incorrectly to a signal (hazard) car, this shows that they were aware of the hazardous situation, but selected and executed an inappropriate avoidance response. Therefore, by defining even incorrect responses to signals as hits in this measure, we hoped that it would reflect subjects' ability to detect hazards (an aspect of SA) more than their decision/action abilities. The hazard detection measure focuses on subjects' awareness of vehicles in front of and behind their car, since the hazardous cars always entered the driver's lane from a side lane and then approached the driver.

Our second imbedded-task measure focused on subjects' awareness of cars in the blindspots to the right and left. Since our 3D display did not show cars immediately to the right or left of the driver, subjects had a larger blindspot than in real driving. On signal trials, the only way subjects could know about cars in the blindspot was by remembering that a car had entered the blindspot and had not left it. All cars in the blindspot during response intervals were located such that the subjects' car would hit them if the subject tried to avoid the signal car by moving into the blindspot. On a trial where the signal car approached from the front and there were cars in the right and left blindspot, subjects were considered as avoiding 2 of 2 blindspot cars if they braked or accelerated, and 1 of 2 if they went right or left. Overall *blindspot avoidance* was estimated by the ratio of the total number of blindspot cars avoided over the total number of blindspot cars.

As in the hazard-detection measure, scoring high on the blindspot-avoidance measure does not depend on making a correct response, in terms of global task performance. In the above example, a subject could accelerate into the oncoming car and still be credited with avoiding 2 of 2 blindspot cars. Thus we hoped that the blindspot avoidance measure would reflect subjects' awareness of blindspot cars more than their decision/action processes.

Global Measure of Driving Performance

Our main measure of global performance was *crash avoidance*, the percentage of hazards successfully avoided. A correct response on this measure involved avoiding any signal cars without hitting blindspot cars on signal trials, and making no response during the response interval of catch trials.

Experiments 1 and 2

A *primary goal* of the experiments was to compare what the recall and performance measures told us about drivers' SA, in particular to test whether these measures were associated or dissociated. To do this we combined the performance probes and recall probes into a third type of probe, *performance-recall probes*, in which subjects gave driving responses, as necessary, during the moving scenes, and then recalled the traffic car locations at the end of the scene. In the performance-recall probes, the feedback about the correctness of the driving responses was withheld until after the subject had completed the recall probe.

The *second goal* of the experiments was to see how subjects' SA for car locations is affected by changes in working memory load. Thus, across the driving scenarios shown to subjects, we varied the number of cars to be tracked and recalled from 4 to 7.

Method

Experiment 1
The 35 subjects were hired from temporary employment agencies. A within subjects design was used, with 18 subjects (8 males and 10 females) receiving the recall probes on the first day and the performance-recall probes on the second day, and 16 subjects (7 males and 9 females) receiving the reverse order. After the initial instructions, each subject completed one block of 28 trials in the morning, and two blocks of 28 in the afternoon. Each block took between 45 and 60 minutes to complete. On the following day, subjects completed three blocks of 28 trials, following a schedule similar to day 1.

Experiment 2
Forty six paid subjects (25 males and 21 females) from temporary agencies participated. After the initial instructions, each subject completed a block of 42 trials with recall probes. Three days later, subjects completed a block of 42 trials with performance-recall probes.

Results

Comparing Explicit (Recall) and Implicit (Performance) Measures of Situation Awareness
Before comparing the recall and performance measures, we first evaluated whether subjects performed at greater than chance levels on these measures, since some researchers have found that subjects performed at chance levels on situation-awareness measures in real-time tasks (Fracker & Davis, 1991). For both experiments, 85% of the subjects performed better than chance on the recall measures and all subjects did better than chance on the imbedded-task and global performance measures. The recall measures were highly reliable, with split-half reliabilities generally above .90 for both experiments. For performance measures, the reliabilities for Experiments 1 and 2, respectively, were: sensitivity at detecting hazards: .80 and .68; blindspot avoidance: .75 and .68, and crash avoidance: .68 and .54. The reliabilities for the performance

measures dropped in Experiment 2, because it used only half as many trials as Experiment 1.[1] In Experiment 2, blindspot avoidance and crash avoidance were based on only 21 trials per subject, and hazard detection on 42 trials.

Relationship Between Recall and Performance Measures of Situation Awareness

Table 1 shows the correlations (Pearson's r) between the recall measures and both the imbedded-task and global performance measures. The data for Experiment 1 and 2 show a different picture. For Experiment 1, recall and imbedded-task measures of SA correlate positively with each other. That is, better hazard detection and blindspot avoidance is associated with better explicit recall of car locations. (The negative signs on some correlations always occur when a percent-correct score is correlated with an error score.) Thus, we did not find a dissociation between explicit (recall based) and implicit (performance based) measures of SA. These data suggest that subjects were not using significant implicit knowledge in Experiment 1. Rather, both the explicit and implicit measures seem to be tapping into subjects' explicit knowledge of car locations that is used to perform both the recall and the driving tasks.

In Experiment 2, the correlations between variables were generally lower than in Experiment 1. In particular, there were no significant correlations between the recall and imbedded-task performance measures of SA. This may have been due to the lower number of trials in Experiment 2 (half of the first experiment), which led to lower reliability of the performance-based situation-awareness measures. Twenty to forty trials may not be enough to estimate performance-based SA.

Table 1. Correlations between recall and performance measures of situation awareness

	Recall Measures of SA			Imbedded-Task (Performance) Measures of SA		Global Performance Measure
	Weighted avg. loc. error	% cars recalled	Avg. location error, recalled cars	Hazard detection	Blindspot avoidance	Crash avoidance
Weighted avg. location error		-.77 (.001) -.84 (.001)	.74 (.001) .51 (.001)	-.55 (.001) .04 (.79)	-.47 (.005) .07 (.62)	-.69 (.001) -.35 (.02)
% cars recalled			-.17 (.33) .00 (.99)	.44 (.01) -.04 (.77)	.32 (.07) .05 (.74)	.46 (.01) .27 (.07)
Avg. location error, recalled cars				-.39 (.03) .10 (.51)	-.38 (.03) .27 (.07)	-.58 (.001) -.22 (.14)
Hazard detection					.56 (.001) .16 (.28)	.63 (.001) .36 (.02)
Blindspot avoidance						.30 (.09) .07 (.64)

Note: Experiment 1 and 2 in rows 1 and 2, respectively; alpha levels in parentheses.

[1] This dropoff in reliability did not occur for the recall measures, because these were based on 4 to 7 observations (cars) for each trial, while the performance measures were based on only 1 observation per trial. Also, there were twice as many recall trials as performance trials.

Effects of Working Memory Load on Recall and Driving Performance

The second goal of the experiments was to see how subjects' SA for car locations was affected by changes in working memory load, that is, by changes in the number of traffic cars to be recalled. The accuracy of subjects' explicit recall decreased as traffic level increased. As the number of traffic cars increased from 4 to 7, the percentage of cars recalled decreased significantly, from about 95% to 80% in both experiments, and the average location error increased significantly ($F > 34$, $p < .001$, in all cases). The data suggested that most subjects could track no more than 5 or 6 cars. The performance-based measures, on the other hand, were not affected by traffic level. These findings make sense, since global driving performance as well as hazard detection and blindspot avoidance depend on awareness of a relatively small number of cars near the driver; and the recall data show that at high traffic levels, subjects focus attention on a subset of cars.

But how do drivers determine where to attend when working memory is overloaded? We analyzed subjects' recall data using regression analysis to investigate the factors that affect where subjects allocate attention. In this analysis, we used various cues about individual traffic cars, such as how close they are to the driver, as predictor variables, and the location error in recalling individual cars as the dependent variable. For both Experiments 1 and 2, subjects recalled cars more accurately when they were: nearby, out of the blindspot, in front, moving towards them, and on the left. We assume that better recall accuracy follows from more frequent attention, and that the above factors are the ones subjects use to allocate attention amongst multiple traffic cars.

For both experiments, the regression analysis also showed that subjects' recall accuracy was lower on hazard trials than on catch trials, even though subjects recalled the car about to hit them (the hazard car) more accurately than other cars. This phenomenon of dealing with stressful situations by focusing attention on hazardous stimuli while de-emphasizing other stimuli has been called *cognitive tunneling* (Dirkin, 1983).

Conclusion

The implicit, performance-based measures of SA described here show promise as alternatives to explicit, recall measures. Conceptually, the imbedded-task measures are based not on overall success in performing the driving task but on subjects' awareness of stimuli at particular points in their environment (ahead/behind for hazard detection, and right/left for blindspot avoidance). Thus, these measures seem to be true measures of SA, as opposed to global performance measures. Empirically, the imbedded-task measures showed above-chance performance on both experiments and adequate reliability in Experiment 1. In Experiment 2, the reliabilities of these measures were somewhat low due to the small number of trials.

In Experiment 1, when reliable imbedded-task SA measures were available, these measures correlated positively with explicit, recall measures of SA. This suggests that both the performance-based and the recall measures were tapping into the same knowledge base -- subjects' explicit knowledge of vehicle locations. At least in the emergency driving situations used in Experiment 1, implicit knowledge does not have a large influence on subjects' driving performance.

In addition to investigating implicit measures of SA, we also used the explicit measures to assess how drivers allocate attention when working memory is overloaded. At high memory loads, the accuracy of recall and the number of cars recalled declined. Subjects seemed to be able to track no more than 5 or 6 cars. When working memory was overloaded, subjects used environmental cues to allocate attention, focusing more on cars that were nearby, in front of them, moving towards them, and on the left. The use of these cues suggests that subjects are adaptively allocating attention to cars that are potentially most hazardous. Drivers attention allocation was also affected by stress. When a car was about to hit them, subjects focused on that car and forgot

information about other traffic cars. All of these results concerning attention allocation were replicated in Experiment 2.

References

Adams, M. J., Tenney, Y. & Pew, R. (1995). Situation awareness and the cognitive management of complex systems. *Human Factors, 37*(1), 85-104.

Buchner, A., Funke, J. & Berry, D. (1995). Negative correlations between control performance and verbalizable knowledge: Indicators for implicit learning in process control tasks. *Quarterly Journal of Experimental Psychology, 48A*(1), 166-187.

Dirkin, G. (1983). Cognitive tunneling: Use of visual information under stress. *Perceptual and Motor Skills, 56*, 191-198.

Endsley, M. R. (1995a). Towards a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32-64.

Endsley, M. R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors, 37*(1), 65-84.

Fracker, M. L. & Davis, S. A. (1991). Explicit, Implicit, and Subjective Rating Measures of Situation Awareness in a Monitoring Task. (Technical Report AL-TR-1991-0091). Wright-Patterson Air Force Base, OH: Air Force Systems Command.

Grier, J. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75*(6), 424-429.

Perruchet, P & Amorin, M. A. (1992). Conscious knowledge and changes in performance in sequence learning: Evidence against dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(4), 785-800.

Roediger, H. & McDermott, K. (1993). Implicit memory in normal human subjects. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology, Vol. 8*. Amsterdam, Netherlands: Elsevier Science Publishers.

Watson, C. & Nichols, T. (1976). Detectability of auditory signals presented without defined observation intervals. *Journal of the Acoustical Society of America, 59*(3), 655-667.

Cognitive Correlates of Explicit and Implicit Measures of Situation Awareness

Leo Gugerty and William Tirre

Brooks Air Force Base

In this study we examined the correlation between various ability factors and explicit and implicit measures of situation awareness (SA) obtained from a driving simulator. Previous research has examined the correlates of SA measures in aircraft pilots. For example, Carretta and Ree (1995) reported a study in which a large battery of cognitive, perceptual, and psychomotor tasks were administered to a sample of 171 F-15 pilots. The dependent variable was the first unrotated principal component found on a set of peer and supervisor ratings of SA. The dependent variable was predicted well by F-15 flying hours (experience) but the only individual difference variable that had any incremental validity was a general cognitive ability composite. When F-15 flying experience was partialled out of both the dependent variable and the predictor variables, significant correlations were found for working memory, divided attention tests, and two psychomotor tests.

Of course, it is questionable whether the Carretta and Ree dependent variable actually reflected SA. Another interpretation of their dependent variable was that it reflected only general airmanship (piloting skill) as perceived by other pilots who presumably had ample occasion to make informed judgments. Objective, performance-based measurement of SA, perhaps in a flight simulator, would probably be much more informative. Along these lines, Endsley and Bolstad (1994) examined correlates of SA using the Situation Awareness Global Assessment Technique (SAGAT) and a battery of 18 cognitive, perceptual, and psychomotor tests. Because of the small sample (N = 21), strong conclusions about correlations with SAGAT are not warranted. But it is interesting to note that a psychomotor tracking task correlated 0.72 with SAGAT. Endsley and Bolstad suggested that pilots with superior psychomotor abilities had spare attentional capacity that could be devoted to situation assessment; consequently they demonstrated better SA.

The present study was similar to the Endsley and Bolstad study in that SA was measured objectively. It was dissimilar in that SA was measured in a less complex but still dynamic environment (simulated highway driving) on a sample of adults with a broader range of abilities than considered in either the Carretta and Ree or Endsley and Bolstad studies. The main hypothesis we sought to test was that implicit and explicit measures of SA would be correlated most highly with a working memory factor since SA is dependent on keeping a situation model updated in a changing environment. In addition, we were interested in correlations of SA with other ability factors that might be important in dynamic operator tasks, viz., temporal processing, visual search, and multilimb coordination. We examined the correlations of both implicit and explicit SA measurements with ability factors in three experiments.

Experiment 1

Method

This study was conducted in conjunction with a larger factor analytic study of the Cognitive Abilities Measurement (CAM) battery (Kyllonen, 1994). CAM attempts to comprehensively

measure the human abilities essential to the acquisition of knowledge and skill (Kyllonen, 1994). CAM, a battery of 59 computer-administered tests, was created through the use of a taxonomy. The six rows of the taxonomy reflect the major abilities suggested by cognitive psychology, viz., working memory, processing speed, induction, declarative knowledge, declarative (associative) learning, and procedural (skill) learning. The columns reflect three information domains suggested by psychometric analyses, viz., verbal, quantitative, and spatial. Within each cell are three or four minor rows which correspond to task paradigms or item types. These are repeated across columns. In the larger study, 230 civilians of both sexes between the ages 18 to 30 were administered the CAM and the Armed Services Vocational Aptitude Battery (ASVAB) over five days. A subset (N = 34) was also administered the driving simulator task (see Gugerty & Tirre, this volume, for full description).

Test battery scores were reduced to a manageable number in two steps. First, 18 cell scores for the six rows and three columns of the CAM taxonomy were formed, e.g., all spatial working memory tests were combined into one score. Second, the 18 CAM and 10 ASVAB scores were factor analyzed in separate runs. Factor analysis resulted in three CAM factors (general cognitive ability (g) / working memory, processing speed, and declarative knowledge) and two ASVAB factors (g and perceptual speed). Each factor set was orthogonally rotated using the quartimax, procedure which emphasizes the amount of variance explained by the first factor. The first CAM factor, g/working memory, had its highest loadings on the working memory and procedural learning tasks. This finding is consistent with prior research (e.g., Kyllonen & Christal, 1990; Kyllonen & Stephens, 1990; Tirre & Pena, 1993) that indicates working memory and tasks with heavy working memory requirements might be the core of the psychometric phenomenon known as general cognitive ability. The first ASVAB factor had its highest loadings on tests requiring quantitative reasoning skills, viz., arithmetic reasoning and math knowledge, both of which benefit from education. As such, ASVAB *g* reflects crystallized instead of fluid ability (Cattell, 1971).

Results

Both implicit and explicit measures of SA (see other Gugerty & Tirre article in this volume for definitions) were highly correlated with the CAM g/working memory factor and slightly less so with the ASVAB *g* factor. The implicit SA measure of hazard detection was correlated .60 ($p < .0005$) with g/working memory, and .39 with ASVAB *g* ($p < .03$). Likewise, blindspot avoidance correlated .62 ($p < .0005$) with CAM g/working memory and .44 ($p < .02$) with ASVAB g. The explicit measure of SA showed correlations of similar magnitude: average location error in recall correlated -.73 with g/working memory ($p < .0005$) and -.50 with ASVAB *g* ($p < .003$). Global performance as indexed by crash avoidance was correlated equally with CAM g/working memory and ASVAB *g* (.63, .60). None of the remaining ability factors correlated significantly with the SA measures. These results are consistent with the hypothesis that SA depends critically on the working memory system (Endsley, 1995).

Experiment 2

Method

The second experiment (N = 46) was also conducted as part of a larger factor analytic study, this time of cognitive, perceptual-motor, and temporal processing abilities. Two computer-administered test batteries and one paper-and-pencil battery (the Air Force Officer Qualifying Test, AFOQT) were administered to a sample of paid civilians of both sexes between the ages of 18 and 35. The computer-administered cognitive battery was a subset of the CAM 4.1 battery consisting

of one test from each of the 18 cells of the taxonomy. The perceptual-motor battery consisted of 17 tests designed to measure four of the Fleishman factors: multilimb coordination, control precision, rate control, and response orientation (Fleishman & Quaintance, 1984); and six tests measuring temporal processing. The temporal processing tasks required the subject to estimate short time intervals (e.g., click mouse when counter reaches 100 after seeing 0 - 25), estimate the arrival time of an object "moving" across the screen after its disappearance, or select the winner between two objects racing across the screen after both disappear behind a wall. The third test battery, the AFOQT, was not administered in its entirety -- four of the sixteen subtests were omitted because their content was more appropriate for a college-educated sample. These were reading comprehension, math knowledge, general science, and aviation information.

Because the subject to variable ratio at the time of this writing was too small to warrant factor analysis of the total dataset, we created composite scores corresponding to factors found in previous analyses of the CAM 4.0, perceptual-motor, and AFOQT batteries in which sample size was no problem. Considering CAM first, we created composites for working memory, processing speed, procedural learning, declarative (associative) learning, inductive reasoning, and declarative knowledge that correspond to factors found by Kyllonen (1993). For the perceptual-motor battery we simply created composites corresponding to Fleishman and Quaintance's (1984) factors since we had deliberately attempted to simulate original apparatus tests used by Fleishman in his factor analytic research. We made a separate temporal processing composite, since Tirre (1995) found that a similar factor loaded by two of the temporal processing tests was distinct from four perceptual-motor factors including multilimb coordination. AFOQT variance resolved down to seven factors in analysis by Goff and Tirre (1995): general cognitive ability, verbal, quantitative, spatial, technical knowledge, visual search, and academic/science knowledge. The only factor that might require some explanation is what we called visual search. The visual search factor was loaded by the table reading , scale reading, and block counting subtests. Each of these subtests appears to require the examinee to visually scan a printed stimulus (a table of numbers, a picture of stacked blocks, or a scale with markings) for a certain element or set of data. Each subtest requires attention to visual detail and in a way might be regarded as an occulomotor test since eye movements must be carefully controlled.

Given the results of these previous factor analyses, we then selected certain composite scores to serve as predictors of SA according to Endsley's model of SA (Endsley, 1995). In particular, we were interested in the correlations of working memory, temporal processing, visual search, and multilimb coordination with SA. The reasoning behind selection of these predictors is as follows:

- Working memory has a limited capacity that must be shared between current operations and temporary storage of intermediate results and freshly encoded data. It is the central bottleneck in information processing and its capacity varies considerably across individuals. Thus, we should expect SA to be limited by working memory capacity.
- Temporal processing is probably a component of performance in dynamic visual environments such as driving. For example, temporal processing is probably involved in maintaining a safe stopping distance between cars and in noticing and passing slower vehicles on the highway.
- Visual search might be involved in the perceptual level of SA in that the safe driver scans the environment, checking his mirrors and forward view to monitor the locations and actions of other vehicles.
- Lastly, multilimb coordination might be involved in SA in driving, not because psychomotor control is required, but rather because it reflects multi-tasking ability. If the subject understands his task to be one of monitoring other vehicles to avoid crashes while encoding and possibly rehearsing the locations of other vehicles for later recall, he is forced to time-share between activities. In the multilimb coordination tasks used in this study, the examinee must attend to moving stimuli and coordinate movements of hand and feet. This is essentially a multiple task situation.

The dependent variables were the same as Experiment 1 with the exception that the number of trials in the driving simulator was about half of that in Experiment 1.

Results

The correlations between the four SA measures and the predictor set (see Table 1) indicate that the explicit measure of SA (average location error in recall) correlated significantly with all predictors except temporal processing. In contrast, the implicit measures of SA (hazard detection and blindspot avoidance) were not correlated with any of the predictors. The global performance measure, percent hazards avoided, nonetheless correlated significantly with all predictors except temporal processing. Recall that crash avoidance correlated significantly with both location errors in recall ($r = -.35$, $p < .02$) and hazard detection ($r = .36$, $p < .02$).

Table 1. Correlations of Ability Predictors with Situation Awareness Measures (Experiment 2)

Predictor	SA Measure			Global Performance
	Avg. location error (in recall)	Hazard detection	Blindspot avoidance	Crash avoidance
Working memory	-.46 (.001)*	.05 (.748)	-.26 (.084)	.53 (.001)*
Multilimb coordination	-.50 (.001)*	.19 (.206)	-.23 (.132)	.41 (.005)*
Visual search	-.49 (.001)*	-.05 (.749)	-.12 (.411)	.30 (.045)*
Temporal processing	-.20 (.182)	.19 (.223)	-.03 (.848)	.22 (.141)

Note. N = 46. Pearson's r with probability level in parentheses.

The poor prediction of the implicit measures of SA is probably due in part to the lower reliability of these measures, which had been shortened by 50% in Experiment 2.

The best predictor of average location error in recall was multilimb coordination and the best predictor of crash avoidance was working memory (as indicated by multiple regression analyses), but this is the opposite of what we predicted. Keep in mind that all the predictors were substantially intercorrelated (.53 to .81) and the differences among correlations with the criteria were small. This degree of multicollinearity combined with the small sample size suggests caution in interpreting correlational patterns.

Experiment 3

Method

The third experiment was conducted in conjunction with the same factor analytic study as the second experiment. In this case, 88 civilians of both sexes between the ages of 18 and 30 were paid for their participation. In this experiment, we were interested in the effect of signal rate on detection of hazards. The driving simulator was configured for implicit SA measurement only -- no recall trials were administered. There were two groups, one which experienced hazards on 75% of the trials and a second which experienced hazards on only 25% of the trials. Both groups

first experienced practice in which hazards occurred on 50 to 60% of the trials. As it turns out, signal rate did not have a significant effect on any of the dependent variables, so we collapsed across groups for the correlational analyses.

Results

The SA and global-performance measures were significantly intercorrelated among themselves. Hazard detection was correlated .42 ($p < .001$) with blindspot avoidance and both correlated with crash avoidance ($r = .67$, $p < .001$, $r = .33$, $p < .002$). The correlations of the four ability predictors with the SA dependent variables (see Table 2) indicate that: (1) hazard detection is correlated significantly with all predictors, with working memory having the highest correlation, and (2) blindspot avoidance is correlated significantly with only two predictors, viz., working memory and visual search. The global performance measure, crash avoidance, was correlated with all predictors, but especially multilimb coordination.

Table 2.. Correlations of Ability Predictors with Situation Awareness Measures (Experiment 3)

Predictor	SA Measure		Global Performance
	Hazard detection	Blindspot avoidance	Crash Avoidance
Working Memory	.44 (.001)*	.32 (.002)*	.43 (.001)*
Multilimb Coordination	.31 (.004)*	.20 (.064)	.54 (.001)*
Visual Search	.25 (.02)*	.23 (.028)*	.28 (.007)*
Temporal Processing	.36 (.001)*	.16 (.131)	.46 (.001)*

Note. N = 88. Pearson's r with probability level in parentheses.

Multiple regression analyses with the dependent variables reflect the pattern displayed in Table 2. That is, working memory was the only significant predictor in equations for hazard detection and blindspot avoidance, and multilimb coordination was the only significant predictor of crash avoidance, though the contribution of working memory was nearly significant ($p < .08$).

Conclusions

Because of the small sample sizes available, multivariate analyses relating situation awareness variables to ability variables were not attempted. Since ability data are inherently multivariate, we cannot make strong conclusions about the correlates of SA. Some conclusions about how situation awareness relates to human abilities are nonetheless warranted when we search for relationships that emerge consistently across the three experiments. In drawing conclusions, we chose to exclude experiment 2 results with the implicit measures, since these were unexplainably discrepant.

- Our global performance measure (crash avoidance), which probably reflects both explicit and implicit components of SA as well as decision-making and response

execution, was moderately related to both working memory (.43 < r < .63) and multilimb coordination (.41 < r < .54). These two variables were the strongest predictors of crash avoidance.

- Our explicit measure of SA, location error in recall, was correlated -.46 to -.50 with working memory, visual search, and multilimb coordination in experiment 2. The true correlation with working memory may be considerably higher given an r of -.73 in experiment 1.
- Hazard detection was the better implicit SA measure, and it correlated best with working memory in experiments 1 and 3 (r = .60, .44, respectively).
- Though less reliable, blindspot avoidance also correlated with working memory in experiments 1 and 3 (r = .44, .32, respectively).

Given these results, a fairly safe conclusion is that individual differences in working memory are predictive of both implicit and explicit measures of situation awareness. This finding is consistent with the notion that working memory is pervasive in complex cognitive activities (Baddeley, 1986).

In future research we hope to increase the number and variety of SA measurements we obtain through the driving simulation. The ideal individual differences study would include multiple indicators of situation awareness in order to identify the factor structure. It is unknown whether a single factor or multiple factors are needed to account for the observed variance in SA measures. Until we know the factor structure, we cannot be certain whether relationships found between predictors and SA dependent variables reflect the correlation of the predictor with the core factor (common to all SA measures) or with the unique variance associated with particular SA measures.

References

Baddeley, A. (1986). *Working Memory*. Oxford: Oxford University Press.

Carretta, T.R. & Ree, M.J. (1995, May). *The SAINT program -- prediction of situation awareness in US Air Force F-15 pilots.* Paper presented at the Annual Scientific Meeting of the Aerospace Medical Society, Anaheim, CA.

Cattell, R.B. (1971). *Abilities: Structure, Growth, and Action*. Boston: Houghton-Mifflin.

Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32-64.

Endsley, M.R. & Bolstad, C. A. (1994). Individual differences in pilot situation awareness. *International Journal of Aviation Psychology, 4*(3), 241-264.

Fleishman, E.A. & Quaintance, M.K. (1984). *Taxonomies of Human Performance: The description of human tasks.* Orlando, FL: Academic Press.

Goff, G.N. & Tirre, W.C. (1995). *Confirmatory factor analysis of the Air Force Officer Qualifying Test, Forms O, P1 and P2.* Unpublished manuscript.

Kyllonen, P.C. (1993). Aptitude testing based on information processing: A test of the four-sources model. *Journal of General Psychology, 120*, 375-405.

Kyllonen, P. C. & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity? *Intelligence, 14*, 389-433.

Kyllonen, P.C. & Stephens, D. (1990). Cognitive abilities as determinants of success in acquiring logic skills. *Learning and Individual Differences, 2*(2), 129-160.

Kyllonen, P.C. (1994). CAM: A theoretical framework for cognitive abilities measurement. In D. Detterman (Ed.), *Current topics in human intelligence: Volume IV, Theories of intelligence* (pp. 307-359). Norwood, NJ: Ablex.

Tirre, W.C. (1995, May). *Steps toward a model of pilot aptitude and learning.* Paper presented at the Annual Scientific Meeting of the Aerospace Medical Society, Anaheim, CA.

Tirre, W. C., & Pena, M. C. (1993). Components of quantitative reasoning: General and group factors. *Intelligence, 17*, 501-522.

Situation Awareness Evaluation for an Operator Support System in a Nuclear Power Plant

Geert Uytterhoeven[1], Michel De Vlaminck[1] and Denis Javaux[2]

[1] Tractebel Energy Engineering, Belgium.
[2] University of Lige, Belgium.

Introduction

A new generation of nuclear power plant supervision systems, called DIMOS (DIstributed MOnitoring System) has been developed by TRACTEBEL/BELGATOM and installed at the DOEL nuclear power plant (Belgium). It has been monitoring units 1 and 2 since August 1991 and unit 3 since July 1993. It is presently in development for unit 4 (De Vlaminck and Gilliot, 1994)

DIMOS is a Computerized Operator Support System that is able to handle the raw information available within the plant in order to help the operator crew accomplish their supervisory task. Therefore, no control facilities must be provided. Controls are performed via conventional control panels.

A new functionality that has recently been developed is aimed at reducing the number of present alarms. It has been decided to validate those masking mechanisms on a full scope training simulator. This paper describes the experiments that have been carried out, starting in June 95, and the results that have been obtained.

The evaluation methodology attempts to compare two versions of DIMOS (with alarm-treatment and without alarm-treatment) on several normal situations as well as on incidental and accidental scenarios. Alarm treatment consists in assigning priorities to events allowing important process changes to catch the operator's eye faster than less important ones. Alarm-treatment also means that certain events will not be presented in certain circumstances, and that at other moments they will give an alarm-message on one or more specific operator stations. This mechanism is called 'alarm reduction', and is aimed at reducing useless information, mainly during avalanches when it is difficult to extract the information which is actually useful. Automatic sequence supervision is another mechanism to reduce the number of messages with might appear during an incident. It synthesizes the results of any automatic sequence (e.g. containment isolation) and verifies the actuator performances or verifies the chronology of the actions and events (e.g. diesel sequence after safety injection signal).

The comparison between the two versions of DIMOS is focused on the ability to maintain or enhance *Situation Awareness* (SA), as well as to sustain *performance* and provide *satisfaction* to the operators.

- SA has been chosen as the main criterion for comparing both systems, since an obvious danger with the new alarm-masking version is that it can hide some critical information that normally contributes to SA. SA itself is widely considered as "an essential prerequisite for the safe operation of any complex dynamic system" (Sarter & Woods, 1991), and achieving or enhancing SA is a major objective in the design of dynamic process control interfaces (Endsley & Bolstad, 1993; Wickens, 1992). The measuring method for measuring SA is based on an interruptive technique inspired from existing methods (Endsley & Bolstad, 1993, Hogg & al, 1994). Operators have to provide

answers to a set of questions focused on the current state of the process. Neutral (non-pertinent) questions have been added to the set of questions to avoid biases and attentional focusing on the portions of the process under evolution. As a commonly accepted and integrated definition of SA is still to come (Sarter & Woods, 1991), we have simply investigated two different attributes of situation awareness that we considered as important for our purpose : *scope* measures the range of the SA possessed by operators and *correctness* measures its quality. Scope has been measured as the proportion of the No idea answers to the questions administered to operators since we had introduced this item in the possible answers. Correctness has been evaluated in real-time during the execution of scenarios by a qualified instructor. The complexity of the experiment (variability of the possible process evolutions and operators actions) does not allow to easily implement an automatic and computerized verification of the answers.

- Performance has been chosen as another important criterion for comparing both versions of DIMOS because, besides safety, it is obviously a second factor that must be optimized in modern process control situations (Amalberti, 1995). Proving that one version of DIMOS is actually better in terms of performance than the other is a critical information for deciding which one must be introduced in a real power plant. Since experiments take place in incidental and accidental situations, performance has been measured as the ability to quickly, safely and economically restore the normal state of the process. In particular, it has been rated as an evaluation of the amount of money needed for replacing the hardware or equipment altered, broken or lost during the execution of the scenario.

- Subjective satisfaction of the operators is the third criterion used in this research. Satisfaction is usually viewed as a main contributing factor in usability evaluation methods (Nielsen, 1994), and it must be considered as mandatory if one wants to introduce in real power plants a system that will actively be used.

Most measures have been taken in situations where no diagnostic or corrective procedures are supposed to be applied. Maintaining a correct situation awareness, both in scope and correctness, is mandatory in these situations for dealing with incidents or accidents. Situations where use can be made of existing procedures have been studied with less involvement since it is considered that SA is less important in these situations where operators merely follow procedures. In such cases however, procedures have been adapted in order to take into account the availability of the monitoring system in the control room.

Experiment conditions

As explained before, the main goal of the tests was to find out whether one monitoring system (alarm-masking and non alarm-masking) was 'better' than the other one or not. In order to get the most accurate information, the following decisions have been made:

- tests had to be prepared in close cooperation with nuclear power plant experts;
- tests had to be done on the full-scope simulator of the nuclear power plant of Doel 4 at Kallo, Belgium in order to simulate normal operation as well as incidental, and accidental operation;
- test subjects had to be real power plant operators, who were used as well to the operation of the power plant, as well as to the monitoring system to be tested.

However, several constraints had to be taken into account. There was a poor availability of the operators and the full-scope simulator. Tests moreover had to be done on a small-scale mock-up of the monitoring system: the test results had to contribute to the decision whether full scale configuration could be started or not.

Experiment Scale

Situation Awareness tests have taken place during 9 not always subsequent days. 4 Operator teams or 17 different operators participated to the tests. 29 Scenarios have been 'played', which resulted in 7300 questions out of which 4640 were used to calculate the SA.

Experiment description

Briefing

Before starting the tests with a new operator team, operators were first briefed about the aim of the tests. They were explained that the test-results were used not to evaluate them personally but to evaluate the monitoring system. No information was given on the contents of the scenarios, neither on the sequence.

Questionnaires

At the beginning of the tests each operator team received a dummy questionnaire in order to make them familiar with the type of questions and the method of answering. All questions were multiple-choice questions with one of the options being 'No Idea'. The operator was asked to give this answer rather than to guess if he did not know the real answer.

Questions were asked by means of personal computers in order to facilitate post-analysis. Since the answers always were context-dependent, the instructor had to answer them in parallel in order to get the right answers. The questions tried to find out the Situation Awareness of each subject in a temporal perspective including the past, the present and the future: some questions checked if the operator was aware of some recent events, others tried to find out whether he had an idea about what could happen in the near future and still other questions just asked direct information on the actual state of the process. All these questions together were meant to build an image of the operator's SA. Special care had to be taken not to influence the operator's SA by means of those questions.

Test structure

Each operator team spent 2 days on the simulator for those tests: during one day, the four different scenarios were presented in a random order using either the old (non alarm-masking) or the new (alarm-masking) version of the DIMOS monitoring system. The second day, the same scenarios were presented but in a different order and with the other version of the monitoring system than the one used the first day for each scenario. Special care was taken in order to avoid that scenarios could be recognized the second time they were 'played'.

At pre-defined moments during each scenario, the simulator was 'frozen' and the test-subjects were asked to leave the control room in order to answer some pre-defined situation-dependent questions. Each interruption took about 6 minutes after which the scenario continued. Those questions had to be answered out of the control room in order to avoid that information could be taken from the display panels after the question had been asked.

About 3 question-interruptions were planned for each scenario. At the end of each scenario, the operators' opinion about the scenario, the questions and the monitoring system was asked. During the tests, all kinds of special events, like unexpected or wrong operator actions where recorded and discussed after each scenario. These records were used to evaluate the operator teams' performance.

Debriefing

At the end of all tests, all operators were asked to fill out a questionnaire that was meant to get their subjective judgments about the two versions of the monitoring system.

Results

Test results were divided into three groups: the Situation Awareness, the subjective judgment of the operators and the performance of the operators.

Situation Awareness

As explained before, Situation Awareness was measured using two parameters: *scope* and *correctness*. Results show that both parameters always vary in the same way, but sometimes scope improvement is higher than correctness improvement, meaning that the operators think they are much more aware of the process, where in fact they are not.

The variations were analyzed globally over all available data, per operator team, per scenario and finally per operator-role (e.g. primary circuit operator). We took the results for all operators for each version of DIMOS. Assuming that each operator is a representative sample of the population of the users, and assuming that all other test-parameters remained the same for all test-subjects (which obviously is almost impossible), we see that the significance-test on the hypothesis that the alarm-treatment version of DIMOS scores better than the version without alarm-treatment is not significant. This can be explained very easily because of the complexity of the tests which makes that we do not have enough samples to make sure that random events can be neglected.

The following conclusions can be made:

- On a global scale (meaning all data put together) an improvement of both parameters, scope and correctness was detected for the monitoring system using alarm-treatment.

Table 1. Global data

no alarm-treatment	81 %	64 %
alarm-treatment	86 %	68 %

- However, when those data are split-up to find out more detailed information, some remarkable facts can be noted: 3 out of the 4 teams make a very positive score, meaning an improvement with the monitoring system with alarm-treatment for scope as well as for rightness. The fourth one makes a negative score. This last team, however was the only team still using every day (in the Doel 2 unit) an older type of process-computer that resembles the version of DIMOS without alarm-treatment. After tests they explained that since everything was new for them, they even were more lost with the new monitoring system hiding information than they were with the one putting tons of messages on their screens just like their older system.

Table 2. Scope for SA per team

no alarm-treatment	86 %	78 %	79 %	84 %
alarm-treatment	78 %	80 %	88 %	93 %

Table 3. Correctness for SA per team

no alarm-treatment	70 %	59 %	66 %	66 %
alarm-treatment	65 %	67 %	67 %	72 %

- During the accident-scenario a notable improvement of both scope and correctness was found, while during the incident scenario the opposite seems to happen. Also, one normal operation scenario seems to give an SA-improvement, while the other does not at all.

Table 4. Scope for SA per scenario

no alarm-treatment	81 %	87 %	75 %	87 %
alarm-treatment	84 %	83 %	66 %	93 %

Table 5. Correctness for SA per scenario

no alarm-treatment	68 %	73 %	60 %	61 %
alarm-treatment	76 %	59 %	55 %	72 %

- All operators roles give a better score of SA with the monitoring system with alarm-treatment. One role catches the eye since both scores are much better than the others': the primary circuit operator (PRIM). This can be explained since this operator is the only one still using the monitoring system during incidents or accidents (the others are executing and reading procedures) and he gets much more useful and accurate information at these moments than he did before.

Table 6. Scope for SA per role

no alarm-treatment	80 %	84 %	79 %	82 %
alarm-treatment	85 %	87 %	84 %	86 %

Table 7. Correctness for SA per role

no alarm-treatment	60 %	67 %	63 %	67 %
alarm-treatment	70 %	69 %	66 %	67 %

Subjective Judgment

Two types of subjective information were found: oral information during and after the tests, and written information by means of a formal questionnaire. Both lead to exactly the same conclusions.

Although no information was given on which monitoring system was being used each time, operators could clearly distinguish the different versions, just by watching the quantity of the messages during the scenarios.

All users agreed that the reduced information quantity (by masking irrelevant information) is the only way not to get overwhelmed by tons of messages which will never be read and which make the monitoring system useless for incident or accident cases.

The improved quality of the information makes the operator feel much more comfortable in his corrective decision making process, leading to faster and better decisions.

Pre-treating information can help operators to remind to do some actions which otherwise and especially in stress-conditions might be forgotten.

Operator performance

Finally, a remarkable improvement of performance was detected using the monitoring system with alarm treatment. While several important and expensive equipment broke down (on simulator!) when use was made of the system without alarm-treatment, they were saved during the scenarios using the new version, since it gave clear and correct information at the right moment and at the right place.

Conclusions

One can conclude from the results that the alarm-treatment mechanism has no detrimental effect on SA. On the contrary, by allowing operators to focus attention on the important information, it actually helps to enhance situation awareness and improve performance. This is confirmed by the subjective opinion of operators : the alarm-treatment version of DIMOS is preferred to the other version.

Although the SA-method was very useful to help decide whether or not to implement the alarm-treatment on site, one should be aware that it was only one part of the decision, and that results would not have been sufficient if they were not combined with observation of operator performance and operators' acceptance.

Acknowledgment

The authors wish to acknowledge the Management and the personnel of the Doel nuclear power plant and the Scaldis training center for their efficient assistance.

References

Amalberti, R. (1995). *La Gestion des Systmes Risques*. Paris: Presses Universitaires de France.

De Vlaminck, M., Gilliot, B. (1994). *DIMOS: A New Generation of Nuclear Power Plant Process Monitoring Systems*. Proceedings of the PSAM II Conference, San Diego, CA.

Endsley, M., Boldstad, C. (1993). *Human Capabilities and Limitations in Situation Awareness*. AGARD Conference Proceedings 520.

Hogg, D.N., Folleso, K., Volden, F.S., Torralba, B. (1994). *SACRI: a Measure of Situation Awareness for Use in the Evaluation of Nuclear Power Plant Control Room Systems Providing Information about the Current Process State*. Paper presented at the IAEA Specialists Meeting on Advanced Information Methods and Artificial Intelligence in NPP Control Rooms, Halden, Norway, 13-15 Sept. 1994.

Nielsen, J.N., Mack, R.L. (1994). *Usability Inspection Methods*. New York: John Wiley & Sons.

Sarter N.B. and Woods D.D. (1991). Situation awareness : A critical but ill-defined problem. *The International Journal of Aviation Psychology*,1, pp. 45-57.

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

Evaluating Team Situation Awareness through Communication

Judith Orasanu

NASA-Ames Research Center

Accurate and timely situation awareness (SA) is widely acclaimed as a prerequisite of good performance in high risk dynamic environments like aviation (Billings, 1994; Hitchcock, 1994). However, identifying valid and reliable methods to assess SA has proven to be a challenge. One feature of performance common to many complex environments is that tasks are often performed by teams rather than by single individuals. A focus on teams rather than individuals operating within their natural contexts offers the unique opportunity of using communication as an unobtrusive and possibly meaningful measure of the team's SA.

If we accept the definition of SA as perception and comprehension of environmental events and projection of their future status (after Endsley, 1988), we can ask to what extent such perceptions, interpretations and projections are reflected in a team's communication. This paper focuses on the performance of commercial flight crews and the role of SA as they make flight-related decisions. Of greatest interest is communication in abnormal and emergency situations. In these cases it is critical that all crew members and extended resources beyond the flight deck know that there is a problem, what it is, and what the plan is for dealing with it. SA is the starting point.

The importance of establishing common SA between a flight crew and air traffic controllers is illustrated in the crash of the Avianca B-707 flying from Medellin, Colombia to JFK in New York (NTSB, 1991). After several holds that depleted their fuel, and despite several exchanges with ground, the crew was not successful in communicating the seriousness of their fuel situation to the controller in NY. He failed to give the flight the kind of special handling it required and it crashed due to fuel exhaustion. Had the controller fully understood the fuel condition, it is likely he would have treated the flight differently.

Three questions illustrated by the Avianca crash are addressed in this paper: (1) What is the role of SA in crew decision making? (2) What can we learn about team situation awareness from team communication? (3) Can communication be used to measure team situation awareness?

The Role of Situation Awareness in Aviation Decision Making

A model of aviation DM has been developed and described elsewhere (Orasanu & Fischer, in press). It involves two major components: Situation Assessment and Choosing a Course of Action. SA is the ingredient that launches the entire process and that serves as the link between the two components (Figure 1). Situation *assessment* is distinguished from situation *awareness* in that it refers to the active *process* by which situation awareness (the *state*) is achieved after initial perception of cues that signal the problem (Sarter & Woods, 1991). The left-hand loop of Figure 1 refers to the situation assessment process; the right-hand loop to the choice of a course of action. The process of assessing a situation and making decisions is iterative, reflecting both the dynamic nature of the situation (e.g., weather changes) and the effects of decisions and actions on the evolving situation. As illustrated in Figure 1, initial perception of a problem drives the situation

assessment process to clarify the nature of the problem and primes the crew to monitor or seek further information.

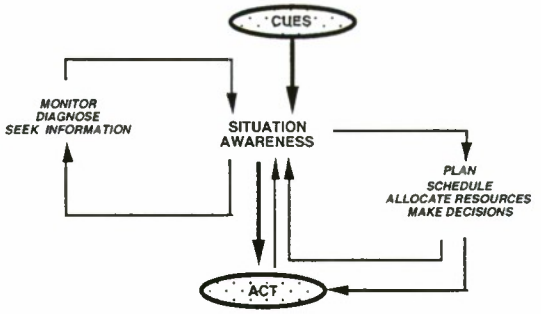


Figure 1. Situation awareness primes situation sssessment and grounds choice of s course of action

Three components characterize the situation assessment process in aviation: Defining the *nature* of the problem, determining how much *time* is available for coping with it, and assessing the level of *risk*, both immediate and in the future. Each of these components may change over time, requiring frequent updating. Once the situation has been perceived and interpreted, the updated situation model provides a basis for planning action and for making decisions (see also Klein, 1995).

Assessing Team Situation Awareness

How can we know the nature of a crew's situation awareness at any point in time? In principle, it can only be known indirectly; it is inferred from behaviors related to events in specific contexts. All methods for assessing SA (Endsley, 1994) yield only partial views, i.e., a discrete piece of information available at a point in time, self-reports filtered through memory and motive, or behaviors interpreted by an observer. The method we have used in our laboratory is essentially verbal process tracing (Ericsson & Simon, 1980; Woods, 1993). Professional flight crews flew a mission in a full-fidelity simulator. Their flight scenario was fraught with weather problems and system malfunctions. Domain experts defined cues needed for detecting and understanding the problems, likely developments, further actions to take, and additional cues to watch for. Crew communication was used instead of think-aloud protocols to assess the crews' situation understanding, goals, strategies and knowledge use. Presumably what crew members choose to talk about reflects what they think is important for handling the problem.

Two types of utterances were of particular interest--those involved in situation assessment and those related to planning a course of action. Situation assessment primarily includes observations on the state of the aircraft or flight, what we call *monitoring* utterances, and those that *seek information* that may be necessary for diagnosing a problem. In our analysis of crews "flying" the simulator, we found that the rate of monitoring utterances went up significantly following a system malfunction (hydraulic system failure) and bad weather which required a go-around at the original destination airport, and a decision about whether and where to divert (Orasanu, Gaddy &

Rodvold, 1995, based on Foushee, et al., 1986). Examples of monitoring utterances include, "OK, here comes the left main (gear)," "Not picking up a DME," or "Landing check's complete."

In the same experiment, the amount of information requested by the captain also increased following the onset of the problems. This information related directly to the condition of the aircraft and to decisions they needed to make about diversion, such as remaining system capabilities and limitations following the hydraulic system failure, weather and runway conditions at potential alternates, and fuel remaining (see also Mosier & Chidester, 1991). For example, "Do they expect any improvements in the weather around here?" or "Ask him how far he shows us from Lynchburg."

Captains of more effective crews (who made fewer operational or procedural errors) verbalized a greater number of plans than those of lower performing crews and requested and used more information in making their decisions. Was their SA greater? We don't know for sure, but can say that their processing was more information-intensive and forward looking.

Awareness of the situation can also serve as a basis for predicting possible future events or alerts or warnings. For example, "Keep your eye on that temperature." or "Coming up on the marker pretty soon." These types of utterances also increased following onset of the system failure and bad weather.

Crew SA can extend beyond the flight deck to ground personnel who may be helpful in an abnormal or emergency situation (McCoy, et al., 1995). Communication between the flight deck and ground (Dispatch or ATC) was examined to determine what information was shared and requested (Orasanu, Gaddy, & Rodvold, 1995). Three categories of information were examined: Telling about a Problem, Requesting Information, and Requesting Operational Assistance. Again, higher performing crews requested significantly more information (especially up-to-date runway wind reports) and operational assistance (i.e. vectors, long final approach, or holding patterns), which reflected awareness of the workload and temporal requirements of the problems and the resources available within the flight deck.

Basis for SA Analysis

What can we learn about team SA from analyzing team communication? Communication is an indirect method that relies on linguistic and logical analysis of precursors of an utterance. If we assume that utterances reflect perceived events rather than figments, then we may take an utterance "I see X," to mean that X exists and has been perceived by the speaker. Observations may be direct, such as, "Looks like a cat two up there," which indicates that cues signaling a weather condition have been both perceived and interpreted. Other observations may be less direct: On approaching a thunderstorm one captain commented, "...smell the rain. Smell it?" The first officer replied, "Yup. Got lightning in it too," which could be read as an indirect warning to avoid the storm based on awareness of its severity.

If someone requests additional information about the condition of X, we infer some uncertainty on the part of the speaker ("That was a heading of 265 out of Lynchburgh, wasn't it?") Presuppositions usually are not made explicit. "Ask him if he has any ride reports," may reflect concern over encountered turbulence and possibly deteriorating weather, even though these are not mentioned. "Call Dispatch and ask them if any other airport is open," implies a judgment that weather at the original destination was sufficiently poor that landing would be impossible and a diversion may be necessary, even though these are not stated.

Role of Communication in Enhancing SA

Perhaps equally important, communication is the means by which the SA of the team, both within the flight deck and on the ground, is expanded. Team SA is an emergent product of communication. As team members *verbalize* what they perceive in the environment around them,

they increase the team's knowledge. All members can then offer interpretations and predictions that can be evaluated and accepted or rejected.

Is it possible to improve team SA by training in situation assessment and communication skills? These might include specific labeling of the current situation, interpretation of the problem, assessment of risk level and time pressures, and anticipation of how the situation might evolve. These would serve as a basis for planning a course of action and making any necessary decisions. The question is whether enhanced SA would result from improving crew strategies for assessing the situation and communicating these assessments to other crewmembers. At present we have no data on this issue. Alternative methods would be needed to assess SA at a point in time to determine whether the interventions made a difference.

Can Team SA Be Measured by Communication?

The issue remains: What is the relation between communication and SA? More specifically, how completely does communication reflect underlying SA? The simple answer is that we don't know. Experiments are needed to establish the relations using converging operations. A major caution is that communication is a social phenomenon, influenced by social expectations and constraints. Some flight deck communication is fraught with risk, especially for junior crew members. Challenging a captain's judgment or pointing out errors may jeopardize a young first officer's career.

Communication also reflects personal style; some people are more loquacious than others. Does a higher level of communication reflect a higher level of SA or just a more expressive style? We have tried to address this issue by teasing apart different types of communication within the overall level of talk by each crew member. Problem-relevant talk has been identified and extracted from total talk during simulated flight (Orasanu, Gaddy & Rodvold, 1995). No differences between more and less effective captains were found in total amount of talk. However, a greater proportion of higher-performing captains' than lower performing captains' total talk was devoted to flight problems, i.e., statements of goals, plans, and information requests. This effect was greater during the high-workload abnormal phase of flight.

Pushing the question one notch further: Does a higher level of problem-related talk reflect higher SA? Certain temporal sequences between events, utterances, and actions suggest but do not confirm such a relation. For example, in our scenario all higher performing crews voiced awareness of a limitation on restarting the auxiliarry power unit (APU) in flight (which had malfunctioned at the gate); none of the lower performing crews commented on it, and in fact tried to restart the APU in flight significantly more often than higher performing crews (Orasanu & Jobe, 1994, based on Wiener, et al., 1991), suggesting a relation between awareness, communication and action.

The most serious shortcoming of this approach is that crews may be aware of more than they say. Given their responsibility for flight safety and the time pressure of many situations, crew members are likely to talk about problems they perceive and how to solve them. However, in the absence of converging operations we have no way of calibrating the relation between their actual awareness and their verbalization of it. What we can say is that crews that verbalize more about problems in fact perform at a higher level, commiting fewer operational and procedural errors. Their verbalizations appear to be useful for developing shared models for the problems and for assuring coordinated actions in dealing with them (Cannon-Bowers, Salas & Converse, 1994; Orasanu, 1994).

Communication also fails to reflect moment to moment changes in SA. Presumably it lags behind the event of interest some indeterminate amount of time. It would be necessary to know for sure when a cue has become available to the crew to be able to assess the lag in perception and

interpretation. Communication also fails to inform us of any differences in SA among team members unless they are explicitly stated.

At this point we can only hypothesize a relation between problem-related communication and underlying SA. However, because communication data can be collected unobtrusively, because of its face validity, and because of the established importance of communication to flight safety and performance, we recommend the use of communication as an indirect indicator of SA. Further analyses using converging methods will be needed to determine how much we can trust in what we hear.

Acknowledgements

Funding for the research on which this paper is based was received from NASA Code UL and the FAA Office of the Chief Scientific and Technical Advisor for Human Factors. Thanks are also extended to Jeannie Davison for assistance in preparation of the manuscript and to C. Elaine McCoy for comments on an earlier version.

References

Billings, C. E. (1994). Situation awareness in complex systems: A commentary. In R. D. Gilson, D. J. Garland & J. M. Koonce (Eds.), *Situation awareness in complex systems,* (pp. 321-325). Daytona Beach, Florida: Embry-Riddle Aeronautical University Press.

Cannon-Bowers, J. A., Salas, E., & Converse, S. E. (1994). Shared mental models in expert team decision making. In N. J. Castellan, Jr. (Ed.), *Current issues in individual and group decision making*. Hillsdale, New Jersey: Lawrence Erlbaum.

Endsley, M. (1988). Situation awareness global assessment technique (SAGAT). *Proceedings of the national aerospace and electronics conference (NAECON).* New York: IEEE.

Endsley, M. R. (1994). Situation awareness in dynamic human decision making: Theory. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situation awareness in complex systems,* (pp. 27-58). Daytona Beach, Florida: Embry-Riddle Aeronautical University Press.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215-251.

Foushee, H. C., Lauber, J. K., Baetge, M. M., & Acomb, D. B. (1986). *Crew factors in flight operations: III. The operational significance of exposure to short-haul air transport operations.* (Tech. Mem. No. 88322). Moffett Field, CA: NASA-Ames Research Center.

Hitchcock, L. (1994). Yours, mine, and ours: Some observations on the metaphysics of situational awareness. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situation awareness in complex systems,* (pp. 3-16). Daytona Beach, Florida: Embry-Riddle Aeronautical University Press.

Klein, G.A. (1995). *A user's guide to naturalistic decision making.* Report prepared under contract No. DASW0-1-94-M-9906 for the U.S. Army Research Institute. Fairborn, OH: Klein Associates Inc.

McCoy, C. E., Orasanu, J., Smith, P. J., VanHorn, A., Billings, C., Denning, R., Rodvold, M., & Gee, T. (1995). Situational awareness at different levels of abstraction: The distributed cooperative problem solving domain of ATCSCC-Airline operations. Paper presented at the International Conference on Experimental Analysis and Measurement of Situation Awareness. Daytona Beach, FL.

Mosier, K. L., & Chidester, T. R. (1991). Situation assessment and situation awareness in a team setting. *Designing for Everyone: Proceedings of the 11th Congress of the International Ergonomics Association*, Paris, 15-20 July.

National Transportation Safety Board. (1991). *Aircraft Accident Report--Avianca, The Airline of Colombia, Boeing 707-321B, HK2016, Fuel exhaustion, Cove Neck, New York, January 25, 1990,* (NTSB/AAR-91-04), Washington, DC.

Orasanu, J. (1994). Shared problem models and flight crew performance. In N. Johnston , N. McDonald, & R. Fuller (Eds.) *Aviation psychology in practice*. Aldershot, UK: Ashgate (pp. 255-285).

Orasanu, J., & Fischer, U. (in press). Finding decisions in natural environments: The view from the cockpit. To appear in C. Zsambok & G. Klein (Eds.). *Naturalistic Decision Making*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Orasanu, J., Gaddy, M., & Rodvold., M. (1995). Air-ground communication: Extending resource management. Poster presented at the annual meeting of the Judgment and Decision Making Society, Los Angeles, CA.

Orasanu, J. & Jobe, K. (1994). Adaptive problem solving strategies in air transport flight crews. Paper presented at the annual meeting of the Psychonomic Society, St. Louis, MO.

Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology, 1,* 45-57.

Wiener, E. L., Chidester, T. R., Kanki, B. G., Palmer, E. A., Curry, R. E., & Gregorich, S. E. (1991). *The Impact of Cockpit Automation on Crew Coordination and Communication: 1.* (Contractor Report #177587). Moffett Field, CA: NASA-Ames Research Center.

Woods, D. D. (1993). Process-tracing methods for the study of cognition outside of the experimental psychology laboratory. In G. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok, (Eds.). *Decision making in action: Models and methods*. (pp. 228-250). Norwood, NJ: Ablex Publishers.

Situation Awareness and Older Workers

Cheryl A. Bolstad and Thomas M. Hess

North Carolina State University

Introduction

Since 1900, there has been a "graying of America", with the average age of the population steadily rising. Older adults (over age 65) now comprise over 13% of the population, as compared to 4% in 1900, and this number is projected to be 20% by the year 2030 (Moody, 1994). During the next two decades, population growth will be concentrated among those individuals over the age of 50 as the baby boomers will become senior citizens. Along with this shift in the population structure is an increasing concern with the capabilities of and problems faced by older adults, especially given that a decrease in the size of the workforce and changing retirement practices may lead to greater levels of employment of older adults than seen in the past.

Although typically studied in younger adults, we would like to argue that the study of Situation Awareness (SA) may be a useful way of assessing the specific problems faced by older workers. Obviously, SA is just as important for older and middle-aged adults as it is for younger individuals and future research in SA needs to address these populations. This is especially true now that SA work is being conducted outside of the military, such as in nuclear power plants and air traffic control towers, where the subject population is likely to be more heterogeneous and SA acquisition in older adults may be more of a concern. We would also like to argue that the study of SA is not just useful for work-related activities, but in everyday situations as well. For example, SA is as important to the older adult trying to cross a busy street as it is to a young pilot trying to shoot down the enemy. In both situations a life may be lost, but good SA could lead to more positive outcomes. Thus, the formation of SA may be critical to continued well-being and adaptive functioning in older adults.

The study of aging and SA may present unique problems to researchers. Presently, most of the theoretical work on SA and its measurement methods rely on studies in which the subject population consists either entirely of college students or young military pilots. Due to cognitive changes that may begin occurring during the middle years of a person's lifetime, however, both the formation and assessment of SA could be affected. Age-related changes in cognitive capabilities and experience may force us to examine current operationalizations of SA and their applicability to different age groups. It is the goal of this paper to examine aging-related cognitive changes that may affect a persons' ability to acquire SA. Whereas we will not address specific SA measurement methodologies, much of what we have to say will have implications for the assessment of SA across the adult lifespan.

Endsley's Theory of Situation Awareness

There are several theories that describe the formation of SA (e. g., Endsley, 1995; Adams, Tenney, and Pew, 1995). While each take different perspectives, they embrace some of the same

ideas. For present purposes, however, we felt it was useful to examine age-related cognitive changes and SA acquisition within the context of one model that addresses information processing since much of the research on cognition and aging has been done within this perspective. For this reason, we chose Endsley's (1995) theory because of its close link with information-processing theory, which in turn should facilitate our development of hypotheses about aging. Endsley proposes that situation awareness is comprised of three hierarchical phases. These phases form the primary components of her definition of SA.

> Situation awareness is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future (Endsley, 1988).

Each of the phases of Endsley's definition of SA will be addressed below and how age-related cognitive changes may affect their formation.

Level 1 SA

Currently, it is accepted that one of the first steps towards the formation of SA is the perception of information within a given situation (Salas, Prince, Baker, and Shreshta, 1995). It is this active perception that allows individuals to extract important components from their environment (Dominguez, 1994). Endsley (1988, 1990, 1995) refers to this state as Level 1 SA: "The perception of the elements in the environment within a volume of time and space .." The individual perceives the elements by selectively attending to the incoming stimuli. Through this selective attention, important/essential information is attended to while nonessential items are disregarded.

As we age, certain aspects of our ability to selectively attend to information decline (for a review, see Plude, Schwartz and Murphy, in press). This is especially true in situations where the target information competes with other information in the environment. One problem that older adults have has to do with their ability to inhibit non-selected information. Relative to younger adults, older individuals may have more difficulty suppressing or inhibiting unattended information. This has the effect of limiting working-memory resources as precious capacity is taken up by irrelevant information (Hasher and Zacks, 1988). Studies have also shown that selective attention is negatively affected by aging when the individual must engage in visual search. Basically, older adults exhibit disproportionate increases in time to search for a visual target as the number of items in the visual array increases. Finally, such effects are not just limited to the visual domain. Research has also shown that aging has a negative impact on the ability to actively filter or select out information from multiple sources in the auditory domain. This is especially true in tasks that place a heavy demand on the limited attentional capacities of a person. For example, a familiar auditory selection task includes the "cocktail party phenomenon," in which a person can selectively attend to many of the conversations going on around the room while ignoring others. Relative to younger adults, older adults demonstrate a greater susceptibility to interference in some situations when compared to younger adults (Barr and Giambra, 1990). In essence, then, studies of attention in aging have demonstrated that older adults have problems in selecting information from the environment when they must search the environment and when the amount of irrelevant or competing information is great.

While studies have shown that some aspects of selective attention decline in later life, this does not appear to be the case in every situation. When environmental support in the form of distinctive cues and preexisting information regarding spatial location are available to an adult, they can overcome many of these problems (Plude et al., in press). In addition, it should be noted that the problems just noted typically occur within novel contexts (e.g., standard laboratory tasks). In situations where the subject has expertise, many of these effects may be greatly reduced or eliminated as demonstrated in several studies of expertise that use domain-specific tasks (e.g., Morrow, Leirer, Fitzsimmons, and Altieri, 1994). Experts have the ability to activate appropriate

schemas from long-term memory to aid in performance in domain-specific tasks. The schemas allow them to focus their attention on the appropriate information as well as help direct their attention to where information may be presented through use of probabilistic information. These schemas also aid in the inhibition of irrelevant, nonessential items.

Thus, if SA is measured using experts for the given situation, as it currently is, attentional deficiencies associated with aging may not be much of a concern. If and when SA is measured for novices, however, these decrements may play a part in a persons' formation of SA. It should be noted also that the ability of certain tests to predict one's ability at SA formation may be age-related. For example, if standardized tests assessing individual differences in basic skills thought to be associated with the formation of SA are used in selection or performance assessment, older experts may be penalized by their poor performance on content-free tasks. The inability of such tests to assess performance in context would be an important concern in any attempts to promote their use.

Level 2 SA

Once individuals have perceived the information in the environment, the next step is to integrate and comprehend the information in working memory (Salas et al., 1995). The information is brought into consciousness, thus allowing the person to meld the information into a coherent picture (Dominguez, 1994). Endsley (1988, 1990, 1995) refers to this state as Level 2 SA: "..the comprehension of their (elements) meaning". Many authors refer to this "product" of SA as a mental model. It is the elements in the environment that a person perceives which activate the schemas in working memory and updates their mental model (SA) (Dominguez, 1994). In essence, they use their mental models/schemata from long term memory to enhance/clarify the situation (Salas, et al., 1995). However, SA is temporal in nature and this process would need to be continuous in order for a coherent picture of the situation to remain.

It is this Level 2 SA that may be the most troublesome for older adults. While older individuals may have no difficulties perceiving relevant elements within the environment, they may have problems retrieving this information from memory as well as perceiving from where it came. Many researchers believe retrieval problems are due to the incomplete encoding of contextual information in the first place. Studies have shown that there is an age-related impairment in the processing component concerned with the retrieval of information from memory (Salthouse, 1992).

In addition, once information is activated in memory, older adults may have difficulty identifying the source of the information and, thereby, discerning both where the information came from and what actually occurred versus what was inferred from existing schemas. An example of this type of problem is clearly demonstrated in studies of false fame (e.g., Jennings and Jacoby, 1993), where older adults are more likely than younger adults to misattribute fame to nonfamous but previously viewed names because they have difficulty identifying the source of their feelings of familiarity. In general, age has a more negative effect on memory in situations involving conscious recollection processes (i.e., direct tests) than in those involving automatic retrieval (i.e., indirect tests) (Howard, in press). Interestingly, it might be hypothesized that experts would be more susceptible to false familiarity problems associated with aging than would be novices. Experts possess many mental models of various situations and may have difficultly discerning between what actually occurred and what was activated in their memory by these models.

Such problems may prove particularly troublesome in situations where it is important not only to be aware of information essential to SA, but also to know the origins of such information. Through the operation of automatic encoding and retrieval processes, older adults may have much of the same information available to them as younger adults. However, as Endsley (1995) points out, it is important in many situations to be able to identify not only what information is available, but how this information became available. Uncertainty about the validity of the information used in SA formation may result in the construction of more tenuous mental models that may have a negative impact on decision making. The automatic processing associated with many scenarios

may result in problems for individuals of all ages, but these problems may be exacerbated with age as the ability to consciously control and monitor working memory processes suffers.

Thus, Endsley's Level 2 SA has the greatest potential to be effected by age related changes. This may be especially apparent when using many of the SA measurement methods. In particular, explicit tests of SA will be the most affected as these directly test an individuals' SA. While expertise may be able to negate some of these more negative effects associated with aging, older novices, will be at a disadvantage relative to younger novices.

Level 3 SA

As we have noted, changes in cognitive abilities with age may have important consequences on the formation of Level 1 and Level 2 SA, but what happens to Level 3 SA? Endsley (1988, 1990, 1995) refers to Level 3 SA as the "..the projection of their (elements) status in the future." Through the comprehension of the activated schemata, the individual is guided in their projection of future status as well as their selection of future actions (Endsley, 1995).

Both Level 2 and Level 3 SA involve the continuous extraction and updating of information from working memory. Numerous studies have shown that working-memory capacity and the efficiency of associated processes declines with age (Salthouse, 1991; Salthouse and Babcock, 1991). It may be this decline in working memory capacity that has the greatest effect on Level 3 SA. As we have already noted, these problems in working-memory functioning are most evident in tasks that require a great deal of deliberate, conscious processing (Smith and Earles, in press). Age-related limitations in working-memory processes may lead to difficulties in gaining a coherent picture of the environment as well as making future projections since it can be reasonably assumed that at least some aspects of these functions require deliberate processing. For many experts, many of these processes may be automatic. Thus, with extensive practice, older adults may be able to compensate for some of their working memory declines as they develop automaticity of a skill (Bosman and Charness, in press). Endsley (1995) also believes that attentional limitations can also be somewhat overcome through automaticity of a skill. Such optimism may have to be tempered, however, based on evidence that age may even limit the development of automaticity in certain types of situations (e.g., Rogers, Fisk, and Hertzog, 1994).

Conclusion

The age-related cognitive changes affecting the formation of SA that we have discussed above exhibit a great deal of interindividual variability. Therefore, not all middle-aged or older individuals will exhibit these changes, nor will they be affected in the same manner. However, researchers in the area of SA, particularly in fields were SA may be studied with novices, should be aware of these cognitive changes. They can not only affect the formation of SA, but they can influence the utility of many measurement methods.

References

Adams, M. J., Tenney, Y. J., & Pew, R. W. (1995). Situation awareness and the cognitive management of complex systems. *Human Factors, 37* (1), 85-104.

Barr, R. A., & Giambra, L. M. (1990). Age-related decrement in auditory selective attention. *Psychology and Aging, 5,* 597-599.

Bosman, E. A., & Charness, N. (in press). Age differences in skilled performance and skill acquisition. In F. Blanchard-Fields & T. M. Hess (Eds.), *Perspectives on cognitive change in adulthood and aging*. New York: McGraw-Hill.

Dominguez, C. (1994). Can SA be defined? In M. Vidulich, C. Dominguez, E. Vogel, & G. McMillan, *Situation awareness: Papers and annotated bibliography* (AL/CF Publication No. TR-1994-0085), pp. 5-15). Armstrong Laboratory: Wright Patterson Air Force base.

Endsley, M. R. (1995). Towards a theory of situation awareness. *Human Factors, 37(1),* 32-64.

Endsley, M. R. (1990). *Situation awareness in dynamic human decision making: Theory and measurement* (NOR-DOC 90-49). Hawthorne, CA: Northrop Corporation.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society 32nd Annual Meeting,* (pp. 97-101). Santa Monica, CA: Human Factors Society.

Hasher, L. & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 193-225). Orlando: Academic Press.

Howard, D. V. (in press). The aging of implicit and explicit memory. In F. Blanchard-Fields & T. M. Hess (Eds.), *Perspectives on cognitive change in adulthood and aging*. New York: McGraw-Hill.

Jennings, J. M. & Jacoby, L. L. (1993). Automatic versus intentional uses of memory: Aging, attention and control. *Psychology and Aging, 8(2),* 283-293.

Moody, H. (1994) *Aging: Concepts and Controversies*. Thousand Oaks, CA: Pine Forge Press.

Morrow, D. G., Leirer, V. O., Fitzsimmons, C., & Altieri, P. A. (1994). When expertise reduces age differences in performance. *Psychology and Aging, 9,* 134-148.

Plude, D. J., Schwartz, L. K., & Murphy, L. J. (in press). Active selection and inhibition in the aging of attention. In F. Blanchard-Fields & T. M. Hess (Eds.), *Perspectives on cognitive change in adulthood and aging*. New York: McGraw-Hill.

Rogers, W. A., Fisk, A. D., & Hertzog, C. (1994). Do ability-performance relationships differentiate age and practice effects in visual search. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 710-738.

Salas, E., Prince, C., Baker, D. P., & Shreshta, L. (1995). Situation Awareness in Team Performance: Implications for Measurement and Training. *Human Factors, 37(1),* 123-136.

Salthouse, T. A. (1991). *Theoretical perspectives on cognitive aging*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Salthouse, T. A. (1992). The information processing perspective on cognitive aging. In R. J. Stemberg & C. A. Berg (Eds.), *Intellectual development.* New York: Cambridge University Press.

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology,* 27, 763-776.

Smith, A. D. & Earles, J. L. (in press). Memory changes in normal aging. In F. Blanchard-Fields & T. M. Hess (Eds.), *Perspectives on cognitive change in adulthood and aging.* New York: McGraw-Hill.

Expertise and Chess: A Pilot Study Comparing Situation Awareness Methodologies

Francis T. Durso[1], Todd R. Truitt[1], Carla A. Hackworth[1], Jerry M. Crutchfield[1], Danko Nikolic[1], Peter M. Moertl[1], Daryl Ohrt[1], and Carol A. Manning[2]

[1] University of Oklahoma
[2] F.A.A. Civil Aeromedical Institute

Situation awareness (SA) has received a considerable amount of attention in the recent literature. However, no agreed upon definition or methodology for the measurement of this phenomenon currently exists. Many definitions have been proposed and have provided perspectives varied in scope. For example, the definitions provided by Endsley (1988) and Mogford (1994), although not incompatible, present different viewpoints regarding SA. Despite the lack of a common definition, our experiment focused on finding a sensitive procedure that would best differentiate among levels of expertise or skill and hence, SA.

A number of different methodologies have been explored in an effort to understand how operators develop and maintain a "picture" of the situation in which they are involved. Methodologies previously used have included verbal protocol analysis (e.g., Ericsson & Simon, 1994; Ohnemus & Biers, 1993; Sullivan & Blackman, 1991), retrospective event recall (e.g., de Groot, 1965; Kibbe, 1988), concurrent memory probes such as the *Situation Awareness Global Assessment Technique* (SAGAT; Endsley, 1988), and physiological measures such as eye movements (e.g., Moray & Rotenberg, 1989; Stein, 1989; Wierwille & Eggemeier, 1993). Beginning with the assumption that experts have better SA than novices, we compared five very different procedures in one study: verbal protocols, eye movements, on-line queries with the situation present, on-line queries with the situation removed (as in SAGAT), and post-hoc recollection.

We chose to look at chess expertise for several reasons. First, chess has been correctly called the drosophila of cognitive psychology (Charness, 1989), and its long history of study should serve us well in understanding SA. Second, it seems to us that, perhaps more than most activities, differences in chess expertise are differences in SA. The entire game involves assessing the relationship between existing pieces and predicting impending moves. For example, input and output processes are simple and are unlikely to distort the internal model of the situation. Third, chess players are ranked by the United States Chess Federation (USCF) and the differences among rankings are well understood. Finally, our ultimate interest is in understanding SA in air traffic controllers, and chess provides a God's-eye perspective of a number of different entities that make it a nice, albeit limited, laboratory analog of air traffic control.

General Methodology

All participants monitored four chess games. We asked participants to monitor, rather than play, the games to allow control over the entire game. All players experienced exactly the same game

regardless of their skill level, allowing us, for example, to insert queries at identical points for each player. To ensure involvement, the participants were asked to monitor the game for imminent material losses. This is a clear component of chess and allowed us to engage the participants even though they were simply monitoring an existing game.

Table 1. Overview of experimental procedures.

	Game 1	Game 2	Game 3	Game 4
Opening	Queen's Gambit (Slav defense/ Main line)	Queen's Indian (Main line)	Caro-Kann (Capablanca /Main line)	Sicilian (Alapin)
Outcome	White mates in 60	Black mates in 74	White mates in 52	Black mates in 61
Number of captures	8	7	12	8
Monitor material loss	Yes	Yes	Yes	Yes
Method tested	Verbal protocols	Eye movements	Situation-present queries	Situation-absent queries (SAGAT like)
Post-hoc recall	Yes	Yes	Yes	Yes

In each game, the participant sat 42" from a projected image of a chess board that subtended a visual angle of 40(, with each square subtending about 5(. At the beginning of each ply, a piece blinked twice, moved, and then flashed twice again. The position remained for 15 sec. The game was stopped after 80 plies (40 moves).

A chess expert (USCF 2100; Expert), an intermediate (USCF 1607; Class B), and a novice (USCF 1374; Class D) monitored four high-level games generated by a commercial computer program, *Chessmaster 4000* (. Characteristics of each game, and what methodological procedures were employed, appear in Table 1. The games were generated by setting both white and black on *Chessmaster's* Chessmaster-level and having the computer play itself. Games were randomly assigned to order of presentation. In all games, moves of both white and black were made while the participant monitored the game to determine when a piece was about to be captured. Across games there were 7 to 12 plies or sequences of plies during which pieces were captured.

In this pilot study, the use of only three participants makes the statistical discovery of the most sensitive procedure problematic. In part we relied on visual inspection of the data, looking for clear differences across levels of expertise. In addition, we conducted item analyses using materials (e.g., queries) rather than participants as the random variate. This allows us to generalize to other probes or situations for these participants, but does not allow us, for example, to discriminate between effects caused by the particular participant and effects caused by his level of skill. All tests were conducted at an alpha level of .05.

Anticipating Material Loss

As the cover task, the participant used a joystick to register his judgment about the imminent loss of material. If he believed that a piece would be taken "in the near future", he was to pull the joystick back; if he believed that a piece would not be taken in the near future, he pushed the stick forward. "Near future" was intentionally left vague so as not to influence the distance in the future that the participant normally considered. Confidence was indicated by the extent to which the stick

was pushed or pulled. If the participant did not have any feeling about the upcoming state of affairs, he was to rest the stick in the middle, neutral position.

We began analysis by determining the plies during which a take occurred. To control for strategies (such as always anticipate loss of material) we chose, for each game, a comparable number of plies in which a take did not occur. Thus, if a person did invariably pull the joystick back, he would do well anticipating loss of material, but poorly anticipating when no material loss would occur. Performance of a theoretical perfect player who moved optimally from take to no-take and back was calculated. For each critical ply in the game, we computed the point prior to that occurrence when the participant changed his judgment from "no take" to "take," or from "take" to "no take," and compared it to the theoretical perfect participant. Scores could range from 0 sec, if the ply was never correctly detected, to the theoretical optimum.

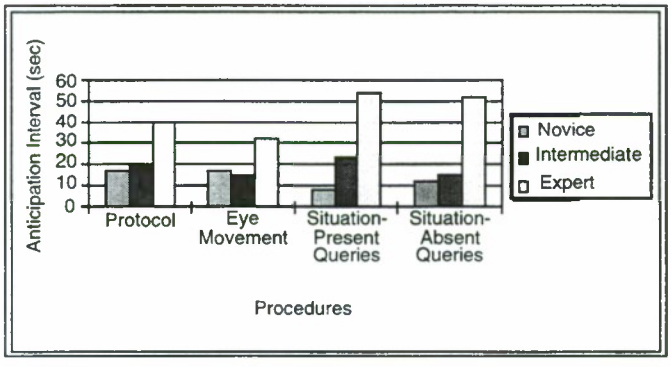


Figure 1. Projecting material loss

Not surprisingly, the expert anticipated material losses (47 sec; 4 plies) sooner than did the class B player (19 sec; 2 plies) or the class D player (13 sec; 1 ply). Results are shown in Figure 1. Regardless of the concurrent procedure, the expert was easily distinguishable from the two nonexperts. A repeated measures ANOVA revealed a significant main effect of skill, $F(2, 132) = 14.82$. Although there was no significant effect of method nor any interaction of method with skill, the figure suggests that only when the participants were in the Situation-present Queries condition were all three levels of expertise distinguishable. This may suggest that the Situation-present Queries condition is most likely to reflect individual differences in SA.

There has been some concern that interrupting the participant could, in itself, interfere with SA (e.g., Sarter & Woods, 1991). However, in the current study, there was little indication for the novice or the intermediate that the particular procedure had any contaminating effect on SA. While verbal protocols have been criticized as obtrusive and unreliable (Nisbett & Wilson, 1977), recording eye movements is touted as a relatively unobtrusive measure of recording behavioral data. However, in our data there was no effect of methodology on anticipating material loss, and what little effect may have been present argues for the use of query techniques.

Verbal Protocols

Verbal protocols were gathered during the first game. Participants were told to ". . .talk out loud. Try and verbalize your thoughts about the chess game that you are watching. For example, talk about pieces that are about to be taken, tactics or strategies that you notice are being used, or just general comments about the game such as which side currently has an advantage. We do not expect you to comment about anything in particular, just tell us basically what you are thinking about regarding the game. . . ."

Protocols were transcribed, segmented into plies, and then categorized by the experimenters. Comments relevant to the game were coded into four categories: Predictions, Assessments, Identifications, and Other. Predictions were sentences that described possible future moves (e.g., "The rook will go to H3 and attack the queen."). Assessments were sentences that characterized the ongoing flow of the game without making any predictions (e.g., "White needs to avoid exchanges because he has the developmental advantage."). Identifications were utterances that identified specific moves or tactics (e.g., "The knight is pinned; This is a Queen's gambit."). Comments that did not fit these categories were rare and were excluded from further analyses.

Overall, the intermediate participant made the most utterances (N = 184), the expert the least (N = 94), with our novice (N = 106) falling between these two. This pattern of overall utterances was consistent with the intermediate effect (Grant & Marsden, 1988; Schmidt & Boshuizen, 1993). When the utterances are classified as in Figure 2, it becomes apparent that the participants uttered different types of comments depending on their level of expertise, $\chi^2(4) = 9.49$. Our expert produced mostly predictions (64%), more than either the intermediate or the novice, $\chi^2(2) = 12.46$. The intermediate, on the other hand, produced the largest number of statements assessing the situation, $\chi^2(2) = 76.03$. His corpus was 62% assessment, with only 27% predictions and 10% identifications. Finally, the novice's protocol consisted of 50% assessment, 25% prediction, and 19% identification, a profile too similar to the intermediate's to allow recommendation of this procedure, resulting in a nonsignificant $\chi^2(2) = 4.0$, *NS*.

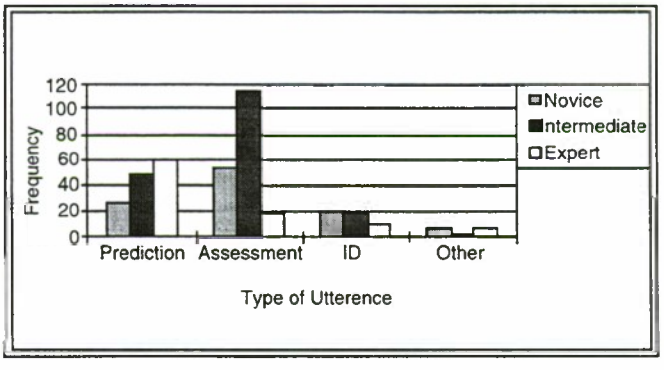


Figure 2. Protocol analysis

Eye Movements

During game two, eye movements were recorded using an Applied Science Laboratories (ASL) series 4000 eye tracker. The eye tracker utilized a magnetic headtracking apparatus in order to compensate for any head movements made by the participant. Eye movement data were recorded in two, 10 minute sessions. The first session included plies 1-40; the second session included plies 41-80. A 10 minute break was allowed between sessions to prevent participants from experiencing any discomfort resulting from the headband of the eye tracker. Participants were instructed to watch the game while using the joystick as in the previous game. Fixations and saccades were recorded. To qualify as a fixation, the eye had to remain looking within .5(visual angle for at least 100 msec.

Unfortunately, in all of our analyses, differences among skill levels were quite small. For example, the largest difference in the fixation rate was 0.4 fix/sec, and the difference in average saccade distance was 1(. A more promising finding may be that the novice player had longer fixation durations (M = 283 msec) than did the expert (M = 233 msec) or intermediate player (M = 198 msec), but even here it was difficult to clearly classify the players. We looked at a myriad of other measures (e.g., fixation duration on critical pieces, area of the board covered) with little evidence that eye movements would be a consistent predictor of expertise. It certainly would not be an easy one to ascertain. Unfortunately, a clear picture of SA, as reflected in eye movement differences, was not apparent, although some interesting trends did emerge. Obviously, this is not to say that such differences do not exist, merely that we could not find them.

Situation-present Queries

During the third game, the participant responded to questions about current and future chess positions. On some trials during the chess game, a tone sounded and a question was presented auditorily while the chess board and pieces remained in view. All participants were asked the same questions on the same plies. Eighteen questions were asked, six from each of three categories: 1) Perceptual, 2) Present Conceptual, 3) Future Conceptual. In this version of the on-line query methodology, the situation remains present while the participant responds. Clearly, the proportion of correct responses should be quite high, given that the participant can determine the correct answer from information still being displayed. Thus, the primary dependent variable was response time.

We looked at this variation of more traditional on-line querying techniques for a number of reasons. First, unlike looking at mistakes in SA, response time allows us to investigate successful SA, rather than inferring characteristics of SA from its failures. Second, our ultimate interest is in exploring SA among air traffic controllers, where any technique that removes them from the radar screen is likely to be disruptive and viewed with suspicion. An auditory query allows the situation to continue and uses an input mode that can be fit into the controller's existing work scheme (e.g., by querying over a telephone line).

The participant answered orally while the experimenter recorded his response. The screen in front of the participant went blank and the response recorded by the experimenter was then displayed, allowing the participant to confirm or change the experimenter's input. A question was asked during 18 randomly selected plies during the game. Participants were queried about: 1) Perceptual characteristics(Where is the white queen?; What piece is adjacent to the black rook?; 2) Present conceptual relations (What piece is the white bishop attacking?; What piece is defending the white knight?; 3) Future relations(What piece can white move to pin black's rook?; What piece can black move to prevent a back rank mate?

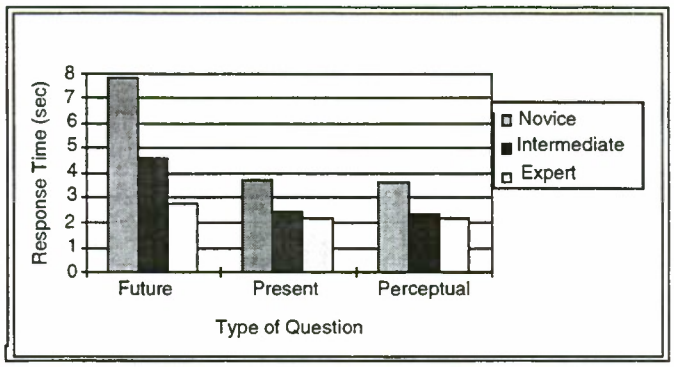


Figure 3. Situation-present queries

As expected, accuracy was quite high and did not distinguish among levels of expertise. Response time, on the other hand, appears to be a good index of expertise, and presumably, SA, $F(2, 24) = 13.12$. Results are shown in Figure 3. Overall, the expert responded faster than the intermediate, who responded faster than the novice. Latency also varied as a function of question type, $F(2, 12) = 4.78$, with future questions taking longer to answer than did the other types. Skill also interacted with question type, $F(4, 24) = 2.32$, $p <$. 10. Differences as a function of skill were present most clearly in questions about future events, $F(2, 8) = 7.20$ and perceptual events, $F(2, 10) = 5.28$.

Situation-absent Queries

More typically, on-line queries freeze the simulation of interest, remove information, and ask the participant a question or series of questions (i.e., Endsley's (1988) SAGAT). In the fourth game, the pieces were removed from the board as a tone sounded, and a question was presented visually to the right of the now empty board. Participants read and answered the question, the screen went blank, the experimenter typed the response to allow the participant to verify the entry, and then the board reappeared as it was when the question sequence was initiated. As before, 18 questions were asked.

The data from situation-absent procedures are the number of questions responded to correctly. Those data appear in Figure 4. It is worth noting that the figure presents results similar to those found in the verbal protocol: an expert advantage for predictions (future) and an intermediate advantage for assessments of the (current) position. This methodology showed a marginal skill effect, $F(2, 30) = 2.94$, $p < .10$. Questions about the future distinguished most clearly among the three skill levels, $F(2, 10) = 3.18$, $p < .10$. Skill differences were not reliable for the perceptual or present questions, despite the graph's suggestion of an intermediate superiority for present-queries.

Given the success we had with the response time analysis in the situation-present procedure, we looked at the response times associated with the correct responses in the situation-absent procedure. Unlike percent correct, visual inspection of the response time data revealed consistent differences among skill levels (although not as clear as the situation-present condition). Unfortunately, the error rate prevented any meaningful analysis of correct latency. Nevertheless, this may suggest that the differences between the situation-present and situation-absent procedures

may be relatively unimportant, provided that response time is used to assess differences among levels of expertise.

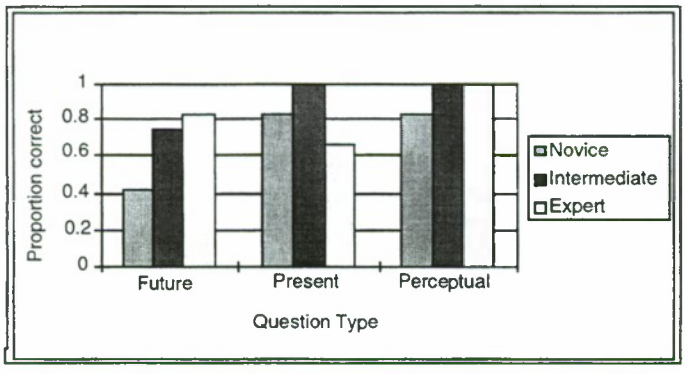


Figure 4. Situation-absent queries

Memory

The final analysis involved consideration of the recall protocols participants gave after the game. Although recall occurred after each game, we considered only the first game here, since it would be uncontaminated by other recalls. The participants differed dramatically in the length of their recall protocols, with the expert saying very little and the intermediate saying quite a bit. However, the expert's recall consisted primarily of general abstractions that characterized large segments of the game. The novice's recall consisted of several move by move recollections of the game. The intermediate's recall fell between these two. Unfortunately, the succinctness of the expert's recall made it difficult to analyze the protocols beyond this overall classification. However, it does suggest that experts tend to convey labels of encapsulated information (Schmidt & Boshuizen, 1993).

Discussion

We investigated five procedures. Of those, eye movements seemed to be the most complicated and yielded the fewest insights. Analysis of memory protocols and on-line protocols were also problematic. The two query procedures seemed to supply some useful information. Across the procedures, we found considerable evidence that questions about the future are most likely to discriminate among all three levels of expertise. Number of predictions in the verbal protocols, proportion correct for future-oriented queries (situation-absent), and response time for future-oriented queries (situation-present) all suggest that researchers interested in distinguishing levels of SA would do well to focus their efforts on future events. Even our cover task, which required participants to predict material loss, showed clear expertise effects. Information about the current state of affairs, as measured by assessment utterances in verbal protocols and proportion correct in

situation-absent queries, did not discriminate across skill levels. This was due to the finding of an intermediate effect where the intermediate chess player outperformed both the expert and the novice. The problem in assessing SA with these types of information is that the novice and the expert are more similar than their ranking suggest they should be.

The current work also suggests that the best procedure for measuring SA is to ask questions of the participant, but to rely on the response time for correct responses, rather than on the proportion of correct responses. In the situation-present procedure, response time (especially for future events) was a good discriminator of skill level. Even in the situation-absent procedure, response time may prove to be a fairly good discriminator. The query procedures also did not seem to fetter the participants' ability to monitor the task, as evidenced by their continued ability to project material loss throughout the procedure. This suggests that interrupting participants with a query does not seriously disrupt the task. Further, there seems to be no reason to remove the situation in order to assess SA. Of course, proportion correct will be near ceiling since the participant can simply reexamine the display; however, response time will capture SA differences. In fact, it will capture it more clearly than removing the situation. We believe this finding to be quite encouraging for efforts, such as ours, that hope to be applicable to air traffic controllers. We believe, in such an environment, an important part of SA is knowing where to find the information when you need it; in air traffic control, not all information needs to be memorized, nor should it be. With the situation present, the controller can continue to control traffic while SA queries from other facilities are occasionally presented over one of the controller's telephone lines. Attempts to extrapolate these results to other arenas should take into consideration the small number of participants used in this study, the differences between task domains, and the differences due to monitoring versus active control. Nevertheless, the current work suggests that the best way to assess SA is to ask questions about impending events while the situation remains available to the operator and to time how long it takes the operator to make the correct response.

References

Charness, N. (1989). Expertise in chess and bridge. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 183-208). Hillsdale, NJ: Erlbaum.

de Groot, A. (1965). *Thought and choice in chess.* Paris, France: Mouton.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, (pp. 97-101). Santa Monica, CA: Human Factors Society.

Ericsson, K., & Simon, H. (1984). *Protocol analysis - Verbal reports as data.* Cambridge, MA: MIT Press.

Grant, J., & Marsden, P. (1988). Primary knowledge, medical education and consultant expertise. *Medical Education, 22,* 746-753.

Kibbe, M. P. (1988). Information transfer from intelligent EW displays. In *Proceedings of the Human Factors Society 32nd Annual Meeting,* (pp. 107-110). Santa Monica, CA: Human Factors Society.

Mogford, R. H. (1994). Mental models and situation awareness in air traffic control. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems,* (pp. 199-207). Embry-Riddle Aeronautical Press.

Moray, N. & Rotenberg, I. (1989). Fault management in process control: Eye movements and action. Special Issue: Current methods in cognitive ergonomics. *Ergonomics, 32,* 1319-1342.

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

Ohnemus, K., & Biers, D. (1993). Retrospective versus concurrent thinking-out-loud in usability testing. In *Proceedings of the 37th Annual Meeting of the Human Factors Society*, (pp. 1127-1131). Santa Monica, CA: Human Factors and Ergonomics Society.

Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology, 1,* 45-57.

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On the origin of intermediate effects in clinical case recall. *Memory & Cognition, 21,* 338-351.

Stein, E. S. (1989). *Air traffic controller visual search* (DOT/FAA/CT-TN89/9). DOT/FAA Technical Center: Atlantic City International Airport, New Jersey.

Sullivan, C., & Blackman, H. S. (1991). Insights into pilot situation awareness using verbal protocol analysis. In *Proceedings of the Human Factors Society 35th Annual Meeting,* (pp. 57-61). Santa Monica, CA: Human Factors Society.

Wierwille, W., & Eggemeier, F. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors, 35,* 263-282.

This research was supported by contract # DTFA-02-94-T-80261 from the Federal Aviation Administration to Francis T. Durso. Address correspondence to Frank Durso, Department of Psychology, University of Oklahoma, Norman, OK 73019-0535 or e-mail: fdurso@uoknor.edu.

Perspectives on the Appreciation of Team Situational Awareness

Iain S MacLeod[1], Robert M Taylor[2] and Colin L Davies[1]

[1] Aerosystems International
[2] DRA CHS

A Metric on Applied Skills?

"Situation awareness is related to the managerial skill dimension of perception."
(Prince, Chidester, Bowers, & Canon-Bowers, 1992).

General

To be aware a person must possess a pertinent knowledge and appreciation of their immediate environment. To be situationally aware a person must also understand the implications of the perceived environment, with regard to their current and future status in that environment, and be able to selectively and dynamically focus their perception and skills in support of safe and effective goal attainment.

In the UK Royal Air Force, some SA environmental basics were taught under the banner of 'Airmanship' and 'Rules of the Air'. In flying terms, SA is also closely related to the skills associated with aircrew mission planning, system tactical direction and control, and team work. Elsewhere, SA has been known under many 'slang' names depending on the era and societal group, for example, canny, nonce, street-wise, bad.

Perspectives

Perspectives suggests different views on the same thing. This paper will consider perspectives related to the continuous subjective assessment of SA through observations of work performance and, therefore, is primarily concerned with the concurrent evaluation of operator and team performance. Subjective rating techniques of SA (Endsley (1988), Taylor (1990)) will not be discussed in detail. Further, whilst it is accepted that subjective rating assessment of SA is an important approach to the understanding of SA, it is basically a retrospective and global consideration of past work aspects. It is suggested that rating scale application during the normal course of a subject's work could disrupt their primary task performance.

Individual SA

Situational awareness (SA) is associated more with complex 'open skills' than with 'closed skills' (Poulton, 1957), in that its maintenance requires a good deal of appreciation of, and interaction with, external events. SA is a necessary and intrinsic property of applied skills. It is also a

product of the quality and experience of skill application, but is variable under the influences of certain personal and organisational factors.

The SA capabilities of each individual is different, partly through variations in their levels of basic ability. In some cases the ability levels of the individual will bar improvement in their skills or SA. See Phillips, 1995 as an illustration of the seriousness of such limitations. However, the SA level of the individual can usually be raised through the teaching of associated skills, and SA should improve in parallel to work practice and a gain of relevant experience of work environments.

Team SA

Team SA involves a joint management and sharing of information, and the collective projection of that information into a future context. The SA possessed by a team not only relies on the knowledge and skills of the individual team members, and their conception of the perceived working environment, it involves a team consensus on the understanding of that environment.

This shared consensus implies a within team continual awareness of each others' proficiencies and limitations plus a collective appreciation of the standard of current team performance. Considering future Human-Machine System (HMS) teams, it is suggested that good synergy between human team and machine involves the establishment of extended joint cognitive man-machine systems where physical interfaces are not obvious barriers to smooth and proficient system performance.

Chosen Perspectives

This paper will present an appreciation of approaches to the observational assessment of team SA assessment from three perspectives, namely:

1. the perspective of expert assessment;
2. the perspective of the utility of the 'Teamwork Model' to direct observations on the contributions of different aspects of teamwork to teamwork performance (Taylor & Selcon, 1992);
3. from perspectives suggested through the application of the 'Homeostat' tool, a tool for the determination of team cohesiveness or 'gomphotics' [1]. (Novikov, Bystritskaya, Eskov, Vasilyiev, Vinokhodova & Davies, 1993).

The three perspectives will be discussed and their relative usefulness to the assessment of team SA will be assessed, considering the present and the future.

Perspective of Expert Assessment

Expert assessment of aircrew normally serves one or several of the following functions:

1) certification;
2) standardisation;
3) assessment;
4) training.

---

1 *Gomphotic. Meaning strongly connected - from the Greek 'gomphos', a bolt or nail.*

In the aviation world, the expert does not assess SA as such but generally appreciates SA as a quality product of the suite of various skills applied by aircrew during their work.

Further, the conduct of a flight can be separated into phases that include take-off, transit and descent to a destination. Assessment of SA is made with relation to the conduct of the various phases of the flight, in connection with any particular notable occurrences during that flight, and under the umbrella of the overall assessment of crew performance. Usually, the assessments of SA are couched in terms of qualitative assessments of aircrew performance with relation to the employment of the aircraft, their use of aircraft equipments, and their appreciation and application of evidence available from the aircraft's environment.

There are many influences on aircrew performance; the individual's and the team's moods & motivations, and organisational influences to name but a few. All influences can effect the flight and mission conduct. In the authors' experience, the crew performance on a flight can often be accurately predicted from the observed performance of early flight stages, for example from the quality of the conduct of mission planning.

A lazy start makes you late for the rest of the day (old Scottish Proverb).

However, early assessment must not create assessor complacency, but should give indications of the set of primary performance cues that should be looked for throughout the remainder of the flight or flight simulation exercise.

Large crew military aircraft are normally employed in air / ground surveillance or long range operations such as maritime patrol. With large crew aircraft, such as the Nimrod MR and AWACS, there is a wealth of information on which to assess crew performance. However, to prevent the assessor being overloaded with too much information, appraisals of large crews must be focused on an assessor selection of primary cues.

The assessor of such crews must possess several skills including an acknowledge high skill in the area of the assessment, accepted standards on which to base the assessment, and a trusted fairness in their assessment of others. Military assessment and standardisation of aircrew is performed on the professional level by trade, and at the crew level considering levels of crew teamwork and overall performance.

McMillan et al (1995) , reporting the results of a major USAF study, conclude that SA evaluations that involve complex tactical scenarios should include structured expert observations as part of the measurement process. Objective measures alone were considered to be not yet up to the task.

The Crew Proficiency at Routine Tasks

The assessment of crew proficiency at routine tasks can be based on many forms of cues and crew activity areas. The few areas that will be discussed are:

- Intercom Discipline;
- Cooperation;
- Advice and Aiding.

Intercom Discipline

An aircraft intercom is a voice communication net that allows all its participants full conferencing facilities. The aircraft types under question are normally fitted with several intercom nets to prevent individual intercom net overload. Moreover, these intercoms are usually divided between a crew intercom net and tactical intercom nets, these latter nets being utilised by operators of the specialised mission equipments of the aircraft or by operators residing in one aircraft area such as the cockpit. All nets can be overridden in an emergency.

The skill in using these nets is highly dependant on crew SA. Crew SA is manifest in the appropriateness and form of the traffic used in the nets, appropriateness being associated with the

team awareness of mission context and the aircraft situation in the light of available evidence. Thus, idle chat or excessive interruptions of other users is usually a sign of poor crew discipline or boredom.

The expert assessor will use an assessment of intercom discipline as one of many methods of assessing crew performance and SA. Generally, the less disciplined the use of intercom the poorer the level of teamwork and team SA. Further, the pertinence of reporting is supported by crew SA in that good reporting indicates an awareness of others' needs in a team or the relative importance of a report or new information to the team and aircraft system performance.

Cooperation

If one or more team members decide not to cooperate with the rest of the team, the effects are obvious to the expert observer. The lack of information or support from the non cooperating members can seriously disrupt the flow of information between the crew members and mar the cohesion of the team. With willing cooperation the reverse is true in that cooperation becomes an aim of the crew, this accompanied by a motivation to improve their team work. Such good cooperation is usually accompanied by a good intemal team SA and genuine efforts to improve the system performance within the appreciated individual and collective environment.

Advice and Aiding

Advice and aiding are associated with team cooperation but are useful indicators of individual aspirations, not only to cooperate as a team, but also to improve the team cohesion and cooperation by proffering useful advice and giving aid to the less able of the team. The activity in this area indicates an important characteristic needed for the support of team work and SA. This characteristic is concern for the other members of the team, a willingness to assist them through difficulties, and is an important prerequisite for trust within a team.

Forms of Expert Assessment

The expert assessment is in the form of expert opinion and must be open to personal bias. However, it can be conducted in parallel to crew performance and is the only method that can allow non obtrusive and concurrent assessment of crew performance (team and individual), crew teamwork and crew SA. Further, expert assessment allows rapid feedback to the crew in debrief after a standardisation or assessment check. Expert assessment is the method normally used for the regular appraisal of the overall quality of operational crew performance.

The expert, whether assessing aircrew teams in the air or in the simulator, is not concerned with the definition of teamwork, proactive or reactive planning, or SA. He is concemed with the performance of the whole system, in a potentially hostile environment, and the effects that sub optimal performance of parts of the system have on the whole. Nevertheless, under some form of convenient classification, he is aware of SA and is continually assessing the quality of SA possessed by the crew, and the influence it has on their performance.

The coverage of areas of crew activity has of necessity been sparse. Other common areas considered include: the use of initiative, anticipation of events, availability of alternative plans, general updating and briefing within the crew, the use of awareness games (e.g. weather assessment by crew members other than the pilot).

The main drawback of expert assessment is that it has to be performed on known systems that fit within the skill and experience remit of the assessor. Therefore, some form of assessment should be devised for the consideration and assessment of the skills and SA required for future systems, especially systems where the machine has elements of SA and becomes a team member with the human. One such approach is illustrated by the Teamwork Model.

The Perspective of Observations using the Teamwork Model

The Teamwork Model considers that a team had three distinctive characteristics, namely:

1. Coordination of team activities based on trust;
2. A well defined organisation and structure with specific assigned individual roles;
3. Team processes that depend on communication and interactions between team members.

The introduction of team concepts provides a broader framework for thinking about human-machine cooperation. Consideration of the machine as a teaming resource raises a number of issues. Foremost among these must be considerations of trust between team members, functionality of team members, communications within the team, and where authority should be vested within the team. (Taylor and Selcon, 1992).

It is interesting to compare the three 'Teamwork' characteristics with the team characteristics previously discussed previously under expert observation and the seven characteristics of a team as given by Alluisi, 1992. The Teamwork Model considers that four basic characteristics are foundation teamwork components. These characteristics are given as team goals, resources, team architecture and inter-team processes. The model is illustrated in Figure 1.

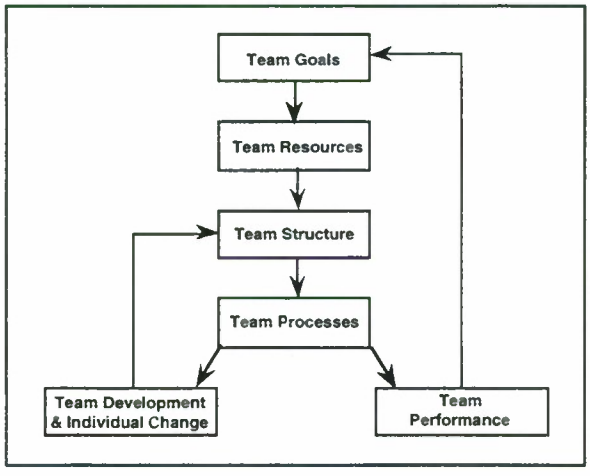


Figure 1. Teamwork Model (Taylor and Selcon, 1992).

A teamwork audit model was constructed to consist of twenty teamwork constructs, elicited from the literature on human-electronic crew teamwork, and linked to the four foundation characteristics of the model. Taylor et al (1995) report ratings on the teamwork model dimensions for levels of 'electronic crewmember' automation aiding against benchmarks for good and poor human teamwork performance.

The model contains ideas common to most team models but is expressed in language that might require translation by even an expert assessor. For example, the assessor might have difficulty considering a human-machine team in terms such as 'Wide Bandwidth'. Indeed the authors

recognised this problem in that they observed that certain model dimensions seemed difficult for aircrew to grasp because the descriptions used theoretical constructs and unfamiliar words.

Nevertheless, this model illustrates the need to develop models as an aid to the appreciating and explanation of current and future HMS teamwork. In addition, it also indicates a requirement to overcome the problems involved in the practical development of such a model, if the model is to be used for the guidance of future 'Expert Assessors' of crew performance and SA. In contrast to the Teamwork Model, the Homeostat Tool suggests an alternative approach for the future assessment of HMS teamwork and SA.

The Perspective of the Homeostat Tool

Control theory has long been capable of specifying the dynamic control changes required to direct the necessary machine processes to adapt to changes caused by environmental influences or inputs to the machine processes. The seminal work by Ashby (1956) on cybernetics introduced the idea of homeostasis in man machine systems; adaptive control to maintain a stable and required man-machine system state. In addition, during the 1960s, the Russian Institute of Biomedical Problems investigated the problem of conflicts within small teams over periods of long confinement in space or polar stations. They found that if people in the small group were psychologically compatible, then they would work effectively and happily together, and show few conflicts, even over periods of long confinement.

Arising from the above research, the 'Homeostat' Tool was created and used by Novikov et al (1993) to assess team cohesiveness through the tool based assessment of the quality of adaptive interaction and functioning in a team. Through use of the tool, the interactions between a group of people were assessed, though not the processes behind their interactions. Moreover, it was argued that inferences on human to human interaction processes could be made through the quality of the team members' performance with the tool.

Homeostat is described by Novikov as a biotechnical system. The subject team was required to stop normal work periodically and work with the Homeostat Tool. Each team member had a joystick control and a control dial with two needles. Needle One was under individual control and had to be manipulated through use of the control to place the Needle Two in the zero position. The interrelationship between the control of the Needle One, and the adjustment of the Needle Two to zero, was hidden within the tool. The inputs of each team member influenced the tool's control equations. The actual inter relationship was through a set number of linear equations, and an associated number of unknown quantities, these equal to the number of members of the team.

It was shown that the task difficulty level depended on the strength of the interrelation between the tool activities of the various team members. When the difficulty reached a certain level, the task became unmanageable for a team acting on an individual basis and without attention to their partners. At this stage, to regain control, it was necessary for one of the subjects to change strategy and try and control the whole process. Experiments were conducted to test various hypotheses connected with the examination of team cohesion and team leadership.

One of the results of the experiments was a proposed classification for the level of small team development. The highest was termed the gomphotic group or team, the team where the psychological and physical compatibility was excellent, where the team psychological structure has flexibility, and where the team members were prepared for their work, had common aims, tastes and adopted skilled methods. Figure 2 below gives an illustration of the classification of small team development. The lower classification teams will not be defined as their properties are reasonably obvious and, outside their existence as indicators of levels of team development, the continued discussion of teams has little further utility in the arguments of this article.

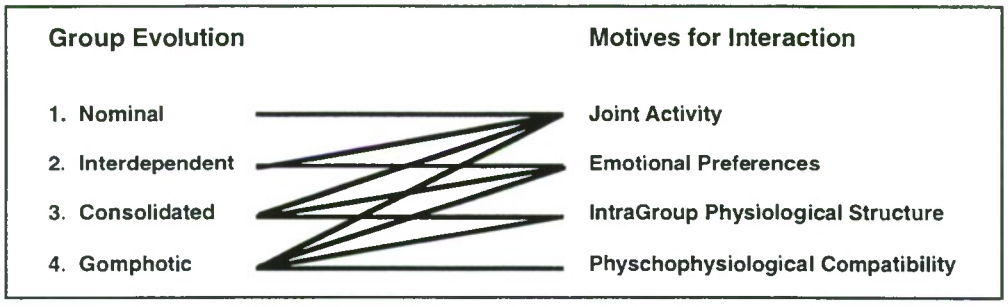


Figure 2. Classification of Small Groups and Teams

The point of particular interest to this article is that the simple Homeostat Tool gave indications to all team members not only of the cohesion of the team but also of the state of the human-machine system that existed as the tool / team system. It is not the purpose of this article to assess the validity of the Homeostat experiment. However, considering the principles of the tool, it may be possible in the future to have a Team-Machine System (TMS) that assesses the quality of the interaction of all the human-machine system components and adaptively aids individual components to maintain an equilibrium and desired quality of system performance. This area is fraught with difficulties and should be approached through an appreciation of the dynamics of the total TMS. Further, the concept of an adaptive controller must cover a TMS interaction policy as well as an HMS control policy (Hollnagel, 1995).

In association with the above would be a need to keep all team members aware of the performance of the system as a whole, of the quality of their own contribution, and that contribution of others. Such a machine supported feedback would aid the maintenance of the team SA. Such avenues have also been suggested in work on the measurement of cognitive compatibility (Taylor, 1995), where it is indicated that future systems will need to consider dimensions of social interactions (e.g. goals, shared functions, aiding, advice) as well as the traditional human factors considerations of compatibility (i.e. modality, movement, spatial, conceptual compatibility) to achieve cognitive quality in joint cognitive systems.

Comments on Perspectives

Three perspectives on SA were considered, namely: i) the usual present approach to the assessment of crew skills and SA through the utilisation of an expert assessor; ii) the development of the Teamwork Model, and its associated constructs, as an aide memoir to the assessment of HMS teamwork and embedded SA; iii) the Homeostat Tool and the suggestions from its use towards the development of machine based assessment of teamwork and system based SA assessment.

The expert assessment approach will always have utility but, in the future, may have difficulty in accurately assessing advanced human-machine system states and SA. The Teamwork Model shows an early attempt to scope the measurement of the advanced human-machine system. Lastly, the Homeostat Tool gives possible indicators to a dynamic SA and team performance assessment subsystem of the future.

The concept of SA is like the concept of Leadership, all believe they exist, but by their very nature both concepts will thwart any secure definition. Whilst SA is a necessary and intrinsic part

of skill application, it should also be considered as a product of that application and not as a separate process.

Team SA relies heavily on the maintenance of common goals. This maintenance can only be achieved by a team through agreed endeavours, a willingness to co-operate, and a consideration of the needs of fellow team members. Thus SA supports the maintenance of trust. Trust is essential in sustainable teamwork. Trust is promoted by awareness of the predicaments of both the team as a whole, and of the individual members, as well as awareness of closure on commonly shared goals, through agreed means of achievement.

References

Alluisi, E.A. (1992) in the Forward to *Teams: Their training and performance*,, Swezey, R. & Salas, E (Eds) Ablex, Norwood, NJ.

Ashby, W.R., (1956), *An introduction to cybernetics*, Methuen & Co., London

Endsley, M.R. (1988), Situation awareness global assessment technique (SAGAT), in *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference - NAECON 1988* (Volume 3, pps 789-795.

Hollnagel, E. (1995) The Art of Efficient Man-Machine Interaction: Improving the coupling between man and machine, in Hoc, J-H, Cacciabue, P.C. and Hollnagel, E (Eds), *Expertise and Technology: Cognition and Human-Computer Cooperation*, Lawrence Erbaum, New Jersey, pps 229-241.

McMillan, G.R., Bushman J., and Judge C.L.A. (1995). Evaluating Pilot Situational Awareness in an Operational Environment, in *Situation Awareness: Limitations and Enhancement in the Aviation Environment*, 79th AGARD AMP Symposium. AGARD Conference Proceedings, AGARD, Neuilly-sur-Seine. April 1995 (In press)

Novikov, M.A., Bystritskaya, A.F., Eskov, K.N., Vasilyiev, V.K., Vinokhodova, A.G. & Davies, C. (1993) 'HOMEOSTAT -A Bioengineering System' in *Proceedings of 23rd International Conference on Environmental Systems*, Colorado Springs, Colorado, July 12-15, 1993, published by SAE Technical Paper Series, Warrendale, PA.

Phillips, E.H., (1995), Eagle Crash Probe Targets Crew, *Aviation Week and Space Technology*, April 24, pp30.

Poulton, E.C. (1957) ' On the stimulus and response in pursuit tracking', *Journal of Experimental Psychology*, 53, pps 57-65.

Prince, C., Chichester, T.R., Bowers, C. & Cannon-Bowers, T (1992) ' Aircrew Co-ordination-Achieving Teamwork in the Cockpit' in Swezey, R. & Salas, E (Eds) *Teams: Their training and performance*, Ablex, Norwood, NJ.

Taylor, R.M. & Selcon, S.J. (1992) *A Teamwork Model of Pilot Aiding: Psychological Principles for Mission Management System Design*, AGARD-CP-504, March 1992, Paper 9.

Taylor, R.M. (1990), Situational awareness rating technique (SART): The development of a tool for aircrew systems design, in *AGARD-CP-478 - Situational Awareness in Aerospace Operations*, pps 3-1 to 3-17, Neuilly Sur Seine, France: Advisory Group for Aerospace Research and Development. (AD-A223939).

Taylor, R.M. (1995) *Experiential Measures: Performance-Based Self Ratings of Situational Awareness* in Conference on Experimental Analysis and Measurement of Situational Awareness, Daytona Beach, 1-3 November. Embry-Riddle Press, Florida (in press).

Taylor, R.M., Shadrake. R. and Haugh. J., (1995), Trust and Adaptation Failure: An Experimental Study of Unco-operation Awareness, in The Human-Electronic Crew: Can we trust the team?, *Proceedings of the 3rd International Workshop on Human-Computer Teamwork*, Cambridge, UK, 27-30 September.

A Methodology for Analyzing Team Situation Awareness in Aviation Maintenance

Michelle M. Robertson[1] and Mica R. Endsley[2]

[1] University of Southern Califomia
[2] Texas Tech University

To assess team situation awareness in an aviation maintenance setting, a methodology was developed for examining situation awareness requirements that incorporates both individual and team situation awareness perspectives. In the present study, inquires were conducted in the field maintenance setting at a major airline. Contextual inquires were combined with a goal directed task analysis to specify the situation awareness requirements involved in each of the interactions (between and within teams) required to perform maintenance tasks. The use of this methodology is discussed along with its application for analyzing team SA in this setting. The methodology is unique in providing a useful technique for examining situation awareness requirements at the team and organizational level as opposed to a purely individual level which has been the hallmark of previous research. .

Introduction

Insufficient attention has been paid to problems involved in aircraft maintenance. While the number of incidents due to mechanical failures that can be traced to maintenance problems are relatively few when compared to other causal factors (e.g. inflight human error), they do exist and can be systematically addressed. Marx and Graeber (1994), for instance, report that 12% of accidents are due to maintenance and inspection faults, and around one-third of all malfunctions can be attributed to maintenance deficiencies. In addition to its impact on safety of flight, the efficiency of maintenance activities can also be linked to flight delays, ground damage and other factors that directly impact airline costs and business viability.

In examining problems that occur within the maintenance arena, several types of difficulties can be identified.

1) The first involves shortcomings in the detection of critical cues regarding the state of the aircraft or sub-system. Several accidents have been traced to metal fatigue or loose and missing bolts that should have been visible to maintenance crews. Incidents exist of aircraft being returned to service with missing parts or incomplete repairs. Frequent errors include loose objects left in aircraft, fuel and oil caps missing or loose, panels and other parts not secured, and pins not removed (Marx & Graeber, 1994). While several factors may contribute to this type of error, in all of these cases the state of the system was not detected prior to returning the aircraft to service.

2) Often, even when important information is perceived, there may be difficulties in properly interpreting the meaning or significance of that information. For instance, Ruffner (1990) found that in more than 60% of cases, the incorrect avionics system is replaced in an aircraft. While the symptoms may be observed correctly, a significant task remains in properly diagnosing the true cause of the failure. While not much data exists regarding the impact of misdiagnoses of this type,

there is a significant increase in the probability of an incident occurring when the aircraft undertakes the next flight with the faulty system still aboard.

3) These problems are compounded by the fact that many different individuals may be involved in working on the same aircraft. In this situation, it is very easy for information and tasks to fall through the cracks. The presence of multiple individuals hightens the need for a clear understanding of responsibilities and communications between individuals to support the requirements of individuals in performing those tasks. In addition to the need for intra-team coordination, a significant task befalling maintenance crews is the coordination of tasks and information across teams, to those on different shifts or in different geographical locations. The Eastern Airlines incident at Miami Airport (National Transportation Safety Board, 1984) has been directly linked to a problem with coordination of information across shifts (along with other contributing factors). In addition, considerable energy is often directed at coordination across sites to accommodate maintenance tasks within flight schedule and parts availability constraints. These factors add a level of complexity to the problem that increases the probability of tasks not being completed, or completed properly, important information not being communicated, and problems going undetected as responsibility for tasks becomes diluted.

Situation Awareness

All of these difficulties point to a problem of situation awareness. That is, maintenance crews need additional support/training in ascertaining the current state of the aircraft system (supplementing current technical training programs). Situation awareness has been found to be important in a wide variety of systems operations, including piloting, air traffic control and maintenance operations. Formally defined, *"situation awareness is the detection of the elements in the environment within a volume of space and time, the comprehension of their meaning, and the projection of their status in the near future"* (Endsley, 1988). In the context of aircraft maintenance, this means being aware of the state of the aircraft system (and the sub-system one is working on). Termed Level 1 SA, this would include perception of the state of the factors listed in item number one above. Level 2 SA would involve the technicians' understanding or comprehension of the significance of observed system states. Specifically this would include their diagnosis of the causal factors associated with observed symptoms.

While SA has generally been discussed in terms of the operation of a dynamic system, such as an aircraft, the concept is also applicable to the maintenance domain. The complexity of aircraft systems and the distributed nature of equipment and system components posses a significant challenge to the technicians' ability to determine the state of the system (Level 1 SA) during diagnosis and repair activities. Putting together observed cues to form a proper understanding of the underlying nature of malfunctions (Level 2 SA) is a significant problem in diagnostic activities. Level 3 SA, the ability to project the state of the system in the near future, is considered the highest level of SA in dynamic systems. In the maintenance domain, technicians may need to be able to project what will happen to an aircraft's performance with (or without) certain actions being taken or with given equipment modifications/repairs/adjustments occurring. This task may be even more difficult for maintenance technicians, as they often receive little or no feedback on the effects of their actions, and thus may have difficulty developing an adequate mental model for making accurate predictions. The ability to project system status forward (to determine possible future occurrences) may also be highly related to the ability to project system status backward, to determine what events may have led to an observed system state. This ability is particularly critical to effective diagnostic behavior.

Team SA

In aircraft maintenance, as in many other domains, the requirement for situation awareness becomes compounded by the presence of multiple team members, and multiple teams, as

individuals need to not only understand the status of the system they are working on, but also what other individuals or teams are (and are not) doing as well, as both factors contribute to their ultimate decision making and performance. Team situation awareness can be defined as "the degree to which every team member possesses the situation awareness required for his or her responsibilities" (Endsley, 1989). In this context, the weak link in the chain occurs when the person who needs a given piece of information (per his or her job requirements) does not have it. The level of SA accross the team, therefore, becomes an issue of some concern. The objective of the current study was to identify situation awareness requirements for aircraft maintenance teams, analyze how SA needs are currently being met in a typical maintenance environment, and establish concepts and requirements for training Team SA in this domain.

Methodology

A Team SA Context Analysis methodology was developed for this project. This method consists of two parts: An SA Requirements Analysis and an SA Resource Analysis, as shown in Figure 1.

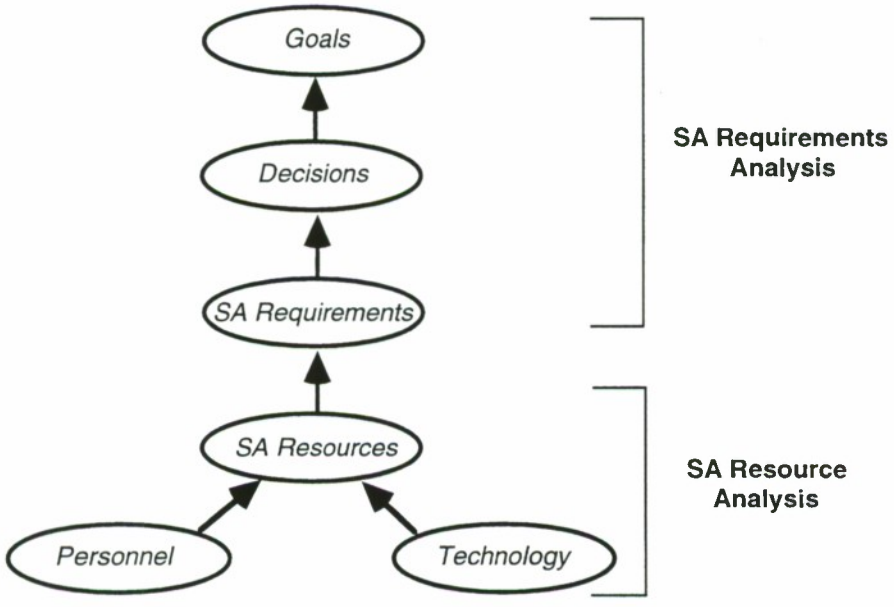


Figure 1. Team SA Context Analysis

SA Requirements Analysis

The first step in addressing situation awareness was to determine the specific situation awareness requirements of individuals in the aircraft maintenance arena. This was addressed through a goal-

directed task analysis which assessed: 1) the goals and sub-goals associated with maintenance crews, 2) the decision requirements associated with these goals, and 3) the situation awareness requirements necessary for addressing the decisions at all three SA levels - detection, comprehension, and projection. This type of analysis has been successfully conducted for several classes of aircraft (Endsley, 1989; Endsley, 1993), air traffic control (Endsley & Rodgers, 1994) and airway facilities maintenance Endsley, 1994 #524.

Analyses were conducted through expert elicitation with experienced maintenance personnel, observation of aircraft maintenance activities, and review of all available maintenance documentation. The analysis concentrated on the B-Check maintenance activities conducted by a major airline company at a major airport. To date interviews have been conducted with 3 maintenance supervisors, 4 lead technicians and 4 A&P technicians at the site.

SA Resource Analysis

The second part of the Team SA Context Analysis concentrated on identifying the SA Resources used in the current environment to achieve the SA Requirements. Two major categories of resources were considered: Other technical operations personnel as a source of information and the technologies used as sources of information.

To provide an assessment of the personnel SA resources in the aviation maintenance setting, an analysis of communications between organizations and individuals was conducted using a contextual inquiry approach. The contextual inquiry approach (Holtzblatt & Jones, 1993; Robertson & O'Neill, 1994) focused on understanding and describing the communication patterns within and between teams as related to their performance goals. The contextual inquiries were conducted simultaneously with the interviews for determining the SA requirements. The contextual inquires involved semi-structured interviews in which each individual was asked to describe his/her major job functions and goals and the organizations, departments or individuals that served as resources in meeting those goals. A context mapping was then determined showing the communication patterns among and between team members. Each individual was asked to make an estimate of the overall frequency of communication with each identified unit or department and the importance of the communication for achieving their goals. Finally they were asked to identify system, technology or personnel barriers to effective communication and performance in the work setting.

In addition to identifying the SA requirements of teams working on each maintenance task, the technologies for obtaining each requirement within the current system are documented. Based on this analysis, an assessment can be made of the degree to which the current system supports Team SA and the skills and abilities that are required for achieving good SA within this environment.

Results

Examples of the results of the application of the Team SA Context Analysis methodology in the maintenance domain are presented here. Job goals in the aircraft maintenance domain appear to be oriented towards the dual goals of ensuring aircraft safety and delivering aircraft for service on time. A breakout of A&P technician goals is shown in Table 1. The major decisions that need to be made for achieving each goal were determined during the analysis and the associated SA requirements were delineated. An example of the output of the SA requirements analysis for one sub-goal is shown in Table 2.

The contextual inquiry depicts the personnel SA resources, in terms of the individuals or units within the maintenance technical operations, that are needed to meet the maintenance team's SA requirements. Figure 2 shows the units and individuals that the A&P technician interfaces with.

Lines show communication patterns among and between units. In addition the importance and frequency of each interaction is depicted in Table 3.

Table 1. A&P Technician Goals

1.0 Aircraft safety
- 1.1 Deliver aircraft in airworthy, safe condition
  - 1.1.1 Make repairs
  - 1.1.2 Service aircraft
  - 1.1.3 Find potential problems
  - 1.1.4 Solve problems
  - 1.1.5 Provide quality workmanship
- 1.2 Keep area clean

2.0 Deliver aircraft on time

Table 2. SA Requirements Analysis: Service Aircraft

1.1.2 Service aircraft
- Perform service activities
  - tasks to be done
  - fuel status
  - lubrication system status
  - lavatory status
- *Are we meeting schedule?*
  - time aircraft due at gate
  - delays to aircraft
    - estimated time of arrival at gate
    - aircraft repair status
- *Where do we need to go?*
  - gate assignments
  - permission to taxi
  - permission to do high power run-up
  - taxi/runway clearances
- Job Status
  - status of other tasks impacting own task
  - other tasks own task will impact
  - who can help
  - who needs help
    - tasks started
    - tasks completed
    - tasks/activities being done next
    - who is doing each task
    - activity currently being performed by others
  - major problems encountered

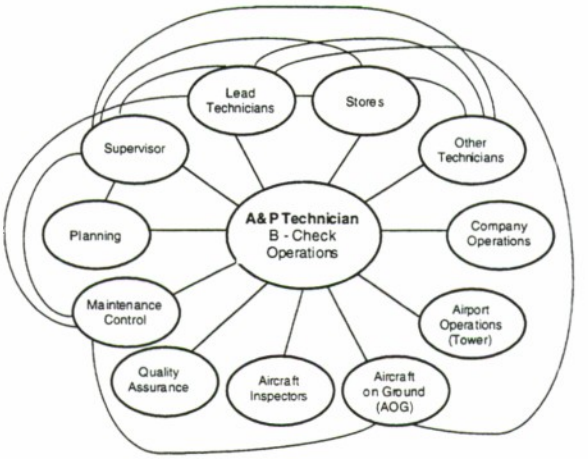


Figure 2. SA Resources: A&P Technicians Communication Patterns

Table 3. SA Personnel Resources: A&P Technicians

SA Resources: Personnel	*Mean Importance*	*Mean Frequency (%)*
Maintenance Control - Maintenance Operations Control - Aircraft on Ground (AOG)	2.5	3.75
Lead Technician	1.5	26.75
Stores	1.5	8.25
Other Technicians	1.0	54.50
Airport Operations	1.0	< 1.00
Company Operations	1.0	< 1.00
Supervisor	3.2	2.25
Quality Assurance	2.5	< 1.00
Aircraft Inspectors	2.0	3.25
Planning	2.0	< 1.00

Discussion

Overall, the applicability of the concept and importance of situation awareness in maintenance teams has been supported by the preliminary data. Teams of technicians are supported by many other personnel and organizational units to achieve their goals, each of which has a major impact on the attainment of maintenance goals. In the maintenance environment it is necessary to examine how information flows between and among team members in order to identify system and personnel factors that will impact on the degree to which team members are able to develop and maintain an accurate picture of an aircraft's status. This knowledge appears to be crucial to the technicians' ability to perform tasks (as each task is interdependent on other tasks being performed by other team members), their ability to make correct assessments (e.g. whether a detected problem should be fixed now or deferred to later (placarded)), and their ability to correctly project

into the future to make good decisions (e.g. time required to perform task, availability of parts, etc...). Further analysis of the other maintenance departments (e.g. material services, planning) is currently being conducted, thus leading to a more robust understanding of the SA requirements and SA resources necessary to effectively perform in this complex, distributed team setting. The methodology presented has been demonstrated to be useful for identifying SA requirements and the resources available in a field setting for meeting those SA needs, and for identifying key issues that impact on people's ability to achieve SA in a team environment.

References

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 97-101). Santa Monica, CA: Human Factors Society.

Endsley, M. R. (1989). *Final report: Situation awareness in an advanced strategic mission* (NOR DOC 89-32). Hawthorne, CA: Northrop Corporation.

Endsley, M. R. (1993). A survey of situation awareness requirements in air-to-air combat fighters. *International Journal of Aviation Psychology, 3(2),* 157-168.

Endsley, M. R., & Rodgers, M. D. (1994). *Situation awareness information requirements for en route air traffic control* (DOT/FAA/AM-94/27). Washington, D.C.: Federal Aviation Administration Office of Aviation Medicine.

Endsley, M. R. (1994). *Situation awareness in FAA Airway Facilities Maintenance Control Centers (MCC): Final Report.* Lubbock, TX: Texas Tech University.

Holtzblatt, K., & Jones, S. (1993). Contextual inquiry: A participatory technique for system design. In D. Schuler & A. Namioka (Eds.), *Participatory design: Principles and practices* (pp. 177-210). Hillsdale, NJ: Lawrence Erlbaum.

Marx, D. A., & Graeber, R. C. (1994). Human error in aircraft maintenance. In N. Johnston, N. McDonald, & R. Fuller (Eds.), *Aviation Psychology in Practice* (pp. 87-104). Aldershot, UK: Avebury.

National Transportation Safety Board (1984). *Aircraft Accidents Report, Eastern Air Lines, Inc., L-1011, Miami, Florida, May 5, 1983.* Washington, DC: Author.

Robertson, M. M., & O'Neill, M. J. (1994). *Development of contextual inquires for examining knowledge workers in technological office environments.* Zeeland, MI: Herman Miller, Inc.

Ruffner, J. W. (1990). *A survey of human factors methodologies and models for improving the maintainability of emerging army aviation systems.* Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Aircraft Recognition Thresholds and Manual Attitude Control: Individual Performance Links that Might be Important to Situation Awareness

Jeremy M. A. Beer[1], Robert A. Gallaway[1], & Fred H. Previc[2]

[1] Systems Research Laboratories, Inc.
[2] Brooks AFB

Abstract

This study tested viewers' recognition of near-threshold aircraft targets, examined the effect of this task on concurrent attitude tracking, and tested whether individuals' performance on each task would predict dual-task performance. An underlying assumption was that efficient multitasking supports SA. Experiment 1 measured duration thresholds for recognition at different nonfoveal locations. Critical tracking ability was assessed also, using a central attitude display. Recognition deteriorated with eccentricity, and a wide distribution of threshold and tracking abilities was found. Experiment 2 combined the tasks, with recognition as the primary task; measures included the dual-task decrease in recognition accuracy and the increase in tracking error. Viewers recognition thresholds (measured in Experiment 1) predicted their dual-task recognition performance but not the increase in tracking error. Notably, critical tracking ability predicted viewers ability to preserve dual-task recognition. Field biases were identified in recognition performance under workload; this is consistent with a spatially biased attention system. Findings are potentially relevant to SA assessment and to display design.

Introduction

Hartman and Secrist (1991) claimed that pilots who can process fleeting or faint (near-threshold) visual images most efficiently will have an advantage when global situation awareness is required. Several questions must be answered, however, before this principle is established. It has not yet been determined how near-threshold recognition of real-world targets, including aircraft, varies throughout the visual field. Neither has it been demonstrated that near-threshold performance predicts the viewer's ability to maintain a global situational percept. This research addressed these questions. An underlying assumption was that to maintain SA, the viewer must perform several perceptual/cognitive tasks effectively at once (Damos, 1978; Endsley & Bolstad, 1994; North & Gopher, 1976). Good performance on two concurrent visual tasks, therefore, would be interpreted as evidence that the viewer possesses skills important for SA. Two experiments tested the distribution of near-threshold performance throughout the visual field, examined the attention drain that near-threshold recognition would place on a simultaneous manual attitude task, and tested whether individuals' performance on each of these tasks would predict their ability to perform both together.

Experiment 1: Individual recognition thresholds and manual tracking performance

Moving stimuli away from the location of gaze is similar to decreasing contrast or viewing duration in its effects, because it reduces available sensory information. In addition, visual performance in the periphery can be subject to directional asymmetries (Previc & Blume, 1993). These effects are important because flight tasks place demands on peripheral processing; scanning a scene or an instrument panel in a succession of glances depends on processing features in the periphery. Experiment 1 tested the spatial distribution of threshold recognition performance, and measured each viewers overall threshold performance, as well as his or her manual tracking ability. Duration thresholds were measured for recognition at three retinal eccentricities in the four visual field quadrants, to determine how performance would fall off in the periphery. Twenty observers (two of whom were left-handed) participated, completing four training sessions and one test session.

Stimulus sequences were generated using an Iris workstation with a collimated display system that simulated distance focus. A joystick was used to collect tracking data and recognition responses. In the recognition task, aircraft were displayed at twelve possible visual locations, and viewers identified the aircraft as belonging to a fighter (e.g. F-16) or a non-fighter (e.g. Boeing 747) target set which had been studied previously. Aircraft were shown in four orientations. Images were size-normalized to a 2-deg width, to eliminate scale cues between different planes. In each trial, a central fixation cross was displayed. An aircraft was then displayed briefly in the upper right, upper left, lower left, or lower right quadrant of the screen, at an eccentricity of 5, 9, or 13 deg (Figure 1). A mask was then displayed at the same location as the plane, after which the viewer pressed one of two joystick buttons to indicate which group the aircraft belonged to. Feedback was provided during training.

The minimum viewing duration required for 75% performance was measured for each screen location. Thresholds were measured using the step method (Simpson, 1989), an adaptive paradigm that used the subject's response history to adjust duration across trials and home in on the target performance level. Viewing duration was adjusted at each location in increments between a possible minimum of 17 ms (one video frame) and a maximum of 250 ms, above which the viewer would be able to make a saccade and fixate the aircraft. A 32-trial run comprising views of eight aircraft (four fighters, four nonfighters) in four orientations was conducted at each of the twelve screen locations. Trials from the twelve runs were intermixed randomly in each 384-trial session. The threshold estimates reached at each screen location were recorded for analysis in a 3 (eccentricity) x 4 (quadrant) within-subjects design. Response times were also recorded.

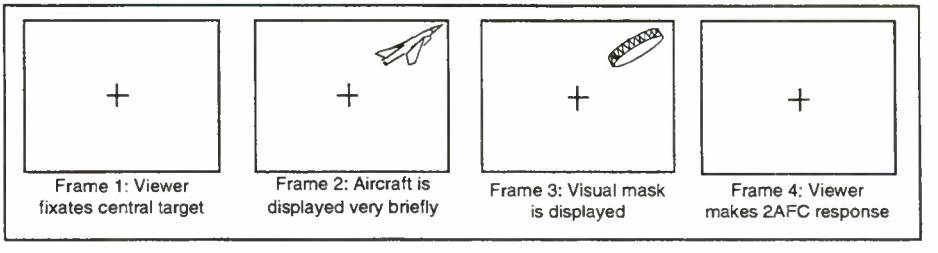


Figure 1. Time course of an aircraft recognition trial.

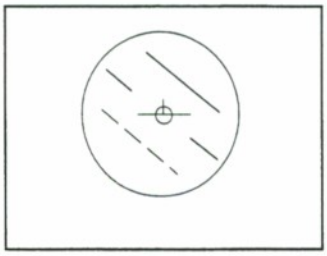


Figure 2. Inside-out roll attitude indicator, used for critical and subcritical tracking tasks.

Two performance measures were used to assess viewers recognition and manual tracking. These enabled the subsequent testing in Experiment 2 (in which all subjects also participated) of whether individuals near-threshold and tracking ability would predict their performance in a dual-task situation. The first measure was the viewers mean threshold, averaged across screen locations. The second measure assessed performance on a critical tracking task (Jex, McDonnell, & Phatak, 1966). Viewers used the joystick to control an increasingly-unstable attitude display located in the center of the screen. The instability that each viewer could tolerate before losing control was measured. The display configuration was inside-out; a stationary aircraft icon was bracketed by parallel line segments which represented the horizon (Figure 2). The horizon lines rolled left and right to indicate attitude changes that would result from random stick perturbations. The viewer corrected the attitude with the joystick. Performance was assessed using the (instability index. Sessions began with (at 1.5; at this value, the display forgives slow attitude corrections. As the trial progressed, (was increased gradually until the viewer lost control. After each crash, (was reset to a lower value and a new trial began. Trials continued for 7500 frames, displayed at 15.5 frames/s. The performance measure was each block's median crash (value.

Results and Discussion

Recognition threshold durations increased with eccentricity, $F(2, 38)=61.3$, $p<.001$ (Figure 3), and viewers took longer to classify aircraft at the greatest eccentricity value, $F(2, 38)=4.45$, $p<.02$. Threshold durations were shorter in the upper and right visual fields; this quadrant effect was nonsignificant, however. Similarly, response latencies were shorter in the upper and right visual fields, but this effect was nonsignificant, $F(3, 57)=2.38$, $p=.079$.

Individual differences were identified. A central clustering of threshold means between 100 and 140 ms was observed. Three subjects fell below this window, requiring markedly less viewing time to classify the planes than did the rest of the subjects. Three means were greater than the central cluster, indicating subjects who required more viewing time. Experiment 2 would test whether this threshold measure would predict the viewers ability to perform two tasks at once. The second performance measure, (, yielded a similar distribution; certain viewers distinguished themselves as better or poorer at tracking. Endsley and Bolstad (1994) observed that pilots' manual tracking performance was correlated with performance on a situational awareness battery, and hypothesized that pilots with good tracking skills can devote more attention toward situation assessment. If this principle applied generally to the performance of multiple perceptual tasks, viewers with the best tracking ability in a population should suffer the least performance loss when required to switch to dual-task conditions.

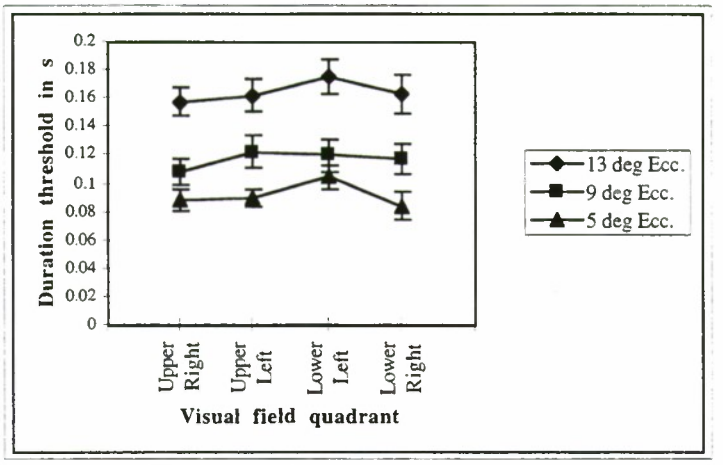


Figure 3. Experiment 1: 75% correct duration thresholds.

Experiment 2: Does near-threshold recognition (or manual tracking performance) predict dual-task performance?

Experiment 2 tested the principle that exceptional ability in processing fleeting targets (or in tracking) confers an advantage for dual-task performance, and tested whether visual field recognition biases would emerge under dual-task workload. The design followed from the theoretical assumption that a pilot's capacity to share attention among several tasks is relevant to cockpit performance (Damos, 1978; Endsley & Bolstad, 1994; North & Gopher, 1976). Weinstein and Wickens (1992) examined this question, and determined that using central and peripheral vision simultaneously to perform two disparate cockpit tasks can cause visual overload.

Experiment 2 tested viewers' ability to manage this overload. The same twenty subjects practiced performing the tasks from Experiment 1 simultaneously, in three training sessions, and then completed single- and dual-task testing in a fourth and fifth session. In dual-task conditions, viewers were instructed to treat recognition as the primary task, and the tasks were interdependent. In order to complete the recognition task, subjects must not crash on the secondary tracking task in the center of the screen; conversely, recognition feedback was included to remind subjects to protect the primary task. The displays and design for the recognition task were as in Experiment 1, except that viewing durations were fixed at 83, 100, and 150 ms, respectively (the average thresholds at each eccentricity in Experiment 1), for targets at 5, 9, and 13 deg eccentricities. Eight aircraft were displayed in four orientations at each screen location. The twelve 32-trial runs, presented randomly, comprised a recognition session. Percent-correct and response latencies were recorded, with number of tasks (single vs dual), eccentricity, and quadrant as within-subject factors. In presenting all viewers with the same objective challenge to recognize near-threshold aircraft as a first priority, this task tested the prediction that individuals with the lowest recognition thresholds would suffer least from the dual-task transition. In contrast, individuals with high thresholds would be taxed more by views of the same duration, so their dual-task performance would be expected to suffer more.

The tracking display also resembled that in Experiment 1; however, instead of increasing, instability remained constant at a manageable, subcritical value, 55% of the subject's peak critical (in Experiment 1. Subcritical tracking (Kenyon & Kneller, 1993; Previc, Kenyon, Boer, &

Johnson, 1993) allowed subjects to complete dual-task sessions without crashing frequently. A forcing function was used to simulate gusts rolling the horizon left and right. Tracking test sessions comprised eight blocks. RMS tracking error was recorded for each block in single- and dual-task conditions. RMS replaced the adaptive (measure, because instability remained constant throughout the task.

Correlations were run on subjects' performance measures, to assess possible links between single-task recognition and tracking, and dual-task performance. These measures included the following: the subjects mean recognition threshold and median critical-tracking ((in Experiment 1); the total percentage of correct aircraft classifications across all screen locations in single- and in dual-task conditions, and the difference between these; the mean recognition response time in single- and dual-task conditions, and the difference between these; and the mean proportional increase in tracking error[1]. calculated as (Dual-task RMS - Single-task RMS)/Single-task RMS.

Results and Discussion

Effects of task, eccentricity, and quadrant.
There were more correct responses in single- than dual-task conditions, $F(1, 19)=51.4$, $p<.05$ (Figure 4). Neither eccentricity[2] nor quadrant influenced percent-correct. No significant interactions were found, although an interactive trend suggested that recognition performance suffered under dual-task conditions in the lower and left visual fields. Response latencies were shorter in dual-task conditions, $F(1, 19)=13.5$, $p<.005$, and increased with eccentricity, $F(2, 38)=18.4$, $p<.001$. Quadrant influenced RT, $F(3, 57)=4.83$, $p<.01$. Eccentricity x quadrant interaction missed significance, $F(6, 114)=2.17$, $p=.051$. No other interactive RT effects were found. RMS tracking error (not shown) was greater in dual-task conditions, $F(1, 19)=76.9$, $p<.001$ and varied with presentation block, $F(7, 133)=5.75$, $p<.001$. There was an interaction between number of tasks and block, $F(7, 133)=2.73$, $p<.05$.

Visual overload clearly hindered dual-task recognition even though viewers were instructed to give it priority. Although the importance of classifying aircraft correctly (not quickly) was stressed in the instructions, an apparent speed-accuracy trade-off occurred whereby viewers responded slightly less accurately, but faster, in dual-task conditions. Similarly, manual attitude control suffered considerably under dual-task loading and deteriorated in the later test blocks. Like the recognition data, the tracking data are consistent with an attention model in which a visual resource pool is shared across disparate tasks performed in central and peripheral vision (Weinstein & Wickens, 1992).

Processing limitations appear to vary with eccentricity and visual field location. Previc and Blume (1993) proposed a visual performance contour for a distance-biased attentional system that favors the upper right quadrant, and whose evolutionary function is to search for and recognize objects. The main quadrant effect on response times was consistent with such a spatial attention bias: Viewers took longer to classify aircraft in the lower and left visual fields. Furthermore, the interactive task x quadrant trend suggests that spatial biases in recognition emerged in high-workload conditions.

The dual-task performance effects might be explained according to the following account. The viewer performs the secondary tracking task continuously, consuming resources from a visual processing pool. At intervals, aircraft targets required the viewer to switch resources to the primary task; good performance on both tasks necessitated efficient attention switching. The lack of resources for the secondary task would relax tracking criteria whenever an aircraft target

---

1 Although dual-task RMS might seem to be the obvious measure for assessing tracking, this would be appropriate only if the objective difficulty of the tracking task was uniform across subjects. Here, (was scaled to the subjects measured tracking ability from Experiment 1; hence the dual-task RMS increase was used. See Beer, Gallaway, and Previc (in press) for detailed description of this design.

2 This indicates that the thresholds from Exp. 1 (on which the display durations at each eccentricity were based) and percent-correct were mutually consistent; thresholds registered the greater viewing time required at wider eccentricities.

appeared, allowing attitude to topple around more. Dual-task responses were faster because this control raggedness was an incentive to divert resources quickly back to tracking.

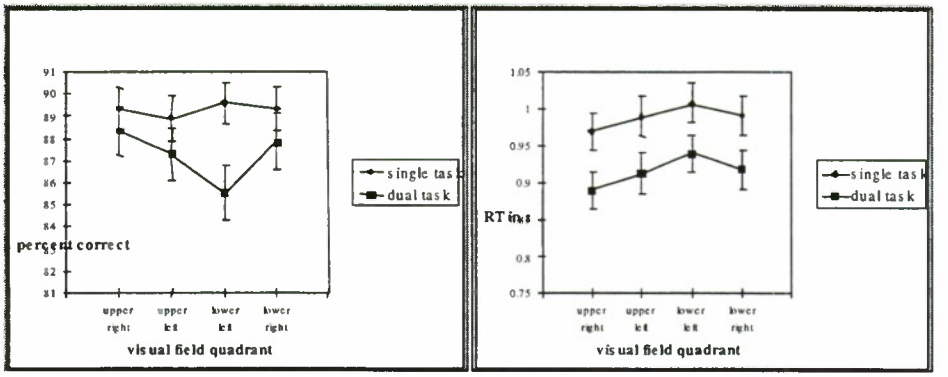


Figure 4. Experiment 2: Aircraft classification performance (left) and response latency (right).

Individual performance measures, single- and dual-task.
Recognition thresholds were correlated with percent-correct in dual-task conditions (Table 1). Certain viewers needed less viewing time than others to classify aircraft reliably, and these viewers identified more aircraft correctly at fixed durations, an advantage that survived the transition to dual-task conditions. However, the nonsignificant correlation between recognition thresholds and the dual-task increase in tracking error failed to provide strong evidence that an individual's near-threshold processing predicts his or her dual-task performance.

The experiment's second goal was to test whether viewers who are good at manual tracking can better handle stressful multiple-task conditions. It is notable that critical tracking, which pressures the viewer continuously until he crashes, appears to involve the same resources that underlie dual-task aircraft classification: Viewers who could withstand greater tracking instability were better at preserving dual-task recognition. This finding is consistent with the hypothesis that tracking ability is associated with efficient use of attention resources (Endsley & Bolstad, 1994).

Table 1. Exp. 2: Correlations Between Performance Measures

Measure 1	Measure 2	Pearson r	Spearman
Single Correct	Exp. 1: Recognition Threshold[1]	**-.619***	**-.728***
Dual Correct	Exp. 1: Recognition Threshold	**-.628***	**-.609***
(Correct	Exp. 1: Recognition Threshold	-.298	-.097
(RMS	Exp. 1: Recognition Threshold	.308	.128
Single Correct	Exp. 1: Critical (	.223	
Dual Correct	Exp. 1: Critical (	.433	
(Correct	Exp. 1: Critical (	**.492***	
Dual Correct	Single Correct	**.848***	
(RT	Exp. 1: Recognition Threshold	**-.579***	**-.695***
(RT	Single Correct	**.600***	
(RT	Dual Correct	**.460***	

1 Threshold means assumed slightly nonnormal distributions; (is therefore included.

In an additional unexpected finding, viewers who were good at aircraft classification were less likely to speed their responses on that task in high-workload conditions. A possible explanation is that viewers' awareness of the effort required to switch attention might influence their allocation strategy and performance. High-threshold, less proficient viewers may have stolen more time from recognition because they were more acutely aware of the resources they were spending on it, and more aware that this expenditure taxed their ability to maintain attitude. In the classroom analogy of attention switching (Sperling & Dosher, 1986), this awareness constitutes a cue to dash from (difficult) aircraft class early to attend (easy) manual tracking. Awareness of multitask workload might prove to be useful for predicting SA; for these tasks, it appears that this kind of situational arousal would be a liability.

Conclusions

The experiments identified three aspects of near-threshold and manual tracking performance that are relevant to pilot performance in stressful cockpit situations. Although adaptive threshold estimation is potentially useful for assessing viewers' recognition of real-world targets, no strong evidence was found to indicate that visual recognition thresholds will predict the viewers performance on a second, disparate cockpit task performed concurrently. The second finding, paradoxically, indicated a converse relation: Critical tracking, which pressures the viewer until he loses control, tapped a competence that also supported the maintenance of visual attention in the periphery. Critical tracking may be a promising tool for examining further the competence framework that underlies multitasking and cockpit SA. The third class of findings suggested that if the pilot is busy, the worst place for a fleeting target to show up is in his or her lower left visual field.

References

Beer, J. M. A., Gallaway, R. A., & Previc, F. H. (in press). Do individuals visual recognition thresholds predict performance on concurrent attitude control flight tasks? *International Journal of Aviation Psychology.*

Damos, D. L. (1978). Residual attention as a predictor of pilot performance. *Human Factors, 20,* 435-440.

Endsley, M. R., & Bolstad, C. A. (1994). Individual differences in pilot situational awareness. *The International Journal of Aviation Psychology, 4,* 241-264.

Hartman, B. O., & Secrist, G. E. (1991). Situational awareness is more than exceptional vision. *Aviation, Space, and Environmental Medicine, 62,* 1084-9.

Jex, H. R., McDonnell, J. D., & Phatak, A. V. (1966). A critical tracking task for manual control research. *IEEE Transactions on Human Factors in Electronics, HFE-7,* 138-145.

Kenyon, R. V., & Kneller, E. W. (1993). The effects of field of view size on the control of roll motion. *IEEE Transactions on Systems, Man, and Cybernetics, 23,* 183-193.

North, R. A., & Gopher, D. (1976). Measures of attention as predictors of flight performance. *Human Factors, 18,* 1-14.

Previc, F. H., & Blume, J. L. (1993). Visual search asymmetries in three-dimensional space. *Vision Research, 33,* 2697-2704.

Previc, F. H., Kenyon, R. V., Boer, E. R., & Johnson, B. H. (1993). The effects of background visual roll stimulation on postural and manual control and self-motion perception. *Perception & Psychophysics, 54,* 93-107.

Simpson, W. A. (1989). The step method: A new adaptive psychophysical procedure. *Perception & Psychophysics, 45,* 572-576.

Sperling, G., & Dosher, B. A. (1986). Strategy and optimization in human information processing. In K. Boff, J. Thomas, & L. Kaufman (Eds.), *Handbook of perception and human performance* (Vol. 1, pp. 2.1-2.65). New York: Wiley.

Weinstein, L. F., & Wickens, C. D. (1992). Use of nontraditional flight displays for the reduction of central visual overload in the cockpit. *The International Journal of Aviation Psychology, 2,* 121-142.

The Virtual Patient – An Application of Situation Awareness for Sysytem Design and Training in Intensive Care

S. Keith Adams[1], Shane P. Babin[2] and Zeinab A. Sabri[2]

[1] Iowa State University
[2] Technology International, Inc. of Virginia

Introduction

The need for applying human factors engineering or ergonomics to medical systems and patient care has been recognized for several decades (Rappaport, 1970; Gopher et al., 1989; and Galer and Yap, 1980). Ergonomic design challenges in medical systems have evolved in patterns similar to those found in other areas of technology such as aviation, nuclear power plants, and defense systems management. Early problems centered on basic control and display design, layout and compatibility have evolved into problems centered on the structure and dynamics of cognitive tasks. As these changes occurred, human operation has evolved from one of direct monitoring, rule-based decision making and control activation to one of information acquisition, integrated knowledge based decision making and systems management or supervisory control. The former plethora of individual measuring and monitoring instrumentation units has, to a considerable extent, been replaced with integrated Intensive Care Unit (ICU) systems capable of monitoring many physiological variables simultaneously and displaying data in real time graphic or tabular form, together with recent or past historical patient data, or even comparative data from other patients all combined in an arranged VDT display for analysis by the attending nurse or technician. The design challenge to the human factors engineer is to develop methods that present medical information in ways that promote correct diagnosis and decision making in a minimum amount of time. Correct diagnosis and decision making may also involve the use of information that is not monitored through instrumentation, but is observed directly, such as the patient's pallor, breathing sounds, perspiration rate, or indication of pain. All of the incoming information to the nurse or technician forms a multi-channeled, interactive set of dynamic data streams describing an evolving response to internal conditions and medications and requiring varying focus of attention to particular variables or combinations thereof and the conditions they may indicate. Such a human/equipment/environment scenario presents the need for situation awareness (SA) in which present elements of information are monitored, integrated into a cognitive pattern and used to predict the future state of the system (patient) (Endsley, 1990, and Adams and Pew, 1989).

Situation awareness requires perception, comprehension (diagnostic ability) and projection (prediction) of future states of the system being monitored and controlled. A well-designed ICU serves as a decision aiding mechanism in facilitating these processes. Final decisions in medicine must always be human decisions.

The patient/equipment/nurse interaction using conventional non computer-integrated equipment is depicted as a system in Figure I. A subset of all physiological events is monitored in terms of known measurable factors. These are interpreted and acted upon by the nurse or technician whose major role is to collect and report data and symptoms. The role of equipment is to sense or detect

physical signals and to present these in real time on a real-time recording (strip chart). The dashed lines indicate boundaries of the domain of traditional medical equipment. The human equipment interface occurs across the upper dashed line and consists of the traditional observe, perceive, interpret, respond activity for a basic control/display system.

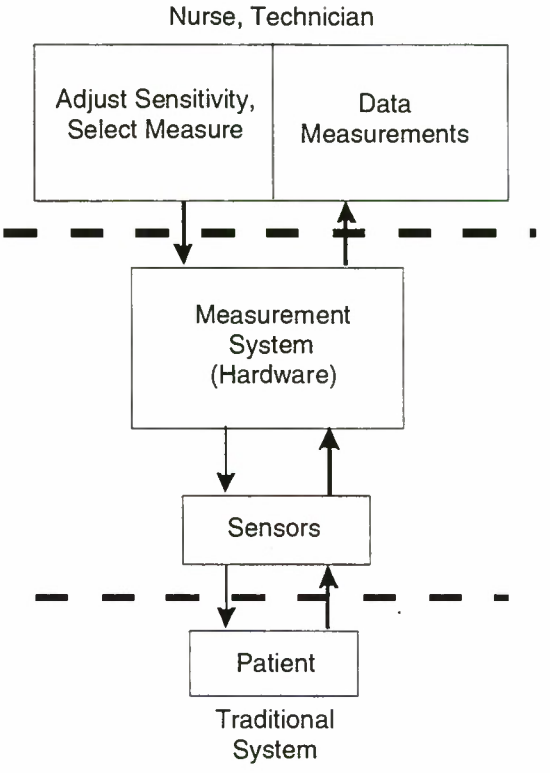


Figure 1. Traditional Role

Figure 2 presents a basic system diagram for a modern ICU in which a set of factors is selected from a larger set of measurable factors and presented to the nurse or technician for analysis, integration and decision making. Also available in this information system are sets of medical facts and procedures as well as data from other hospitals and databases. Since the ICU can store and retrieve information in graphic or tabular form and can summarize and indicate trends in data over time, and can also present menus and factual data from other cases or the patient's past medical history, substantially greater integration and analysis are possible and generally expected in dealing with an acute health crisis warranting the use of an ICU. In this case the nurse or technician can select and manage information as well as monitor and record it. Advanced levels of diagnostic analysis and decision making could require a broader medical knowledge possessed by a physician. The amount of cognitive activity and SA preceding decisions and actions taken is or can be substantially more than occurs using a set of independent parameter monitoring instruments

and machines. Situation awareness in this case extends from the set of selected measurable factors (since they were chosen) through the set of actions taken.

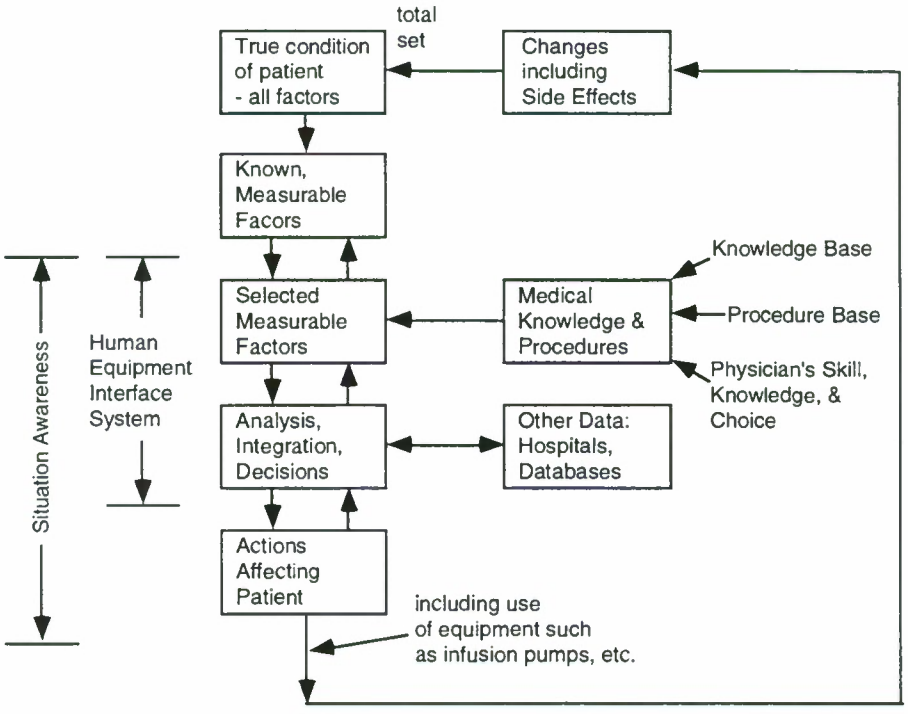


Figure 2. ICU Information Sysytem

A more advanced ICU system is depicted in Figure 3. This system contains more levels of integration and data processing. It has the capability of comparing data from the patient being monitored by the ICU with other data including recent and previous records, typical model examples for comparisons such as ECG patterns or combinations of symptoms or data or diagnoses from other hospitals. Probabilistic risk assessments and lists of advisable medical altematives with advantages and risks could be presented.

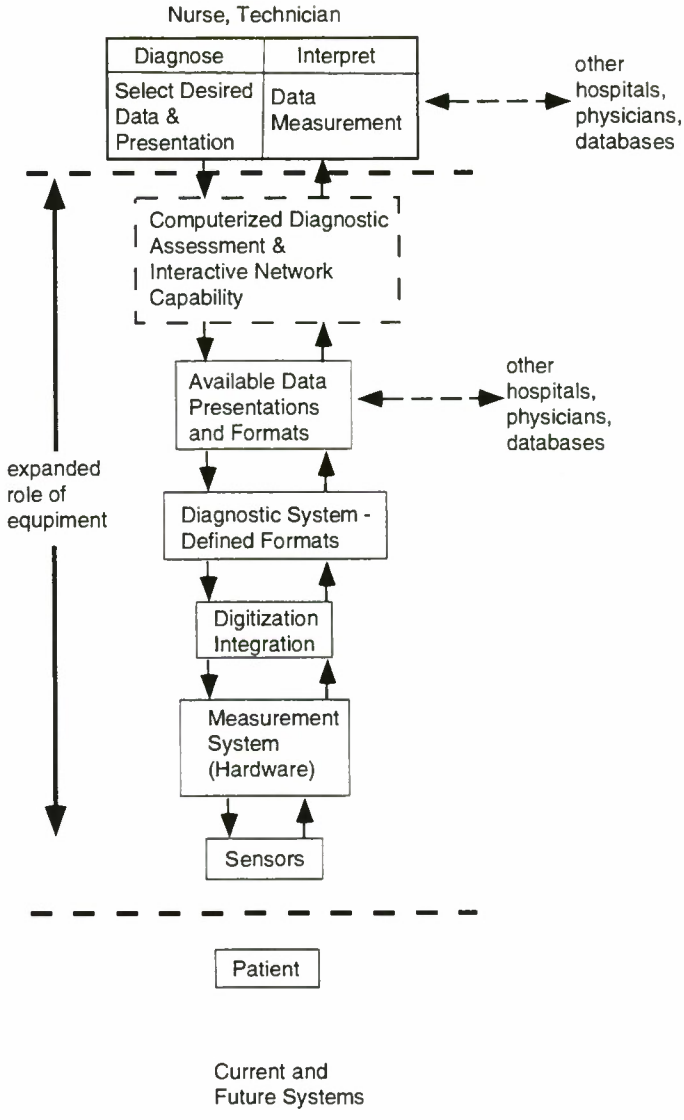


Figure 3. Expanding Role of ICU

The possibility for more sophisticated decision making exists but it comes with the price of a still higher cognitive workload and SA. It is important therefore, that ergonomic techniques be developed that can be used to evaluate and improve the human/ICU interface for present and upcoming ICU hardware and software designs. Also needed are developments and improvements int he use of ICU's as training devices for nurses and technicians, capable of simulating dynamic, interactive and time dependent information as it is presented in an ICU attached to a real patient.

Also possible in such systems is the transfer of knowledge gained through experience and the use of non-measured symptomatic data from senior nurses and technicians to those beginning their professional careers. The topic of this paper is focused on the concept of the Virtual Patient, a dynamic software representation of factors that would be monitored in a given ICU/patient interaction. The scope of application in this discussion is limited to the application of ICU's in intensive cardiovascular care.

Computerized ICU's in Intensive Cardiovascular Care

Cardiovascular monitoring and analysis comprise the major roles of the ICU in modern medicine. The many factors involved in the performance of the heart and the flow and gas content of blood as it delivers oxygen to the brain, organs and muscles are interrelated and form paradigms indicating the need for specific drugs, surgery or therapy.

Upcoming generations of ICU's will be capable of monitoring more variables than current modules including capnography (CO2 monitoring), continuous arterial blood analysis (pulse oximetry, transcutaneous gas measurement), enabling or improving more sophisticated analyses such as intrapulmonary shunting or dead space (Meyer, 1993). Blood electrolytes such as potassium as well as blood urea nitrogen, hematocrit and coagulation can already be monitored. Glucose monitoring is also done routinely in many situations. When these are added to the traditional ECG variables, heart rate and blood pressure data, a substantial increase in cognitive workload and need for SA result. Computer programs such as APACHE (Acute Physiology and Chronic Health Evaluation) provide assistance in decisions regarding the selections of treatment based on probable outcomes (Knaus, et al., 1985 and 1991). Future ICU's will need to provide assistance in the acquisition, integration and management of medical information for patients undergoing intensive care. Logic paradigms (truth tables) have already been developed for diagnosing pulmonary conditions (embolism, pleurisy, pneumothorax, pneumonia, plural effusion and atelectasis) based on immediate clinical assessment of breath sounds, cough, dyspnea, fever, sputum, chest pain, and other factors (Stiesmeyer, 1993). For the cardiovascular system, trend analysis of venous oxygen saturation (SVO2 monitoring) can signal present or upcoming events such as impending metabolic acidosis or acute respiratory failure for earlier than traditional factors such as cardiac output (Hayden, 1993).

Tables have been developed that relate the drop in SVO2 to various routine events and procedures (e.g. coughing, vomiting, daily baths, turning over, linen changes, visiting) as well as more serious events (e.g. hemorrhage) (Hayden, 1993). Administration of certain intravenous fluids (e.g. plasma, vitamin K, protamine) will also decrease SVO2. Digital analysis of waveform data could serve as an additional decision aid in some situations. Analysis of cardiac activity using ECG waveform analysis requires intensive, rapid and accurate visual interpretation and logical analysis, especially when physiological reactions to a present condition or impairment are compounded with those produced by administered drugs. Errors in diagnosis and decisions regarding treatment are possible and even likely to occur under such situation (Graver and Cavallaro, 1993). Space precludes detailed examples, but in summary, the use of logic paradigms and combinational analysis of symptoms are needed. These can be aided through advanced ICU software and appropriate tabular or menu displays. Current nursing textbooks discuss and tabulate ECG measures along with other symptoms and their indications along with artifacts and their probable causes (Bruner and Suddarth, 1986; and Clochesy et al., 1993). It is necessary that nurses, technicians and physicians be able to interpret, integrate, decode and act upon increasingly larger sets of medical data, much of which is dynamic. This requires improved training involving programmed or simulated data presented in real time formats that occur when treating actual patients. In small rural community hospitals, many serious critical conditions may be encountered infrequently enough to make errors in diagnosis and treatment more likely. In a study conducted

by the University of Wisconsin at twelve rural community hospitals in southern Wisconsin, tests were given to 461 practicing physicians, staff nurses and respiratory therapists over a four-year period (Birnbaum, 1994a, and Birnbaum, 1994b). Only 39.6% of the nurses and 64.1% of the physicians correctly identified third-degree A-V block. One third of the nurses and 22% of the physicians did not correctly identify coarse ventricular fibrillation. The effects and purposes of atropine and eponephone were not understood. Propranolol was incorrectly selected for treatment of third degree A-V block by 31.6% of the nurses and 22.8% of the physicians. No improvement in overall performance occurred during the four-year period. The need for portable, cost-effective cardiac training for rural community hospital staff personnel was clearly demonstrated in this study.

The Virtual Patient Concept

The concept of the virtual patient for cardiovascular care is based upon dynamic, real-time simulation of a patient's cardiovascular and case-related date using computer software where output is displayed on an ICU video monitor. Cardiopulmonary conditions can be simulated and interacted with current ICU software or run independently on a personal computer platform. Use of virtual patient software will permit the evaluation or comparison of present ICU displays and human/ICU interfaces based on time and errors in diagnosing and treating simulated cases. Also proposed for development are defined scenarios or cases to be used with testing/evaluation software which will record operator responses as they relate to changing vital signs. The virtual patient system will serve several functions:

1. It will promote familiarity with patient monitoring equipment by presenting the scenarios that require interaction with ICU equipment.
2. It can be used as a training or teaching aid to reinforce diagnostic reasoning and procedures.
3. It can be used to test knowledge by recording operator responses to varying patient conditions and measuring operator performance against changing conditions in real time (or with indicated time spans if real changes occur slowly).
4. It provides a method of evaluating ICU's by determining their effectiveness and user friendliness to facilitating correct diagnosis in a minimum amount of time as a result of display content, interfaces and presentation methods.

In addition to data derived from direct patient instrumentation, symptoms can be listed on the display monitor which incorporate the experience of senior nurses or technicians. In a study of the detection of sepsis in neonatal care, it was found that different nurses used different cues to recognize sepsis (Systematic bacterial infection) in infants in neonatal ICU's (Crandall and Getchell-Rester, 1993; and Crandall and Gamblian, 1991). Many sepsis indicators recommended in textbooks were never mentioned by experienced nurses while others were not mentioned in the literature but very important as directly observed symptoms or trends, were used often and very effectively. Information collected through extensive recorded interviews was used to assemble a comprehensive set of guidelines for the diagnosis and treatment of sepsis. New indicators were identified and incorporated into the new guide which did not exist in the present manuals. As a result, 46% of nurses surveyed gained new information about sepsis assessment. Another 27% stated that the guide helped to reinforce important points. Using this approach, symptomatic information can be incorporated into the virtual patient software.

Several levels of sophistication are possible using virtual patient software in the training or teaching mode. At the first or basic level, conditions and data, ECG, other graphic or tabular displays) will be shown to identify and illustrate symptoms arising from given specified conditions

(e.g. pulmonary embolism, atrial flutter, tachycardia, ventricular fibrillation). At the second level, several conditions can be presented in random order with the operator required to identify each one. At the third level, other factors or symptoms are presented such as recent health history, diseases or complicating factors which would affect the recommended treatment. Also at the third level, the virtual patient will interact with related treatment with the effects of the interaction shown to the operator. At the fourth level, a complete cardiovascular or cardiopulmonary health problem is simulated. The operator must discover and diagnose the problem or condition by selecting and analyzing available data. The operator then makes a final decision and selects a treatment or medication whose effects are presented as in the third level. SA is required when interacting with the virtual patient at the third and fourth levels.

The Virtual Patient and Situation Awareness

Medical treatment requires the perception and integration of many facets of data and symptomatic information in forming a diagnosis, and also the anticipation of the future state of the patient under various treatment options. Situation awareness is already being recognized as a valuable conceptual tool for studying the dynamics of anesthesiology (Gaba and Howard, 1995). Simulators capable of representing patient reactions have been developed along with a clinical mannequin to permit many of the actions needed during surgery (Gaba, 1994, Gaba and DeAnde, 1989). Surgery presents a particularly complex case for SA because of interactions among the anesthesiologist, surgeon, patient and nurses. The virtual patient model under development for cardiovascular and cardiopulmonary application is less complex and intended for use in a single patient/nurse, technician interaction.

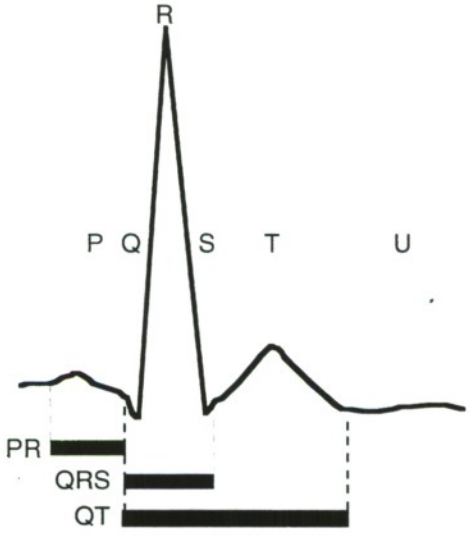


Figure 4. ECG Waveform

One typical example of application for the virtual patent is in creating simulated (or reproduced from actual cases) 10 or 12 ECG pattems representing cardiovascular disorders. The ECG has a standard characteristic waveform consisting of the P, Q, R, S and T portions with an inter-beat U portion (Figure 4). The ECG lengths (duration) and amplitudes of these segments or their distortion or absence indicate specific cardiac malfunctions based on depolarization and repolarization of portions of the heart muscle tissue. For example atrial enlargement (right or left ventricular hypertrophy) is detected by excessive voltage in the P wave. An inverted T wave would indicate abnormal repolarization. For diagnostic purposes, waves are described as upright positive, inverted negative, notched, diphasic (T-phase), flat, tall, broad or narrow. A tall Q wave could indicate a myocardial injunction. Detailed examination of the QRS wave would indicate whether a myocardial injunction, ventricular hypertrophy or ventricular conduction defect (branch block) was occurring, based on QRS wave duration. Heart rate by itself can permit differentiation between conditions. Sinus-tachycardia is indicated by a heart rate of between 100 and 160 beats per minute. This can occur normally as from exercise, anxiety or higher caffeine intake, or it could be the result of decreased cardiac output as caused by hypovolemia or cariogenic shock. Atrial flutter produces a saw-toothed wave in place of the normal P wave typically at a frequency of from 250 to 350 beats per minute. Atrial flutter is caused by valvular heart disease, coronary artery disease and thyroptosises. Atrial fibrillation, on the other hand, produces irregular, rapid waves in place of P waves. It is caused by arteriosclerotic or rheumatic heart disease. Because of the irregular pattern, the atria are unable to eject all the blood inside of them, an effect which could cause stagnation of blood within the atria and possible clot formation. When sinus tachycardia, atrial flutter and atrial fibrillation at rates higher than 100 beats per minute may indicate a broadly defined condition known as supraventricular tachycardia (SVT),a term describing any irregular rhythm originating above the ventricle with a rapid heart rate. SVT is produced by a number of factors including myocardial infraction, valve disorders, and pulmonary embolus. A broad (>0.12 sec) and bizarre QRS wave sequence can be caused by premature ventricular contractions (PVC) or a premature ventricular beat (PVB). T wave inversion is also common under this condition because of unusual depolarization. The condition known as Torsade de Pointes is characterized by alternating up and down pointing of the peaks of the QRS waves and a long QT interval. These disorders and other conditions can be simulated using virtual patient software. Standard decision trees for treatment and medication exist for tachycardia and other general symptoms based on ECG symptoms and diagnosis. Procedures similar to those used in discussing the detection of sepsis in neonatal care can be incorporated into the virtual patient by including symptoms not monitored using instrumentation along with the experienced judgement of senior physicians, nurses and technicians. Reactions of the virtual patient to operator responses regarding medication or treatment, using levels 3 and 4, create the future state of the patient, which must be anticipated by the operator in diagnosing the condition or choosing a course of action to be taken. Thus, SA can be created and hopefully taught and explored using the virtual patient concept.

References

Adams, M. J., R. W. Pew. (1989). Cognitive management of complex, semi-automated systems. *IEEE Conference of Systems, Man and Cybernetics*, 3, pp. 1216-1220.

Birnbaum, M. L., et al. (1994a). Need for Advanced Cardial Life-Support Training in Rural, Community Hospitals. *Critical Care Medicine*, 22 (5), pp. 735-740.

Birnbaum, M. L., et al. (1994b) Effect of Advanced Cardiac Life-Support Training in Rural,Community Hospitals. *Critical Care Medicine*, 22 (5), pp. 741-749.

Bruner, L., D. Suddarth. (Eds). (1986). *Manual of Nursing Practice*, 4th ed. Philadelphia, PA: J. B. Lippincott Co.

Clochesy, Breu, Carolin, Rudy. (1993). *Critical Care Nursing*. Philadelphia, PA: W.B. Saunders Co.

Crandall, B., B. Getchell-Rester. (1993). Critical decision method: A technique for eliciting concrete assessment indicators from the intuition of NICU nurses. *Advances in Nursing Science*, 16 (1), pp. 42-51.

Crandall, B., V. Gamblian. (1991). *Guide to Early Sepsis Assessment in the NICU*. Fairborn, OH: Klein Associates, Inc.

Endsley, M. R. (1990). Design and Evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society*, 34th Annual Meeting, pp. 1509-1513.

Gaba, D. M., S. K. Howard. (1995). Situation awareness in anesthesiology. *Human Factors*, 37 (1), pp. 20-31.

Gaba, D. (1992). Improving Anesthesiologists' Performance by Simulating Reality. *Anesthesiology*, 76, pp. 491-494.

Gaba, D., A. DeAnde. (1988). A comprehensive anesthesia simulation environment: recreating the operating room for research and training. *Anesthesiology*, 69, pp. 387-394.

Galer, I. A. R., B. L. Yap. (1980). Ergonomics in intensive care: applying human factors data to the design and evaluation of patient monitoring systems. *Ergonomics*, 23, pp. 763-779.

Gopher, D., et al. (1989). The nature and causes of human errors in a medical intensive care unit. *Proceedings of the Human Factors Society*, 33rd Annual Meeting, pp. 956-960.

Graver, K., D. Cavallaro. (1993). What did these care teams do wrong? *American Journal of Nursing*, 93 (5), pp. 16E-16J.

Hayden, R. A. (1993). Trend-spotting with an SVO2 monitor. *American Journal of Nursing*, 93 (1), p. 26-33.

Knaus, W. A., J. E. Zimmerman, D. R. Wagner, et al. (1991). APACHE- Acute physiology and chronic health evaluation: A physiologically based classification system. *Critical Care Medicine*, 19, pp. 591-597.

Knaus, W. A., J. E. Zimmerman, D. R. Wagner, et al. (1985). APACHE II - A Severity of disease classification system. *Critical Care Medicine*, 13, pp. 818-829.

Meyer, C. (1993). Visions of tomorrow's ICU. *American Journal of Nursing*, 93 (5), pp. 27-31.Rappaport, M. (1970). Human factors applications in medicine. *Human Factors*, 12 (1), pp. 25-35.

Stiesmeyer, J. K. (1993). A Four-Step approach to pulmonary assessment. *American Journal of Nursing*, 93 (8), pp. 22-28.

An Assessment Of Situation Awareness In An Air Combat Simulation: The Global Implicit Measure Approach

Bart J. Brickman[1], Lawrence J. Hettinger[1], Merry M. Roe[1], Dean Stautberg[1], Michael A. Vidulich[2] Michael W. Haas[2] and Robert L. Shaw[3]

[1] Logicon Technical Services, Inc.
[2] Wright-Patterson A.F.B
[3] F.C.I. Associates, Inc.

Abstract

This paper describes an approach to the assessment of pilot situation awareness (SA) in a simulated air combat task. The technique under investigation, referred to as the Global Implicit Measure approach, is based on the use of objective measures of the status of the pilot-aircraft system taken with respect to previously defined rules of engagement. We believe that this approach provides several advantages over other methods of assessing SA. Specifically, no overt, SA-specific response is required of the subject and no interruption of task performance is required. Therefore, problems associated with altering the nature of the task by interrupting its performance are avoided. We are currently performing an initial investigation of this approach within the context of an evaluation of alternative interface configurations for fighter aircraft cockpits. The goal of our work is to establish the basis for developing advanced crewstation interfaces that will enhance crewmember SA. In addition, the ability to reliably detect degradation in SA on the basis of measures such as those being explored in the current effort may permit the development of adaptive interfaces that can counteract the effects of compromised SA.

Introduction

This paper describes an effort to develop a reliable, objective method for assessing pilot situation awareness (SA) in a simulated air combat task. This work is being conducted at the US Air Force Armstrong Laboratory's Synthetic Immersion Research Environment (SIRE) facility at Wright-Patterson AFB, Ohio, as part of a program to develop future generation airborne crewstations.

The research efforts of the SIRE facility are geared toward the development of virtually augmented, adaptive interfaces. Specifically, our work focuses on developing applications of virtual environment technology for use in future airborne crewstations (Hettinger, Nelson, Brickman, Haas & Roumes, 1995). The underlying assumption of this approach is based on the principle of presenting critical information to the crewmember in as *intuitive* and *salient* a fashion as possible. That is, information is designed to be made available in a manner that minimizes the

need for cognitive processing (i.e., interpretive and decision-making processes), while also taking full advantage of the multimodal information processing capabilities of the human operator. Many current aircraft displays occasionally place high demands on inferential reasoning processes as crewmembers must attempt to quickly comprehend relatively abstract and/or non-spatially localized information. Virtual environments are by definition multisensory, and interfaces developed using this technology can allow critical environmental and/or tactical information to be presented to the visual, auditory, and/or haptic modalities in a fashion that permits relatively easy extrapolation to the three-dimensional real environment. In addition, information can be made available to more than one sensory modality in cases in which a need for redundancy (or additivity) in information delivery is identified.

In addition, we are also pursuing the development of a logical architecture for the design of adaptive interfaces. Specifically, our intent is to develop interfaces whose informational content can vary in real time in response to both *external* features of the tactical environment (e.g., number and proximity of airborne threats) as well as *internal* characteristics of the human (e.g., fatigue, workload, and SA). In the latter case, the development of adaptive interfaces relies on the initial identification of non-obtrusive, reliable measures of pilot status. These measures may, in some cases, be physiologically based (e.g., Wilson, Fullenkamp, & Davis, 1990) and in other cases may be based on measures of overall system performance (e.g., Parasuraman, 1993).

The development of valid and reliable real-time measures of SA for use as inputs to an adaptive algorithm for airborne interfaces is one focus of our work. However, another area of emphasis centers around the use of measures of SA as a means of evaluating the performance of alternative interface configurations in supporting air combat performance. A major assumption of our research and development efforts is that the use of adaptive, virtual interfaces will help to maintain pilot workload and SA in optimal ranges, and will enhance the performance of the overall pilot-aircraft system. Therefore, in evaluating candidate interface configurations in simulated air combat evaluations it is essential to have reliable metrics for the assessment of these aspects of human performance. The development of such an instrument is another impetus for our work in the SA domain.

Past work in our laboratory has employed a retrospective memory probe approach to assess differences in SA associated with the use of different cockpit interface configurations. For instance, Hettinger, Nelson, and Haas (1994) used this method as one means of assessing differences between two cockpit configurations. One configuration, referred to as the "conventional" cockpit, consisted of displays and instrumentation commonly found in modern fighter aircraft. The second configuration, referred to as the "virtually-augmented" cockpit, represented a novel interface design that incorporated virtually-augmented visual displays, three-dimensional auditory information, and a head-slaved head-up display projected onto the transparent visor of a helmet-mounted display.

The interface evaluation consisted of a simulated air combat scenario in which a pilot, flying in one of the two possible cockpit configurations, attempted to intercept four computer-controlled enemy bombers. At the same time, two other pilots actively controlled simulated enemy fighter aircraft in an attempt to protect the bombers. This scenario resulted in frequent dogfights between the three "live" pilots involved in the evaluation. Measures of pilot-aircraft system output, workload, and SA were obtained in an attempt to illuminate differences in performance afforded by the use of the two interface configurations. SA scores were obtained using a set of 12 questions, designed to assess pilots' awareness of the status of various critical aircraft systems (e.g., weapons availability and status, fuel status, etc.) as well as features of the tactical situation (e.g., location of nearest threat, etc.). During each experimental trial, the air combat scenario was interrupted at two randomly selected intervals. During each interruption, six questions (randomly selected from the total set of twelve) were presented to the pilot flying either the conventional or virtually-augmented cockpit. During each of these stoppages, all cockpit and out-the-window displays were blanked out. Pilot's responses to the SA questions were subsequently classified as either "correct" or "incorrect" based on their correspondence to the actual state of affairs at the time the simulation was halted.

An overall SA score was calculated for each trial by summing the correct responses to the 12 SA questions and by converting these combined scores to percentages. The results revealed no statistically significant differences in SA between the two cockpit configurations. The overall SA scores were approximately equal for the conventional (51.39%) and virtually-augmented (53.47%) cockpit conditions. While it is certainly possible that this result reflected a true lack of difference between the two interface conditions, there appeared to be problems associated with the use of the memory probe technique that have motivated us to develop a less intrusive method that relies less on pilot memory. Specifically, the interruptions were a source of annoyance to the pilots, who were often engaged in a dogfight at the moment when the trial was abruptly halted. In addition, the method relies on pilots' conscious recall of mission-relevant information rather than their actual performance relative to the constraints of the mission. It was our opinion that assessment of the latter dimension of human performance rather than the former might provide a more accurate depiction of the state of SA. We were further encouraged in this observation by the finding that several performance outcomes that might reasonably be considered to reflect the outcome of pilot SA showed clear differences between the interface conditions. Specifically, the conventional cockpit experienced catastrophic ground strikes on 12.5% of the trials in which it was used. No such ground strikes were experienced with the virtually-augmented cockpit. In addition, pilots using the conventional cockpit shot down a pre-programmed "friendly" F-15 on 12.5% of their trials. No fratricidal incidents were observed with the use of the virtually-augmented cockpit. Therefore, there appeared to be very meaningful differences between the two cockpit conditions on two performance measures that might reasonably be interpreted as being ultimately based on pilots' SA.

The Global Implicit Measure (GIM) Approach

The experimental analysis and measurement of situation awareness has received a great deal of attention in the human factors literature recently. A recent special issue of *Human Factors* is testament to this increase in research. Many authors have reported on the difficulties associated with the use of traditional SA metrics (e.g., Flach, 1995, Sarter & Woods, 1991, Smith & Hancock, 1995). In addition to the difficulty associated with any particular class of SA metrics (e.g., the intrusiveness of retrospective memory probes), several authors have expressed theoretical concerns about the construct of situation awareness and its measurement in general. For example, in discussing the difficulty in measuring SA in complex, dynamic systems, Sarter and Woods (1991) state: "Attempts to define the critical contents or components of situation awareness in general suffer from the fact that, given the dynamic environment of the flight deck, the relevance of data and events depends on their context...and will therefore vary within and between flights as a function of specific task, the environment, and the tactical objective" (p. 47).

Smith and Hancock (1995) have also recently argued that understanding the context within which situations occur, as well as understanding the goals of the operator, is critical to the examination of SA. They state: "Until an external goal and criteria for achieving it are specified, examination of greater or lesser degrees of SA or even loss of SA remains impossible" (p.139). It has also been argued that laboratory research on SA should be conducted under conditions that afford as much realistic behavior as possible (e.g., Flach, 1995).

To address these concerns, the Global Implicit Measure (GIM) approach (see Vidulich, this volume) has been developed as a new metric for assessing SA. This approach has been designed to incorporate the strengths of performance-based measures while minimizing the weaknesses associated with explicit memory probe techniques. The GIM approach assesses SA by comparing measures of human performance in complex tasks, such as air combat, against previously defined behavioral constraints or "rules of engagement." In our application, measures of pilot-aircraft system performance during air combat simulations are continuously recorded on-line. These

performance measures then serve as the basis for comparing actual performance versus the specified rules of engagement. Our intent is to refine the method to allow on-line calculation of these differences as a means of assessing SA in real-time.

The GIM approach is "implicit" in the sense that SA is probed by continuously assessing the degree to which human performance corresponds to previously defined constraints on functional behavior. Functionally adequate SA is argued to occur when objectively measured aspects of human performance are in correspondence with these constraints. Since it is based on quantifiable measures of human performance, the GIM approach allows for precise measurement. In addition, since the criteria against which performance is assessed using this approach are derived from detailed task analyses, it benefits from high operational validity. The implicit probes are non-intrusive and their associated data are collected without any unusual interruptions, performance of secondary tasks, or additional subject memory load.

The GIM approach is currently being used primarily as a means of examining differences in the level of pilot SA afforded by the two cockpit configurations evaluated by Hettinger et al., (1994). The current evaluation also employs the same simulated air combat mission used by the above authors, as described earlier in this paper.

However, in order to provide a greater level of specificity of the nature of pilot tasks involved in the simulated air intercept mission, a detailed task analysis of the general characteristics of the mission was conducted in association with a subject matter expert. This task analysis provided a detailed level of description of a representative mission broken down into segments. The major segments included a combat air patrol (CAP) phase, an intercept phase, a maneuver phase, a weapons employment phase, a defensive reaction phase, and egress. These segments were further broken down into three major subsections: aviating, navigating, and communicating. Specific mission tasks were then listed under each subheading, and reflect the goals and sub-goals of the pilot relevant to a generic air intercept mission.

This task analysis was then used to create a highly detailed set of rules of engagement, similar in nature to rules of engagement encountered by fighter pilots in actual missions. The rules derived from our task analysis place constraints on the actual engagement of targets and also place constraints on mission performance parameters. For example, the rules of engagement used in the current evaluation place constraints on speed, altitude, course, radar mode and antenna coverage, weapons selection, commit decisions, and a variety of other mission parameters. All of these aspects of pilot-aircraft system performance are recorded on-line for each simulated mission, and are saved for later data analysis. Pilots serving as subjects are informed that adherence to the rules of engagement in the current study is as high a priority as successful completion of the mission (i.e., shooting down the simulated bombers and returning to safe air space). Each simulated mission will then be analyzed on a post-hoc basis in order to determine the level of compliance with the rules of engagement. Pilot SA will then be operationally defined as the degree of correspondence between actual pilot-aircraft system performance and the specified rules of engagement.

The task analysis and the rules of engagement represent one method of segmenting the overall mission into phases, each with unique goals and sub-goals. A transition from the CAP phase to the intercept phase, for example, represents a radical change in the pilot's immediate goals, while the overall goal of bomber interception and return to safe air space remains. Implicit comparisons between actual versus instructed performance will then be used to determine the degree to which the pilot succeeded in adhering to each segment's specified performance constraints, as determined by the rules of engagement. For example, the rules of engagement for a particular segment may specify the radar mode setting the pilot must use (i.e., "track while scan" as opposed to "single target track"). If the pilot sets the radar correctly for that portion of the mission, the resulting score on that implicit metric will reflect adequate SA (i.e., a numerical value of "1"). If the pilot sets the radar to any other mode, the score on that item will reflect inadequate SA (i.e., a numerical value of "0").

Each implicit measure will be scored at a rate equal to the frame rate of the simulation. For a specified time period during the mission (mission phase, or segment within phase), a proportion score will be calculated for each of the implicit probes. This proportion will consist of the sum for

each measure over that time period, divided by the number of observations made during that period. The score on each of the implicit measures during any segment can then be combined with other comparable implicit measures collected within that segment. Since they are all proportion scores, the scores can be combined based on the overall goals and sub-goals within each segment to yield a composite SA score for the entire segment, and an SA profile reflecting different goals within the segment. The different segment scores may then be combined to yield an overall mission SA score. Finally, the scores may be weighted in order to yield a more realistic metric of SA based upon the relative importance of each implicit probe as specified by subject matter experts.

Conclusions

This paper has described a current application of a new approach to the assessment of SA, termed the Global Implicit Measure approach, that is based on comparisons of actual versus instructed performance of complex tasks. The measures yielded by this approach offer a new method of assessing SA in the performance of complex, highly dynamic tasks on a post-hoc basis. Our most immediate application of this methodology will be to assess differences in SA afforded by alternative cockpit designs, particularly those that make use of virtual environment technology to present complex information to pilots in a more intuitive, salient fashion. The optimization of user SA is an important goal of our efforts to develop advanced airborne crewstations. Therefore, the development of valid and reliable metrics for assessing SA that do not interfere with primary task performance is highly relevant to our design efforts.

A future goal of our work with this methodology will be to adapt it for continuous, real-time assessment of SA in order to support the operation of adaptive interfaces. Adaptive interfaces consist of ensembles of controls and displays that are capable of being automatically modified in real-time as a function of fluctuations in external task demands and/or fluctuations in the physiological/psychological status of the user. The ability to continuously monitor pilot SA with a valid and reliable methodology would provide one very useful basis on which to build a logical, adaptive scheme.

References

Flach, J.M. (1995). Situation awareness: Proceed with caution. *Human Factors*, 37(1), pp 149-157.

Hettinger, L. J., Nelson, W. T., Brickman, B. J., Haas, M. W., Roumes, C. (1995). Assessing human performance as a design aid for airborne applications of virtual environment technology. In *Proceedings of the Eighth International Symposium on Aviation Psychology*. Columbus, OH: The Ohio State University

Hettinger, L.J., Nelson, W.T., & Haas, M.W. (1994). Applying virtual environment technology to the design of fighter aircraft cockpits: Pilot performance and situation awareness in a simulated air combat task. In *Proceedings of the Human Factors and Ergonomics Society 38th Annual meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.

Parasuraman, R. (1993). Effects of adaptive function allocation on human performance. In. D.J. Garland and J.A. Wise (Eds.), *Human Factors and Advanced Aviation Technologies*. Embry-Riddle Aeronautical University.

Sarter, N.B., & Woods, D.D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology*, 1(1), pp 45-57.

Smith, K., & Hancock, P.A. (1995). Situation awareness is adaptive, externally directed consciousness. *Human Factors*, 37(1), pp 137-148.
Wilson, G.F., Fullenkamp, S.C., & Davis, I.E. (1990). Physiological measures of pilot and WSO workload during air-to-ground missions. *Aviation, Space, and Environmental Medicine*, 61, pp 454-460.

Displays to Enhance Air Combat Situational Awareness

Gerald P. Chubb

The Ohio State University

Abstract

A brief, nine-month study effort was conducted to develop three advanced air combat display concepts that could be further explored by Armstrong Laboratory in the SIRE facility. The study used three sets of subjects: 1) senior military pilots, 2) Ohio Air National Guard Pilots, and 3) Naval and Air Force Reserve Officer Training Corps (ROTC) cadets.

To evaluate the effectiveness of the proposed concepts, a Baseline was established for the importance and quality of information being displayed in each of three mission types: 1) joint, night interdiction, 2) close air support, and 3) offensive counter air. Subjective workload assessment measures were also taken (using SWAT) for selected events within each scenario. The same measures will be taken to evaluate the proposed advanced display concepts, using computer animation to illustrate how the concepts might be mechanized.

The principal display concept recommended for implementation was a Tactical Situation Display (TSD), showing the "Big Picture" of each combat operation. While the content and format varied in each application, the basic concept was similar: a moving map-like display of the air situation around the particular aircraft. This gave pilots an ego-centered battle management display.

Baseline Study results indicated that the proposed TSD concept should significantly impact the quality of information being displayed in support of each mission type. Other display recommendations included a tactile display of missile launch and an interleaved, periodic, background, three dimensional, audio display of own and enemy aircraft during air-to-air engagements. Since neither of these two display concepts could be implemented in this nine-month study, their effectiveness will need to be evaluated in follow-on research.

Introduction

The Synthetic Immersion Research Environment (SIRE) facility at the Armstrong Laboratory will test display concepts using a cockpit configuration which includes the following displays: 1) a monochrome Heads Up Display (HUD) with a wide field of view, 2) a monochrome Helmet Mounted Display (HMD), which essentially repeats the HUD display, except it allows off-boresight viewing angles, 3) a Heads Level Display (HLD), a color, Multi-Function Display (MFD), and 4) two Heads Down Displays (HDD), both color MFDs.

The present, preliminary study did not call for a complete cockpit layout, but only for three novel display concepts. In order to generate three different concepts, the proposed study examined three particular mission contexts: 1) defensive counter air (DCA), 2) close air support (CAS), and 3) joint, night interdiction (JNI). These were believed to be representative and challenging

missions. The final deliverable was a software functional specification for the proposed display concepts.

The study was conducted in four phases. In the first phase, senior military pilots were interviewed to determine the important elements associated with each of three mission scenarios of interest. Then the three mission scenario scripts were developed with the aid of experienced F-16 pilots for use in subsequent phases. Hypothetical missions were laid out over North Korea, since this area is the only one common to both the Falcon Gold (F-16) and the Strike Eagle (F-15E) simulation packages.

In the second phase, the ROTC pilots were trained in the rudiments of each mission type, preparing them for a brainstorming session. Also, in this phase, F-16 Ohio National Air Guard pilots were used in a Baseline study to determine the information deficiencies of present display systems. Finally, in phase two, brainstorming sessions were also conducted with two groups: 1) relatively naive ROTC subjects, and 2) experienced fighter pilots. In both cases, ideas for new display concepts were solicited.

As a result, three display concepts were developed in phase three and implemented in a computer animation for phase four studies. In phase four, animations of the proposed displays are again being reviewed and evaluated by the same Air Guard F-16 pilots who had participated in the Baseline study.

Method

Except for the empirical Baseline and Evaluation studies using Guard pilots, interview and brainstorming techniques were used.

Subjects

The senior military pilots interviewed in this study were all locally available. One was a test pilot. One was a reconnaissance pilot. The other three were fighter pilots. Two of the fighter pilots, the test pilot, and the recce pilot had all seen combat in Viet Nam. The recce and test pilots had also seen combat in Korea.

The Ohio Air National Guard pilots were selected by their respective Directors of Operations (DO) to participate, knowing that whoever participated in the Baseline study would also have to be available later for the Evaluation study, since subjects were used as their own controls. Four pilots were included from the 178th Fighter Group (FG) at Springfield airport, and six pilots from the 180th FG at the Toledo airport. All were qualified in the F-16C. The 178th had block 30 aircraft, while the 180th had block 40 aircraft. Neither group was experienced in night operations.

The Reserve Officer Training Corps (ROTC) cadets were all volunteers from the Navy and Air Force units at The Ohio State University (OSU). Initial volunteers were later self-screened by availability for the brainstorming session. Class and assignment conflicts eliminated some participants. Seven ROTC cadets participated in one of two brainstorming sessions.

Apparatus

Interviews with senior military officers were audio-recorded with a small, pocket recorder using mini-cassettes. The mini-cassette data were later transcribed and edited for review and approval by the interviewees.

An audio tape of the mission scenario was developed to help pilots get a feel for the pacing of the mission events and aid their visualization of activities occurring in the mission. Baseline

studies of the F-16 displays were then video taped. The Evaluation study used computer animation of the proposed advanced display concepts, and a Pentium computer presented these in conjunction with a digitized version of the same mission audio used in the Baseline study. Again, the Evaluation study sessions were video taped.

Both brainstorming sessions were video taped. In preparation for the brainstorming sessions, ROTC cadets were given formal instruction on the basic elements of flight operation associated with the three mission types of interest. To do this, Falcon Gold (F-16) and Strike Eagle (F-15E) software packages were used on a 66 MHz Pentium using Thrustmaster throttle and control sticks. An experienced fighter / instructor pilot monitored the indoctrination, having trained three student assistants who then tutored the cadets becoming familiar with basic combat operation tactics and procedures.

Procedure

Interviews with Senior Military Pilots

Interviews with senior military pilots were constructed using a structured interview which lasted between one and two hours, depending on the subject's interest in elaborating upon the answers to those questions. Transcribed and edited interviews were then given for review. Any corrections offered were incorporated in the final transcripts. Editing simply eliminated certain redundancies and made verbalized statements more readable.

Brainstorming Sessions

The brainstorming sessions both began by reviewing the standard rules for such sessions: ideas are to build upon each other without any assessment of their pros and cons. Suggestions for creative reformulation of ideas were also offered to participants as part of this introduction. Each of the mission scenarios was then reviewed to set the context for the display-concept brainstorming. Results were independently documented by two observers and then consolidated. Display concepts were then developed based upon the remarks solicited in the three brainstorming sessions (two with ROTC cadets and one with experienced pilots).

Baseline Studies

The baseline studies were conducted on-site at the 178th and 180th FG facilities. Pilot briefing rooms were used, so pilots could refer to standard visual aids of the existing cockpit displays. Sessions took approximately four hours each. The session began with an explanation of the purpose of the study and the rank ordering of the SWAT cards, a necessary prerequisite for using SWAT (Reid, et al., 1982), which itself takes about 30-45 minutes. After a break, each of the three mission scenarios was "run" using the audio tape that emulated radio traffic which would occur during the course of the mission.

Each mission took slightly less than twenty minutes to execute. Subjects were encouraged to draw a picture of the display format they would be using. This forced them to identify, from memory, the significant features of that display. Directly after the tape was played, subjects made their SWAT ratings, based on a prescribed list of events which had occurred within the scenario. A separate rating was given for each of these events. A dual-scale Situational Assessment rating was then given for various kinds of information related to mission execution.

After the fashion of Endsley (1993), the first scale rated the importance of the information, as not important, important, or very important (intermediate values were allowed in cases where the subject had difficulties deciding). The second scale rated how well the current aircraft displays provided the information: not adequate, adequate, and well-done. The rating of well-done was to be interpreted as an indication the display should be left "as is." Again, intermediate scale ratings were allowed. Also, radio communication was interpreted as an "inadequate" display of information, since pilots complain too much is being left for verbal reporting. After each mission, pilots took a short break and then repeated this procedure for the next mission type.

Evaluation Studies
The evaluation studies are also being performed on-site at each unit and will use the same pilots as the Baseline study. The basic procedures will be the same. In this instance, pilots will watch a computer animation of the proposed TSD concept as they listen to the same audio script of the mission. The TSD concept incorporated a synthetic, moving map display and symbolic aircraft, as well as ground position indications. The Digital Chart of the World (DCW) was used as the database for this display (Omara, 1995), a publicly available form of Defense Mapping Agency (DMA) level 0 digital terrain elevation data (DTED). It is obtainable on CD-ROM from the U. S. Geological Survey for $200 and provides vector formatted terrain data at a 1:1,000,000 scale, equivalent to an Operational Navigation Chart (ONC).

SWAT and modified SAGAT ratings (Vidulich, et al., 1994) were again obtained after each simulated mission. These data were then compared to the scores obtained in the Baseline studies at each base for these same pilots.

Results

The Senior Military Interviews
The senior military interviews identified major decisions related to mission success (Chubb, 1995). Clearly identification is a key element in every case, whether air-to-air identification or target recognition in the air-to-ground case. Air-to-air situations roughly separate into not only beyond and with-in visual range (BVR vs. WVR) but before and after merge: when opposing aircraft pass each other and then get tangled up in the proverbial "fur-ball" or what the Navy calls Close-in Combat (CIC).

The close air support mission requires coordination with a Forward Air Controller (FAC), airborne or on the ground, who provides essential updates to the situation assessment of a fluid battle-line. Major improvements in communicating the standard "nine-lines" of tactical information can be proposed, but the key to improvement lies principally with the FAC side of the exchange.

Joint, night interdiction is dominated by a concern for being at the right place at the right time, so situation awareness in this context emphasizes knowing where everyone is relative to pre-planned spatio-temporal event sequences. The principal concern is penetrating enemy defenses and then finding a safe egress route, often to a refueling tanker before returning home. The importance of defense suppression in such joint operations is an increasing concern, as is search and rescue, elements we did not examine explicitly in this study.

Brainstorming Sessions
The brainstorming sessions conformed to the expectation that experienced pilots are better able to make incremental improvements than radical new departures in display concepts. The inexperienced participants were better able to offer "blue sky" ideas, unconstrained by what appeared practical, feasible, or useful. The combination of inputs proved to be useful, since neither group had all of the ideas in-hand. However, both saw the utility of the TSD format which is central to the proposed advanced display concept developed here. This is certainly not a new idea, since the F-15 JTIDS implementation is essentially the contemporary operational version of this concept.

Baseline Studies
The baseline studies provided data which indicated SWAT scores were dominated by the intensity of combat operations themselves. It was not expected that these data would be a sensitive indication of display improvement or its impact on workload, since other factors (like terminal threat encounters) appear to be strong drivers of the SWAT scores. However, indications showed there were ample opportunities to make improvements to the displays, although pilots were in

general satisfied with the F-16 display suite. Most of the information items selected for rating were assessed as important. However, a few items expected to be important (from the literature: like sun angle) proved to be unimportant, at least in a visual display (probably because this information is self-evident when relevant).

Evaluation Studies
The evaluation studies are expected to show that the TSD did in fact significantly alter the quality of displayed information, at least in several key areas. The SWAT scores not expected to evidence any significant change in workload as a result of improvements in situational awareness gains.

Conclusions

A tactical situation display in a moving map format with variable range scales appears to be a useful means for enhancing situational awareness in a variety of mission types. Pilot's desire for all the information they can get, without being overloaded, clearly presents a challenge to understand processing limitations. However, display content and format appears to be driven not only by task within mission phase (and type) but by experience level as well. What competent pilots find cluttered, expert pilots find informative. Conditionally dependent, overlaid, and suitably tailored displays are the next challenge for improving performance effectiveness and mission success.

References

Chubb, Gerald P. (1995), User-centered, object-oriented expertise approach to advanced air combat display design: Phase 1 Interim Report, Ohio State University, Columbus, OH.

Endsley, Mica R. (1993), Survey of situation awareness requirements in air-to-air combat fighters, *The International Journal of Aviation Psychology, 3(2),* 157-168.

Omara, Raymond E. (1995), Digitizing the future defense mapping agency, Fairfax, VA.

Vidulich, Michael, Cynthia Dominguez, Eric Vogel, and Grant McMillan (1994), *Situation awareness: papers and annotated bibliography,* AL/CF-TR-1994-0085, Crew Systems Directorate, Human Engineering Division, Air Force Materiel Command, Wright-Patterson Air Force Base, OH.

Reid, G. B., F. T. Eggemeier, and C. A. Shingledecker (1982), Subjective Workload Assessment Technique, *Proceedings of the AIAA Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics,* Edwards AFB, CA, 281-288.

Measuring Situational Awareness with the "Ideal Observer"[1]

Marc Green[2,3], J. Vernon Odom[2], and J. Terry Yates[4]

[2] University of West Virginia Medical School
[3] York University
[4] Brooks Air Force Base

Background

Situational awareness requires the human operator to quickly detect, integrate and interpret data gathered from the environment. In many real-world conditions, situational awareness is hampered by two factors. First, the data may be spread throughout the visual field. Operators are then limited by attention, memory and ability to combine data seen in the same or different formats. Second, the data are frequently noisy.

We have been investigated application of the "Ideal Observer" model as a means for measuring situational awareness under these conditions. According to the model, task performance is limited only by three factors, "external noise," noise in the environmental data, "internal noise," noise inside the observer and "efficiency," the ability to sample environmental information. The model makes precise predictions about the relationships among these 3 variables.

There are many advantages in using an Ideal Observer model to measure situational awareness. The Ideal observer provides a measure of statistically optimal performance, so that situational awareness can be evaluated against an absolute rather than a relative standard. Moreover, the Ideal Observer model allows independent analysis of both low level information processing and high level cognitive decision-making within the same framework. Further, the Ideal Observer permits direct comparison of optimal behavior across different environments - it is possible to compare situational awareness with vastly different displays and tasks. There are several other benefits of the Ideal Observer, but these will not be obvious until the model is explained below.

We are using the Ideal Observer in developing a battery of tests which evaluate the ability to make decisions when confronted with noisy data. The displays contain information, perturbed by Gaussian noise, which is spread throughout the visual field. The observer must integrate the noisy data and then make a decision. We are investigating factors which minimize observer and decision noise and maximize efficiency.

In this paper, we will describe a preliminary application of the Ideal Observer to a task from our test battery, dot estimation. We have chosen this task both because the application of the Ideal Observer model is simple and direct and because several studies (Endsley and Bolstad, 1994; T. Caretta, cited in Endsley and Bolstad, 1944) have shown that dot estimation performance correlates well with other measures of situational awareness.

The derivation of an Ideal Observer for dot estimation has been described elsewhere (Barlow, 1978; Burgess and Barlow, 1983, but, for convenience, we will provide a brief review and highlight the significance for measurement of situational awareness.

Derivation of the Ideal Observer for Dot Estimation

In Signal Detection theory, performance is limited only by noise. Given this view and a statistical representation of environmental information, it is possible to construct a model, the Ideal Observer, which has no internal noise and uses all available signal information. Such a hypothetical observer performs optimally in the sense that it is limited only by the information in the environment.

In the dot estimation task, the observer sees two boxes, one contains N dots and the other $N + \Delta N$ (Figure 2). The task is to say which box has the ΔN. Optimal performance can be expressed as:

$$d' = \frac{\Delta N}{\sigma_N}$$

where d' is the measure of detectability, ΔN is the difference in dot number between target and background (noise) and σ_N is the noise standard deviation – a number of extra dots which have been added to or subtracted from each box.

Real observers, however, seldom achieve ideal performance. Previous work has suggested two general factors to account for the suboptimality of human performance. The first factor producing suboptimal behavior is internal noise. Although the Ideal Observer is presumably limited only by external noise, real detection devices also have *internal* noise which decreases performance. In the dot estimation task, this can be modelled as:

$$d' = \frac{\Delta_N}{\sqrt{\sigma_N^2 + \sigma_N^2}}$$

where σ^2_N refers to observer's internal noise. (Of course, this assumes independence of internal and external noise.) In other words, detection is a joint function of external and internal noise, which increases the denominator and decreases performance.

The second factor affecting performance is reduced information gathering efficiency. The Ideal Observer is a Bayesian classifier which determines p(hypothesisldata) for each potential signal. It computes a likelihood ratio for any two signals from the ratio of their probabilities. A decision rule uses the likelihood ratio or its monotonic transform to specify when the observer should say "yes" or "no." The response transition point, the criterion, depends on what aspect of performance the observer wishes to optimize). For most applications, this is assumed to be maximum percent correct.

The main problem, as in all Bayesian classification tasks, is to derive an estimate of p(datalhypothesis). The Ideal Observer obtains p(datalhypothesis) by sampling the data in the environment and comparing them to an internal model. The importance of proper sampling is apparent from Barlow's (1978) definition of sampling efficiency, F, as:

$$\frac{\textit{sample size required by the ideal observer}}{\textit{sample size required by the actual observer}}$$

where F=1 is optimal efficiency and lower values represent suboptimal sampling. The less efficient observer will require more information samples, meaning a longer sampling period, and therefore longer reaction time, or better information. This notion of efficiency reflects how well an observer uses external information. The observer samples visual input and tries to match this information to

an internal model, much like a template matching procedure. In the broad application to situational awareness, the templates are presumably a patterns of information which reflect possible states of the aircraft (or any other dynamic system). Efficiency reflects the observers ability to use external information to detect the appropriate "template."

The F value has two practical uses. One is that it can be used to discriminate people who have high and low ability at accepting and integrating visual information. Another is to use F as a tool for designing information displays. That is. when comparing different display formats, high F also suggests highly efficient information display.

F can be estimated directly from sensitivity as:

$$F = (d'_e / d'_i)^2$$

where d'_e is experimentally obtained sensitivity and d'_i is the sensitivity of an Ideal Observer who uses all available information. Finally, by substitution and solving for a d'_e value of 1, the final model for a real observer, taking into account efficiency and internal noise, becomes:

$$\Delta N_T^2 = (1 / F) (\sigma_N^2 + \sigma_N^2)$$

where ΔN^2_T is the square of the difference in dot number required to produce a d'_e of one - 76% correct in a two alternate forced-choice test. (If this step was too large, see Burgess and Barlow, 1983 for more details.) In English, this says that performance is a joint function of two additive factors, internal and external noise, and a multiplicative factor, efficiency[1].

One way to visualize the relationship between internal noise and efficiency is to plot data from a test which measures the square of the of the threshold (dot difference needed to achieve a performance of d'=1) as a function of external noise variance (dots added/subtracted from the display). Figure 1 shows hypothetical results for such an experiment. Internal noise is a constant factor which alters the X intercept (not shown) while efficiency is a multiplicative term which alters slope. The Ideal Observer, A, who has an intercept of zero (is noiseless) and a slope of one (uses all information). C shows a family of observer with internal noise (change in intercept) but efficiency close to the optimum (slope of 1). The internal noise produces a horizontal shift which can be quantified by the negative of the curve's intersection with the abscissa. Slope increases reflect suboptimal efficiency. Curve B shows an observer with no internal noise but a lowered efficiency, which is quantified as the reciprocal of the slope. Of course, an observer could exhibit both lowered efficiency and internal noise.

There are two important points highlighted by this graph. First, the Ideal Observer provides two different measures, efficiency and internal noise level, for each test. At 0, or any single external noise level, a given level of performance could be caused by different combinations of the two factors. Two observers could perform equally on the task, yet one might have low efficiency and the other high internal noise. Situational awareness might correlate with only one of these two factors. To distinguish the two factors, observers must be tested at a minimum of two noise levels. Once done, however, the Ideal Observer models allows a more precise and detailed analysis of individual abilities because it reveals not just overall performance but also individual differences in the factors underlying task performance. Although situational awareness correlates moderately with dot estimation, for example, there could be a higher correlation with observer efficiency and a lower correlation with internal noise. Looking only at overall performance would then reduce the correlation.

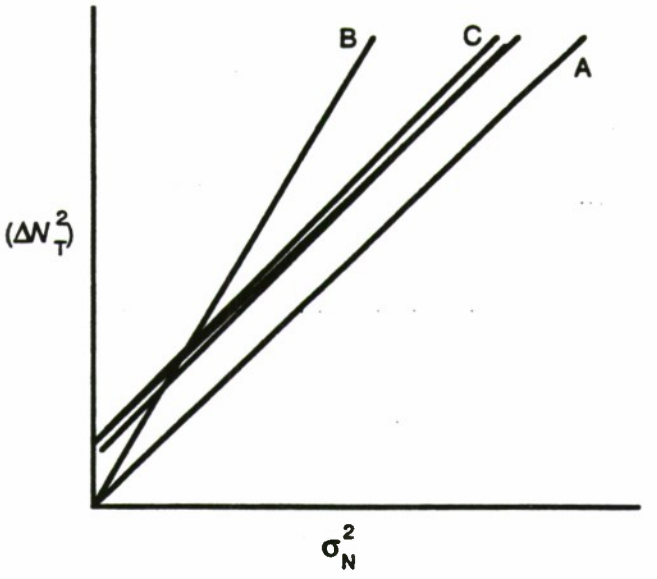


Figure 1. Schematic data from Ideal Observer test.

Second, note that in the low external noise conditions, high and low efficiency observers will be less readily differentiated. At high external noise levels, however, low and high efficiency observers are more easily distinguished. Addition of noise magnifies individual differences in some observers, making it easier to discern those with high situational awareness.

The Dot Estimation Test

Method

In our version of the task, the viewer sees two red rectangular boxes (Figure 2) containing differing numbers of black dots on a grey background. Following each 667 msec exposure, the observer responded by pressing the left or right mouse button to signal whether the left or right box had more dots. Observers were tested in a series of two-alternative spatial, forced-choice trials in which task difficulty, the difference in the number of dots in the two rectangles, was modulated by a tracking rule. The standard number of dots (N) was 100. The dot difference (ΔN) between the boxes was then perturbed by adding "noise" (σ_N), i. e., increasing/decreasing dots from each box. The number of noise nots was randomly chosen from a Gaussian distribution with a mean of 0 and a variance of 0, 25, 100 or 400 dots.

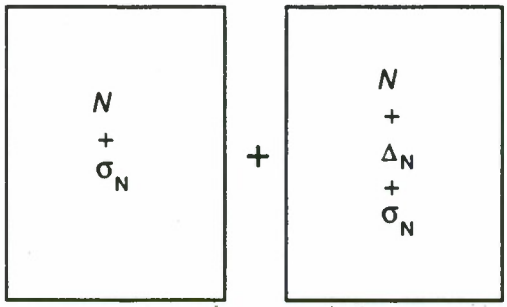


Figure 2. The right hand box contains the signal Δ_N

Results

Table 1 and Figure 3 show the results for three observers. Each observer shows both the presence of internal noise and a slope greater than one indicating suboptimal efficiency. Different observers, however, exhibited different degrees of noise and efficiency. Observer 3 had the lowest efficiency, for example, while observer 2 had highest efficiency but greatest internal noise.

Figure 3 shows the results plotted as a function of the squared dot threshold variance. and external noise The regression lines through the points were fit by least-squares method. For comparison, we also show the predicted performance of the ideal observer. The points cluster close to the regression line, showing that the data are in good agreement with the model for observers with internal noise and suboptimal efficiency.

Table 1.

Observer	X Intercept	Slope
1	290.5	1.77
2	343.7	1.82
3	197.5	2.12
Mean	273.1	1.90

Conclusion

Our goal has been to both describe the advantages of Ideal Observer analysis and to demonstrate its application to a test which is known to correlate with situational awareness. By decomposing test scores into subcomponents of efficiency and internal noise, Ideal Observer models may provide a finer grained analysis of individual skills and capacities and reveal more fundamental components of high situational awareness ability. Moreover, by stressing visual performance with noise, individual differences in ability should become more apparent.

We do not claim that the particular task demonstrated here, dot estimation, would alone be sufficient to measure situational awareness. A complete test battery would be comprised of several perceptual and attentional tests. The ideal Observer, however, can be applied to virtually any test where it is possible to create a statistical description of the task and to detection, reaction time and even physiologically derived dependent measures. We have already applied the Ideal observer to a

large battery of behavioral and physiological tests which are being used to evaluate pilot visual abilities.

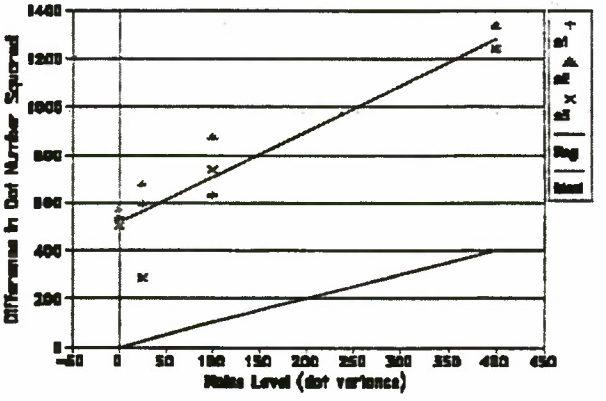


Figure 3.

References

Barlow, H. (1978) The efficiency of detecting changes of density of random dot patterns. *Vision Research, 18*, 637-650.

Burgess, A. and Barlow, H. (1983) The precision of numerosity discrimination in arrays of random dots. *Vision Research, 23*, 811-820

Endsley, M.R. and Bolstad, C.A. (1994). Individual differences in pilot situation awareness. *International Journal of Aviation Psychology, 4*, 241- 264.

Hartman, B and Secrest, G (1991). Situational awareness is more than exceptional vision. *Aviation, Space and Environmental Medicine*, November, 1094-1089

Secrest, G. and Hartman, B. (1993). Situational awareness: The trainability of near- threshold information acquisition dimension. *Aviation, Space and Environmental Medicine*, October, 885-889.

Footnotes

[1]Funded by contract F41624-92-DS-4001 DO-0017 from the U. S. Air Force Human Systems Center.

[5]There is another multiplicative parameter, decision noise, which we have not explicitly modelled. Because it would require far more involved testing procedures, we have chosen, as most researchers do, to collapse it with efficiency in the F parameter.

Towards a Robust, Quantitative Measure for Presence

Jerrold D. Prothero, Donald E. Parker, Thomas A. Furness III, and Maxwell J. Wells

University of Washington

It is argued that an understanding of presence is necessary for an understanding of situation awareness. A model is described which combines presence and vection within a single framework, and previous experiments based on this model are briefly covered. The model is the basis for a planned series of experiments designed to develop an objective measure for presence based on the degree of identification with virtual over conflicting real cues. A possible role of adaptation in situation awareness, presence and vection is considered.

Introduction

"Presence" and "situation awareness" are overlapping constructs. Presence implies that observers perceive their self-orientation and self-location with respect to an environment. Although it includes additional complex aspects of human performance, adequate situation awareness presupposes appropriate environmental orientation. Consequently, development of robust, quantitative measures for presence is of fundamental importance for research on situation awareness.

In presence research, as in the field of situation awareness, the search for relevant measures is a crucial topic. Ideally, one would like to think of the virtual environments industry as being in the business of "presence engineering": i.e., systematically inducing a sense of presence in particular virtual environments. But before one can do engineering, one needs an underlying science; in this case, a "presence science" which explains the origin and nature of presence and the factors on which it depends. Developing a science in turn depends on reliable measures which will allow one to do the experiments to build the science; hence, it is important to develop a reliable measure for presence. The same motivation drives the search for measures of situation awareness which has led to this conference.

For a measure to be useful, it must be closely tied to a successful scientific theory: otherwise one will be mainly measuring noise. Since, in our case, the measure is also necessary to develop a theory of presence, there is an unfortunate circularity. We resolve this circularity by suggesting a process of convergence: we begin with a rough theory and a measure based on this theory, check this measure against existing validated measures for presence, and refine the measure and the theory in parallel. In the end, we hope to have a measure which is consistent with but more sensitive than existing measures, which can be used as the basis for developing a refined theory.

It seems best to begin with the simplest possible theory, since its simplicity makes it easiest to refute. Furthermore, one can add complexity to a simple theory if it is shown to be wrong; if one begins with a complex theory the next step is less clear. In the case of presence, the simplest theory would seem to be a direct formalization of the idea that presence has to do with "being in" a virtual environment. The formalization is that presence is an illusion of position and orientation:

orientation, from using cues defined by the real environment to using cues defined by the virtual environment.

Given the idea that presence is an illusion of position and orientation, it is natural to ask whether it is related to vection (visually-induced illusory self-motion). Such a connection would be very useful, since it would allow results from the vection literature to be applied to presence. In an earlier paper (Prothero *et al.*, 1995) we introduced the "presence rest frame hypothesis" as a possible link between presence and vection. Briefly, this hypothesis states that we maintain a subjective coordinate frame with respect to which we determine positions, orientations and motions. Disturbances to this rest frame may result in either illusory motion (vection) or illusory position and orientation (presence). If presence and vection are as closely related as this implies, we would expect presence to depend on the same factors as determine vection. Recent research, summarized in Prothero, *et al.* (1995), suggests that vection is heavily influenced by one's apparent relative motion with respect to what one takes to be the background. A previous experiment in the vection literature reported changes in level of perceived vection by manipulating perceived background, for constant field-of-view (Mergner & Becker, 1990). We repeated this experiment for level of perceived presence, finding a similar relationship. This provides initial (although far from conclusive) support for the rest frame hypothesis.

Prothero *et al.* (1995) reported an initial test of the link between presence and vection. We propose here a test of another tenet of the rest frame hypothesis, namely that presence is an illusion of position and orientation. This test, if successful, will also provide a possible objective measure for presence. The basis of the test is as follows. If presence has to do with a switch in the cues used to define one's position and orientation, from cues provided by the real environment to cues provided by the virtual environment, then the level of presence should correspond to the level of identification with virtual cues over real cues. We can therefore set up an experiment in which participants are asked to null conflicting virtual with real cues. We would predict from the rest frame hypothesis that the degree to which the virtual cues dominate the real cues should be related to existing subjective measures of presence. We give the details of the planned experiments after reviewing the literature on measures for vection and presence.

Method

As described in the Introduction, our research is based on the hypothesis that vection and presence are closely related. Since vection has been studied longer and more thoroughly, we begin with a summary of the relevant vection literature.

Vection

"Vection" refers to a sense of self-motion induced by visual cues. Vection can be either angular or linear. To induce angular vection subjects are seated in a chair surrounded by a cylinder (often painted with stripes) which rotates around the subject. Linear vection is typically induced by a display in which objects seem to be approaching or receding.

The literature on measures for vection is summarized in Carpenter-Smith *et al.* (1995). Most previous vection studies have been based on a measure of magnitude estimation, in which a subject is requested to assign numbers or joystick positions to perceptions. Magnitude estimation is problematic due to issues such as adaptation to the stimulus, differences in subjective scales between subjects, and "range effects".

A more desirable measure is one in which subjects make a comparison between stimuli, rather than comparing a stimulus to a percept. This eliminates errors due to subjective interpretation and estimation. Measures of this type constitute the so-called "Class A" observations (Brindley, 1970).

Work on Class A measures can be thought of as dividing into threshold (Young *et al.*, 1973; Berthoz *et al.*, 1975) and nulling (Zacharias & Young, 1981; Huang & Young 1981; Huang & Young 1987; Huang & Young 1988) studies. In threshold studies, one looks at how visual stimuli affect the minimally detectable magnitude of inertial motion (or conversely, how inertial motion affects the onset of vection). Young *et al.* (1973) looked at the interaction of visual and vestibular rotation cues, by placing subjects on a rotatable chair surrounded by a stripe pattern rotating at constant angular velocity. Among their findings are higher thresholds for the detection of inertial acceleration when the inertial cues conflict with the vection cues. Berthoz *et al.* (1975) placed subjects in a cart which moved linearly and induced vection by providing moving images in the lateral visual field. They reported vection thresholds in the range of image motion detection, and dominance of visual over conflicting inertial cues. A disadvantage of threshold studies is possible variance due to subjects assuming different confidence criteria for threshold.

In nulling studies, one set of stimuli are opposed by another and subjects are asked to determine the point at which the two stimuli counterbalance each other. Zacharias and Young (1981) set up a circular vection nulling experiment similar to the one we propose. Subjects were asked to maintain a stationary position by adjusting their inertial rotation, in the presence of a rotating visual surround and inertial disturbance. Other research using visual-vestibular nulling is described by Huang and Young, for yaw rotation (Huang & Young, 1981), lateral motion (Huang & Young, 1987), and roll and pitch (Huang & Young, 1988). These papers developed models for visual-vestibular interaction.

Related but distinct is the research by Carpenter-Smith *et al.* (1995). Prone subjects were translated along their head x-axis (fore-aft). In the presence of various inertial and visual surround conditions, subjects were asked to report their direction of motion. By running many trials for each subject in each condition, a point of subjective equality (PSE), at which subjects would think themselves at rest, could be determined mathematically for each condition. Shifts in PSE as a result of changes in the visual surround were used to develop, for the first time, a scale for linear vection. (This is not to say that a scale for vection cannot be developed using the more traditional nulling techniques, as we intend to do in our research.)

The above research indicates that a Class A measure for vection is possible. It says nothing about whether such a measure will also work for presence or whether there might be a relationship between Class A measures for vection and presence.

Presence

The literature on the psychology of presence is not as well developed as that of vection. A summary can be found in Hendrix (1994).

Presence frequently becomes an issue because it is thought to correlate with improved task performance in virtual environments. One approach to finding objective measures for presence is therefore to look at task performance. In the possibly-related area of mental workload, Jex (1988) divides objective measures into task demands, task results, and correlated measures (the latter referring to things which might be correlates of the phenomenon of interest, such as heart-rate or muscle tension).

Physiological measures for presence are in principle very attractive, as at least some of them can be recorded fairly unobtrusively as the subject is participating in the virtual environment, potentially allowing for a real-time response to the subject's level of presence. A list of possible physiological measures for presence is given by Barfield & Weghorst (1993), and includes posture, muscle tension, and cardiovascular and ocular responses to virtual events. Neurological measures might also be considered.

Unfortunately, there is currently no evidence that physiological measures correlate well with presence. Sheridan (1992) (p. 209) mentions physiological measures, stating that "It is natural to seek an objective measure or criterion that can be used to say that telepresence or virtual presence have been achieved. However, telepresence (or virtual presence) is a subjective sensation, much

like mental workload, and it is a mental model--- it is not so amenable to objective physiological definition and measurement."

One can also apply subjective measures to presence. For instance, Slater & Usoh (1993) recommend asking whether "the person can later report the sense of having been somewhere other than where they really were at the time."

Other possible measures proposed by Barfield & Weghorst (1993) are examining virtual world task performance and natural world task performance, frame of reference conflict resolution (if the virtual world and the real world conflict, how does the subject resolve the conflict), and context reorientation time or degree of disorientation when moving between virtual and real worlds.

Held & Durlach (1991) suggest a measure for presence based on the ability of an environment to produce a "startle response" to unexpected stimuli. More generally, can we get people to respond to a virtual environment in a way which would only make sense if they interpreted it as the real environment?

An example of a Class A measure for "simulation fidelity" (which seems closely related to what we call presence) is discussed by Nemire *et al.* (1994). They report an experiment to measure "simulation fidelity" in terms of the ability of a virtual environment to induce a change in perception of gravity-referenced eye level (GREL). A pitched optical array can bias a subject's estimate of eye-level. Nemire *et al.* (1994) report that a physical array biases GREL more than an identical virtual array. However, the addition of longitudinal (into the distance) lines to the virtual array removed the performance difference.

An earlier study by Hatada *et al.* (1980) looked at the ability of a display to induce a change in perceived vertical as a function of horizontal field-of-view and content.

Planned Experiments

As with the Nemire *et al.* (1994) and Hatada *et al.* (1980) studies, we propose a measure for presence based on conflicting virtual and real cues. However, our measure is based on conflicting inertial and visual yaw oscillations at close to the detection threshold. We think these conditions are worth investigating for two reasons. The first is that oscillations avoid the adaptation problems of constant stimuli. The second is the hope that supra-threshold stimuli will produce a stronger signal, since there will be less need to overcome conscious expectations. The idea behind the planned experiments is to set up a situation in which both virtual (visual) and real (inertial) sinusoidal oscillations are present with the same frequency. Participants will be asked to adjust the inertial amplitude so that they believe they are at rest. To the extent that the visual cues are sufficiently compelling to induce vection, we expect that participants will pick a non-zero amplitude for the inertial oscillation to counteract their perceived motion due to vection. It is hoped that the magnitude of inertial oscillation choosen can be used as an objective measure for presence. (For technical reasons, in the actual experiment we plan to keep the inertial cues constant and vary the visuals, but the principle is the same.)

An inertial rotation will be provided by a chair mounted on a Contraves Goertz Corporation Direct Drive Rate Table Series 800, under the control of a Neuro Kinetics, Inc. Motion Simulator Controller. The virtual environment will be a scene displayed in a Division PROVision 100 with a Division dVisor HMD. No tracking will be used. The virtual environment will be oscillated sinusoidally. The same oscillation frequency (possibly with a different phase) will be used to drive the rate table. Participants will be strapped into the chair, and will have their heads restrained to avoid head rotations.

Three experiments are planned. The first experiment will determine whether it is possible to induce yaw-rotation vection using available virtual environments apparatus, and to adjust for phase differences between the human visual and inertial systems. The second and third experiments develop an objective measure based on the participant's perceived yaw-rotation. The basis for Experiments 2 and 3 will be the background manipulation discussed previously (Prothero *et al.*, 1995) which has been shown, using subjective measures, to affect levels of both vection and presence. Experiment 3 will be a replication of Experiment 2, except that a previously-validated

subjective questionnaire will be used as a measure for presence. This will allow us to look for a relationship (predicted by the rest frame hypothesis) between vection, subjective presence, and the objective nulling measure.

A theoretical model of cupula dynamics derived from the torsion-pendulum equation indicates that over the range of 0.1 Hz to 5.0 Hz the detection of inertial oscillations should have a gain of very close to 1 and very little phase offset (Howard, 1986). (This frequency range corresponds to natural head motions.) In this frequency range, therefore, we would expect that participants asked to adjust the phase difference between the visual and inertial cues to minimize apparent peak-to-peak displacement would pick a relative phase angle close to zero. (E.g., a visual stimulus moving to the right should induce vection to the left, causing the participant to favor an inertial motion to the right to counteract the vection cues.) Unfortunately, a second constraint is that we need to use frequencies at which both the visual and inertial cues contribute roughly equally to the sense of motion. The gain for the visual and inertial systems are equal at about .02 Hz, with the inertial system having higher gain above this frequency (Zacharias & Young, 1981). While we don't expect to use such low frequencies (.02 Hz is 50 seconds/cycle) we will probably be in the range .02 Hz to .2 Hz. In this range, we have to adjust for phase lead by the human inertial detection system. This is the point of the first experiment. The phase angles from the first experiment will be used in the subsequent experiments. If no consistent phase relations can be achieved, this would indicate that the visual stimuli are not sufficiently cogent to produce a vection effect. It is anticipated that we will use angular velocities of about 1-2 degrees/sec, as the vestibular detection threshold is in this range (Benson *et al.*, 1989). Using the phase angles from Experiment 1, the second and third experiments will ask participants to adjust amplitude to minimize apparent peak-to-peak motion.

As mentioned previously, the gain for the vestibular system is larger than the gain for the visual system for frequencies above approximately .02 Hz. Therefore, we expect that it will be increasingly difficult for visual cues to "beat" the inertial cues at higher frequencies. This implies that smaller inertial amplitudes will be necessary to null visual oscillations at higher frequencies.

We will look for a relationship between subjective presence level and the objective nulling measure using correlations (Anastasi, 1988). This study should replicate the Prothero *et al.* (1995) results on the ability of foreground/background manipulations to affect presence. If the correlations between self-rated presence and the nulling measure is high, this study would support the rest frame hypothesis, the use of rotation nulling as a measure for presence, and the thesis that foreground-background manipulations influence fundamental processes related to presence.

Discussion

We have outlined a series of experiments which we hope will establish an objective measure for presence and we have argued that such a measure is necessary for the development of a science of presence. In this section we speculate on what form a science of presence might take and its possible relationship to situation awareness.

We have suggested that presence and vection are closely related and that they have to do with maintaining a subjective rest frame with respect to which position, orientation and motion are determined. Further, prior research has suggested that both presence and vection are heavily influenced by one's relation to the perceived background. Why would background play a role? An intriguing possibility is that the subjective rest frame is set by the perceived background, the most stationary stimulus available.

This possibility bears an interesting likeness to adaptation. Adaptation is a general property of the nervous system in which constant stimuli become less likely to be perceived. It is characterized by "after-effects": adaptive changes which persist for a time after the stimulus which induced the adaptation is removed. Adaptation is often attributed to fatigue of the sensory receptors, although

an information-filtering interpretation seems more compelling and has a long history (Helson, 1964). In the information-filtering viewpoint, adaptation has to do with moving one's points of minimum and maximum sensitivity to reduce redundant information and increase sensitivity to rarer information. According to Helson (1964), adaptation both reduces the response to dominant stimulation and heightens response to complementary stimulation. One example given is that decreased sensitivity to a stimulus color is coupled with increased sensitivity to the complementary color. Helson (1964) postulates that "all responses can be viewed as positive or negative gradients from equilibrium conditions", and extends adaptation to deal with perception, affectivity and motivation, learning and performance, cognition and thinking, personality, and interpersonal behavior. An example of more recent research along these lines suggests that spatial frequency adaptation should be interpreted in terms of contrast gain control, rather than neuronal fatigue (Wilson & Humanski, 1993). It may also be useful to distinguish between "channel adaptation" (changes in the sensitivity of a sensory channel) and "sensory-motor adaptation" (changes in a model of expected consequences). An example of the latter is alteration of the vestibulo-ocular reflex as a result of a systematic change in visual-vestibular relationships when wearing prism glasses (e.g., Gonshor & Jones, 1976).

Perhaps presence and vection result from a form of what one might call "spatial adaptation" in which the rest frame "adapts" to the constant stimulus of the background?

Evolution works under a severe constraint, which is that at every stage it has to have an organism capable of surviving and reproducing in a competitive environment. This constraint has been compared to the (easier) problem of converting the Wright brothers' plane into the Space Shuttle in such a way as to have something capable of flying after every change. It forces evolution to favor conservative, incremental improvements over radical redesigns. For instance, aside from evolutionary conservatism there is no obvious reason why fish, reptiles, birds and mammals should all have the same basic body plan.

Sensory adaptation is presumably the first form of information filtering which nature discovered. Given evolution's necessary conservatism, it is reasonable that adaptation would have been carried over into the spatial awareness tasks of which presence and vection are derivatives. The more interesting question is whether adaptation was also carried into situation awareness. This possibility is plausible for two reasons. The first is that situation awareness depends on presence which (we suggest) is related to a form of spatial adaptation. The second (and more fundamental) reason is that situation awareness depends heavily on information filtering: extracting a useful mental model from a wealth of mostly useless information.

Given a relationship between situation awareness and adaptation, what form might that relationship take and can any use be derived from it? The most obvious possibility is that we adapt to a mental model of our situation in a way similar to that in which we adapt to a steady sensory stimulus, with enhanced sensitivity to variations from what we expect based on the mental model. A line of research suggested by an adaptation view of situation awareness is to look for adaptation after-effects: changes in threshold detection or time-to-detect performance depending on how closely something is related to one's mental model of a situation. This might provide an experimental technique for extracting the mental models which underlie situation awareness.

Acknowledgements

This research is supported in part by the Air Force Office of Scientific Research (contract #92-NL-225 and INST PROP NO:78216) and the National Aeronautics and Space Administration Grant NAS 9-703. The inertial rotation equipment is on loan from the NASA Johnson Space Center. The Division equipment was paid for by the US West Foundation. The assistance of the HITL staff is gratefully acknowledged, in particular Robert Burstein and Arturo Gonzalez for hardware

support and Paul Schwartz for software support. This paper benefited from discussions of adaptation with Robert Patterson at Washington State University.

References

Anastasi, A. (1988). *Psychological Testing*. New York,NY:Macmillan Publishing Co.

Barfield, W. & Weghorst, S. (1993). The Sense of Presence Within virtual environments: a conceptual framework. In *Human Computer Interaction: Software and Hardware Interfaces*. UK:Elsevier Science Publishers.

Benson, A.J., Hutt, E.C.B. & Brown S.F. (1989). Thresholds for the perception of whole body angular movement about a vertical axis. *Aviation, Space, and Environmental Medicine*, Mar., 205-213.

Berthoz, A., Pavard, B. & Young, L.R. (1975). Perception of linear horizontal self-motion induced by peripheral vision (linear vection): Basic characteristics and visual-vestibular interactions. *Experimental Brain Research, 23*, 471-489.

Brindley, G.S. (1970). *Physiology of the Retina and Visual Pathway*. Williams & Wilkins.

Carpenter-Smith, T.R., Futamura, R.G., & Parker, D.E. (1995). Inertial acceleration as a measure of linear vection: An alternative to magnitude estimation. *Perception and Psychophysics, 57* (1), 35-42.

Gonshor, A. & Jones, G.M. (1976). Extreme vestibulo-ocular adaptation induced by prolonged optical reversal of vision. *Journal of Physiology* (London), 256, 381-414.

Hatada, T., Sakata, H., & Kusaka, H. (1980). Psychophysical analysis of the "sensation of reality" induced by a visual wide-field display. *SMPTE Journal, 89*, 560-569.

Held, R. & Durlach, N. (1991). Telepresence, time delay and adaptation. In Ellis, S.R. (Ed.) *Pictorial Communication in Virtual and Real Environments*. London,UK:Taylor & Francis.

Hendrix, C.M. (1994). *Exploratory Studies on the Sense of Presence as a Function of Visual and Auditory Display Parameters in Virtual Environments*. Unpublished Master's Thesis, University of Washington. Available online from http://www.hitl.washington.edu/projects/afosr/hendrix/home.html"

Helson, H. (1964). *Adaptation-Level Theory*. New York,NY:Harper & Row.

Howard, I.P. (1986). The Vestibular System. In *Handbook of Perception and Human Performance*. New York,NY:John Wiley and Sons.

Huang, J.K. & Young, L.R. (1981). Sensation of rotation about a vertical axis with a fixed visual field in different illuminations and in the dark. *Experimental Brain Research, 41*, 172-183.

Huang, J.K. and Young, L.R. (1987). Influence of visual and motion cvues on manual lateral stabilization. *Aiation, Space, and Environmental Medicine, 58* (12), 1197-1204.

Huang, J.K. and Young, L.R. (1988). Visual field influence on manual roll and pitch stabilization. *Aviation, Space, and Environmental Medicine, 59* (7), 611-619.

Jex, H.R. (1988). Measuring mental workload: Problems, progress, and promises. In Hancock, P.A. & Meshkati, N. (Ed.s) *Human Mental Workload*. Amsterdam, Netherlands:North-Holland.

Mergner, T. & Becker, W. (1990). In Warren, R. & Wertheim, A.H. (Ed.s) *Perception and Control of Self-Motion*. Hillsdale, N.J.:Lawrence Erlbaum Associates.

Nemire, K., Jacoby, R.H., & Ellis, S.R. (1994). Simulation Fidelity of a Virtual Environment Display. *Human Factors, 36* (1) 79-93.

Prothero, J.D., Hoffman, H.G., Parker, D.E., Furness III, T.A., & Wells, M.J. (1995). Foreground/background manipulations affect presence. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*. Santa Monica,CA:Human Factors Society.

Sheridan, T.B. (Ed.) (1992). *Telerobotics, Automation, and Human Supervisory Control*. Cambridge, MA:The MIT Press.

Slater, M., & Usoh, M. (1993). Presence in immersive virtual environments. *Proceedings of the IEEE Virtual Reality Annual International Symposium*, 90-96. Piscataway, NJ:Institute of Electrical and Electronic Engineers.

Wilson, H.R. & Humanski, R. (1993). Spatial frequency adaptation and contrast gain control. *Vision Research, 33* (8), 1133-1149.

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

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

The Effect of Task Automatisation in the Automotive Context: A Field Study of an Autonomous Intelligent Cruise Control System

Nicholas J. Ward[1], Stephen Fairclough[1] and Mark Humphreys[2]

[1] HUSAT Research Institute, Loughborough University, UK.
[2] Department of Human Sciences, Loughborough University, UK.

Abstract

This study examined the effect of Autonomous Intelligent Cruise Control (AICC) on driver comfort and driving behaviour. A sample of 15 male subjects were divided into groups of high and low sensation seekers. Subjects drove a set route along a highway in moderate traffic with and without a prototype AICC system. Measures were taken of arousal, mood, effort as well as situation awareness and driving behaviour. There were indications that use of the particular AICC tested resulted in reduced levels of arousal and effort in speed and headway control. There were also indications of reduced performance in both proper lane maintenance and yielding to other traffic. This subject group set the AICC to higher speeds and shorter headways than were normally driven, with lesser variation of set parameters resulting in fewer episodes of excessively high speed and short headways.

Introduction

The potential for task automatisation to lead to reduced situation awareness has been demonstrated in the aviation domain (Wiener, 1988). Only recently has the issue of task automatisation in the driving context been systematically examined (Endsley & Kiris, 1995). Autonomous Intelligent Cruise Control (AICC) is currently under development to support the driver in the longitudinal control of the vehicle. Specifically, AICC assumes a control level function over acceleration and braking in order to sustain a specified speed. AICC also assumes a tactical level function by sustaining a specified time headway in relation to detected traffic. By subsuming the functions of monitoring and maintaining speed and time headway, AICC is expected to reduce the effort and stress of long distance driving in traffic. Thus, AICC is intended to function as a 'comfort' system.

However, evidence that AICC provides a comfort benefit has not been demonstrated in a naturalistic setting. Neither has the broader safety effect of automating the driving task. Specifically, for the longitudinal control and tactical functions, the driver is placed 'out-of-the-loop' by the AICC (Endsley, 1995). As a result, there may be a loss of situation awareness (SA) of critical events in the driving environment due either to (i) reduced arousal levels (Mahalel & Szternfeld, 1986), or (ii) an absence of task relevant feedback (Norman, 1990). This study sought to empirically quantify the extent to which AICC influences the comfort and the situation awareness of the driver during highway driving in traffic.

Method

Subjects

In order to avoid novelty effects, the sample was comprised of 15 male subjects who had experience with conventional cruise control and the type of vehicle fitted with the prototype system. The average age of the sample was 39 years (sd = 11.3 years) with an average reported annual mileage of 12500 miles (sd = 3500 miles). Subjects were pre-screened on the basis of the Sensation Seeking Scale (Zuckerman, 1979) in order to obtain two groups which differed significantly in terms of a disposition for sensation seeking [$t(13) = 7.16, p < .05$]: 7 High Sensation Seekers (HSS) [$M = 26$], and 8 Low Sensation Seekers (LSS) [$M = 14$].[1]

Research AICC System

Commercial confidentiality precludes a full description of the research AICC system. The test vehicle was a luxury sedan equipped with an automatic transmission as well as power assisted steering and braking. A colour liquid crystal display (LCD) mounted within the instrument binnacle served as the interface. This interface gave an indication of the active system mode, the user selected set speed, and set time headway (as well as actual time headway). A grouping of six buttons, laid out in two banks of three integrated into the steering wheel, operated the system and input the set speed and headway parameters.

Procedure

All subjects (HSS, LSS) received a half an hour training session followed by two half hour driving sessions with and without the AICC in traffic on a major highway.[2] The order of driving sessions was counterbalanced (AICC, non-AICC). The sessions were scheduled in order that traffic volume was comparable. The study comprised a mixed 2 (AICC: AICC vs. non-AICC) within X 2 (SS: HSS vs. LSS) between subject factor design.

Dependent Measures

(1) Comfort data: Affective state and level of arousal were measured with the UWIST Mood Adjective Checklist (UMACL) in terms of (i) Energetic Arousal, (ii) Tense Arousal, and (iii) Hedonic Tone (Mathews, et al., 1990). Level of arousal was also measured with the Modified Stanford Sleepiness Scale (M-SSS) in terms of (i) Energy and Activation, and (ii) Sleepiness (MacLean et al., 1992) . Mental workload was measured with the NASA R-TLX (Byers et al., 1989). The UMACL, M-SSS and R-TLX were administered after each session. The UMACL and M-SSS were also administered before the first session in order to obtain baseline measures of mood and arousal (pretest).

---

[1] Given the reliance on subjective measures in this study, subjects were also selected on the basis of low lie scale values on the adult EPQ-R produced by Eysenck and Eysenck (1991) in order to eliminate from the sample those individuals with an inclination to fake or give inconsistent answers.

[2] Subjects were fitted with electrodes for the psychophysiological recording of EEG and ECG. The results of this data set are not included in this report.

(2) Situation awareness data: Three techniques were used to measure SA. First, a reaction-time task was designed to simulate brake light detection. The target was a light emitting diode mounted along the sight line of the driver at the end of the hood. The light was randomly activated by the experimenter and responded to by the subject with a finger switch. Second, a self-report questionnaire relevant to the driving task was devised based on the three levels of SA proposed by Endsley (1995). These two measures did not indicate any significant effect for AICC and are not discussed in this report. Third, the experimenter recorded the frequency of driving errors during each session based on an observation record used by Burns and Wilde (1995): (i) poor lane position (i.e. wandering), (ii) unsafe lane change/passing, (iii) failure/improper signal, (iv) failure to yield to others, (v) following too closely, (vi) distraction (poor attention), (vii) confusion (poor judgement), and (viii) excessive speed (> 85 mph).

(3) Vehicle and system operation data were recorded by an onboard computer during driving sessions. The computer recorded the active mode of the system, the current speed and time headway for the vehicle, as well as the set speed and headway currently specified by the subject.

Results

Mental Workload

The average score from the NASA R-TLX and each of the individual items were analysed separately in a 2 x 2 mixed factor ANOVA. The only significant effect of AICC on mental workload emerged as a significant interaction with SS for the 'Effort' item [$F(1,13) = 4.56$, $p < .05$]. As shown in Figure 1A, reported effort in maintaining adequate speed and headway was significantly less [Tukey HSD, $d_{crit} = 17.44$, $p < .05$] with the AICC [$M = 60.86$] than without the AICC [$M = 40.43$] only for the HSS group.

Arousal Level

The two factors of the M-SSS were calculated for each session. Each factor was analysed separately within a 2 x 2 mixed factor ANCOVA using the pretest score as a covariate. There was a marginally significant main effect for AICC for the level of energy and activation [$F(1,11) = 3.98$, $p < .10$]. Subjects reported lower levels of energy and activation with the AICC [$M = .41$] than without the AICC [$M = .54$].

Affective State

The factors of the UMACL were calculated for each session. Each factor was analysed separately within a 2 x 2 mixed factor ANCOVA using the pretest score as a covariate. There was a marginally significant interaction between SS and AICC for the Tense Arousal factor [$F(1,12) = 3.68$, $p < .10$]. Based on a post-hoc comparison of means [Tukey HSD Test, $d_{crit} = 1.84$, $p < .05$], it was apparent that the HSS group reported significantly less tense arousal in conjunction with the AICC [$M = 9.56$] than did the LSS group with the AICC [$M = 13.75$] *and* without the AICC [$M = 12.88$].

Situation Awareness

The total count of observed driving errors in each category of the observational record were analysed together within a 2 x 2 mixed-factor MANOVA. The MANOVA indicated a significant

main effect for AICC[$F(8,6) = 5.26, p < 0.05$]. Univariate analysis indicated that subjects were observed significantly more often [$F(1,13) = 9.32, p < .01$] to have poor lane position with AICC [$M = 2.40$] than without AICC [$M = 1.30$]. Subjects were observed significantly more often [$F(1,13) = 9.93, p < .01$] to fail to yield to traffic with AICC [$M = 1.8$] than without AICC [$M = 0.27$]. Subjects were also observed significantly more often [$F(1,13) = 5.61, p < .05$] to be distracted with AICC [$M = 0.60$] than without AICC [$M = 0.20$]. By contrast, subjects were observed significantly *less* often [$F(1,13) = 13.93, p < .01$] to follow too closely with AICC [$M = 0.67$] than without AICC [$M = 3.07$]. Finally, there was a significant interaction between SS and AICC for episodes of excessive speeds [$F(1,13) = 5.37, p < .05$]. As shown in Figure 1B, the HSS group were observed speeding significantly less often [Tukey HSD Test, $d_{crit} = 2.59, p < .05$] with AICC [$M = 3.43$] than without AICC [$M = 7.29$]

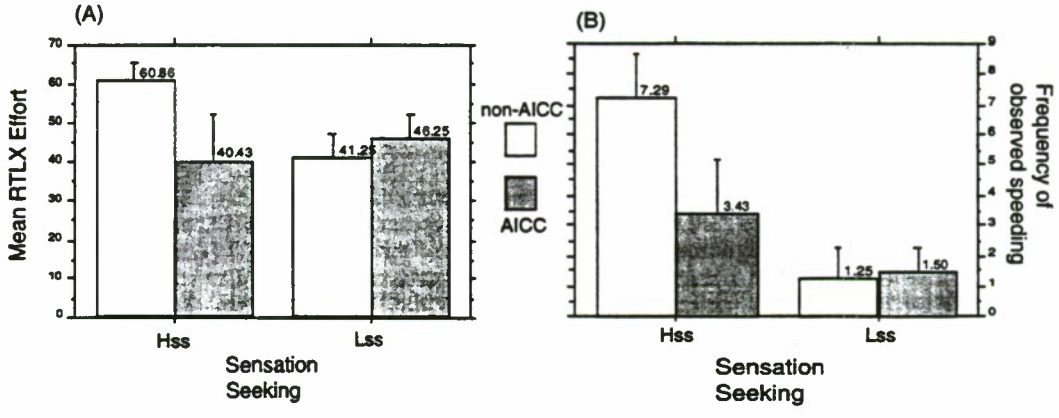


Figure 1. (A) Reported level of effort, and (B) Frequency of observed episodes of excessive speed for the HSS and LSS group with and without AICC.

Vehicle and System Operation

Data for the AICC sessions were filtered to only include recordings taken whilst the system was operating (i.e. data was excluded when system was off or in standby mode).[1] The mean and coefficient of variance for the speed and headway data were then computed for each session. In this manner, *set* speed and time headway in the AICC session were compared to *driven* speed and headway in the non-AICC session. The speed and headway values were analysed separately within a 2 x 2 mixed factor ANOVA.

There was a significant main effect of SS for mean speed [$F(1,13) = 8.60, p < .05$] with the HSS group adopting (either set or driven) higher speeds [$M = 82$ mph] than the LSS group [$M = 74$ mph]. There was a significant main effect of AICC for mean speed [$F(1,13) = 20.71, p < .01$] with subjects *setting* a higher average speed with AICC [$M = 80$ mph] than they *drove* at without AICC [$M = 76$ mph]. There was also a significant main effect of AICC for variation in speed

[1]The same results reported here for this data filtration were also evident when set speed was calculated only when the system was maintaining the specified vehicle speed in 'Cruise' mode, and when set headway was calculated only when the system was maintaining the specified time headway in 'Following' mode.

[$F(1,13) = 122.07$, $p < 0.001$] with subjects varying the setting of their speed less with AICC [$M = 0.04$] than they varied their driven speed without AICC [$M = 0.10$].

There was a significant main effect of AICC for mean time headway [$F(1,13) = 27.87$, $p < 0.01$] with subjects setting a shorter headway with AICC [$M = 1.4$ s] seconds than they drove at without AICC [$M = 1.8$ s]. There was also a significant main effect of AICC for variation in headway [$F(1,13) = 279.88$, $p < 0.001$] with subjects varying the setting of their speed less with AICC [$M = 0.09$] than they varied their headway without AICC [$M = 0.46$].

Discussion

There was some indication that AICC can affect driver comfort. Overall, subjects reported lower levels of energy and activation with the AICC. At least for the group of high sensation seeking subjects, a reduced arousal level may be attributed to a perceived reduction in effort required to maintain an adequate speed and headway with AICC . However, the effect of the AICC was not sufficient to influence the mood of subjects nor the overall level of mental workload.

There was some evidence that poor attention to lane positioning and failure to yield to other traffic was more frequent with AICC. This may be indicative of distraction arising from the prototype interface or an attempt to anticipate the actions of the system. It may also suggest that subjects were relying on the system to handle merges with traffic. There may be less potential for driver distraction with a simplified driver interface combined with a more predictable system response.

The system was used to set faster speeds and shorter headways than typically driven without the AICC. The practical significance of the magnitude of these changes can be debated. For example, it is not clear if the 4 mph increase in set speed and the almost 1/2 second decrease in set headway amounts to a reduction in safety. The AICC did stabilise speed such that the number of excessive speed events were observed to be reduced, at least for high sensation seekers. Similarly, the reduction of set speed and headway variation by as much as a factor of 4 may suggest that AICC can improve traffic flow and harmonisation. However, these suppositions require empirical verification in future research.

Conclusion

The AICC produced a degree of comfort as intended. The system was used in a manner that may improve traffic flow and harmonisation, but only if the selection of higher speeds and shorter headways do not increase the rate or severity of accidents (see Chira-Chavala & Yoo, 1994). Subsequent iterations of the AICC should adopt a simplified driver interface. Algorithms for the handling of the vehicle by the system should lead to reliable and predictable responses. In this manner, the distraction of the driver may be reduced. Future research would be advised to conduct long term trials to examine the effect of AICC on habitual users. This approach will provide more definitive results on system use and the effects on driver expectations and acceptance.

References

Burns, P.C. & Wilde, G.J.S. (1995). Risk taking in male taxi drivers: Relationships among personality, observational data, and driver records. *Personality and Individual Differences, 18*, 267-278.

Byers, J.C., Bittner, Jr., A.C., & Hill, S.G. (1989) Traditional and Raw Task Load Index (TLX) correlations: are paired comparisons necessary? In A. Mital (ed.) *Advances in Industrial Ergonomics and Safety I*, Taylor & Francis, pp481-485.

Chira-Chavala,T. & Yoo,S.M. (1994) Potential Safety Benefits of Intelligent Cruise Control Systems. *Accident Analysis and Prevention, 26*. 135-146.

Endsley, M.R. (1995) Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*, 32-64.

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

MacLean, A.W., Fekken,G.C., Saskin, P. and Knowles, J.B. (1992) Psychometric evaluation of then Stanford Sleepiness Scale. *Journal of Sleep Research, 1*, 35-39.

Mahalel,D & Szternfeld,Z. (1986) Saftey Improvement and Driver Perception. *Accident Analysis and Prevention, 18*, 37-42.

Matthews,G., Jones,D.M., and Chamberlain,A.G. (1990) Refining the measurement of mood: the UWIST Mood Adjective Checklist. *British Journal of Psychology ,81*, 17-42.

Norman, D.A. (1990) The 'Problem' with Automation: inappropriate feedback and interaction, not "over-automation"'. *Phil. Trans. R. Soc. Lond. B327*, 585-593.

Wiener, E.L., (1988) Cockpit automation, In E.L.Weiner & D.C.Nagel (Eds.), *Human Factors in Aviation*, San Diego: Academic Press.

Zuckerman,M. (1979) *Sensation seeking: beyond the optimal level of arousal*. Hillsdale: Lawrence Erlbaum.

The Effect of Automotive Head-Up Displays on Attention to Critical Events in Traffic

Nicholas J. Ward[1], Andrew Parkes[1] and Phil Lindsay[2]

[1] HUSAT Research Institute, Loughborough University
[2] Department of Human Sciences, Loughborough University

Abstract

In order to address the issue of cognitive capture with an automotive Head-Up Display (HUD), this study examined the effect of simulated HUDs on the ability of drivers to selectively attend to safety-critical low probability events in the traffic context of a driving simulator. Sixteen male subjects completed four 20 minute driving sessions in which they had to maintain lateral and longitudinal position behind a lead car of variable speed. In one condition no HUD was present. In the remaining conditions, three HUDs were simulated with incremental levels of task relevant information. At a random period in each session, the lead car would stop abruptly. The reaction times of subjects to this event were recorded. Reaction times were significantly slower with a HUD present. Field dependency was significantly related to slower reaction times only with a HUD. HUDs were reportedly more distracting and resulted in an increased task workload. The results are discussed in terms of the practical implementation of HUDs in road vehicles.

Introduction

With the recent advent of Intelligent Transportation Systems (ITS) to support traffic safety, efficiency, and driver comfort, there are a number of emerging in-vehicle information displays. These displays are in competition for limited dashboard space. Head-Up Displays (HUDs) have been utilised in aviation for many years (Weintraub & Ensing, 1987) and are now being proposed as an option for several forms of information display within the automotive sector. The implications for the transfer of this display technology into the automotive sector from such a disparate context as that of the commercial and military flying environment, is unclear in terms of both design and safety.

Ward and Parkes (1994) have discussed the potential safety implications of using HUDs for ITS applications in automobiles for the general public. A primary concern identified in this review is the potential for automotive HUDs to induce a 'cognitive capture' effect to the detriment of the awareness of safety critical events in the external scene. The potential for HUDs to induce a cognitive capture effect has been demonstrated in simulated aviation applications (e.g. Fischer et al, 1980; Haines, 1980; Wickens and Long, 1995). Ward and Parkes (1994) remark that although a cognitive capture effect of automotive HUDs warrants concern, this hypothesis has not been tested in a driving context with ecologically valid tasks (see Bossi et al., 1994). Thus, the safety consequences of such changes to the allocation of attention, information processing and driving-

behaviour are not yet clear for automotive HUD applications. These consequences may be exacerbated by individual differences in perceptual style. For example, older drivers typified by greater field dependency may be more distracted by the HUD (Ward et al., 1994). Moreover, the distraction effect of HUDs predominates under conditions of high task load (Larish & Wickens, 1995). Thus, it is possible that high levels of HUD information may instill a distraction effect by increasing mental (processing) workload.

To address this issue of cognitive capture with automotive HUDs, this study examined the effect of simulated HUDs (with verying information levels) on the ability of drivers to selectively attend to safety-critical, varying low-probability events in the traffic context of a driving simulator.

Method

Subjects

The study sample consisted of 16 male subjects. The mean age of the sample was 33.5 years (range = 22 to 54). The average annual mileage reported by this sample was 12,000 miles. All subjects had corrected far visual acuity of 20/20 for both eyes and held a valid driving licence. An Embedded Figures Test (EFT) was administered as a measure of field dependency (Melancon & Thompson, 1987). In order to ensure that the subjects were highly motivated to perform well in the experimental task, the guise of performance related payment was introduced. In practice, each subject received the same amount regardless of performance.

The Driving Simulator

This simulation was controlled by a Pentium PC and projected by a SONY Superdata BX Colour Projector onto a 4.5m x 2.5m screen. The simulator was interfaced to a Ford Granada. The vehicle and the screen were enclosed in a blackened room to eliminate ambient light. Steering, accelerator and braking actions upon the vehicle were reproduced in the simulation. The driving simulator depicted a simple line graphic view of a two-lane highway from the perspective of the driver. Apparent speed feedback was provided through the vehicle speedometer.

The Driving Context

The simulator environment consisted of black and white wire frame vector graphics depicting a two-lane dual highway with a single other vehicle. The computer monitored a box region behind the lead vehicle. This region was not directly visible to subjects. The subjects task was to maintain both a constant longitudinal and lateral position within this box despite the independent manoeuvring of the lead car. The lead vehicle accelerated and decelerated sinusoidally [mean speed = 50 mph; amplitude = 10 mph; period of cycle = 30 seconds]. At a randomly chosen point during the trial, the experimenter initiated a low probability event which consisted of the lead car braking abruptly to a standstill, either with or without braking lights. Subjects were instructed to brake to avoid an accident at the instance the lead car initiated the stopping event. Reaction time to the lead car event was recorded by the computer, as the elapsed time from the onset of the deceleration to the instance that the subject activated the brake.

The HUD Symbology and Information Level

A number of information types were generated which were relevant to successful performance of the car following task. The HUD symbology for these task relevant information types is illustrated in Figure 1a. These information types were combined into formats to represent different HUDs of increasing degrees of information and complexity. The combination of information types which defined each display type is illustrated in Figure 1b.[1]

Procedure

Subjects received a briefing on the driving simulator, driving task and HUD information. They then received a 15 minute practice session with the NO HUD configuration. For each test trial, the subject initially drove the vehicle behind the lead car and stabilised their position. The lead car maintained a constant speed of 50 mph for 2 minutes and then started the sinusoidal wave form, at which point data collection commenced. At a random point chosen between 9 minutes and 16 minutes into the trial, the experimenter initiated an event. The trial then continued with the lead car accelerating away again and continuing the sinusoidal motion. After approximately 20 minutes the trial was halted and subjects completed the NASA R-TLX as a measure of task workload. Subjects completed the task with all display types. The order of conditions was random. In order to minimise expectancy effects, only one event occured in each condition. Each subject experienced 4 events in total, comprised of pairings of each of the two event types (brake light, no brake light). The events were scheduled so that one of the events in a pair occured in the NO HUD condition and the second event in a pair in either the HUD 2 or HUD 3 condition. The sample was evenly divided by the type of event pairing, and equally subdivided into which HUD condition the second event in a pair occurred.

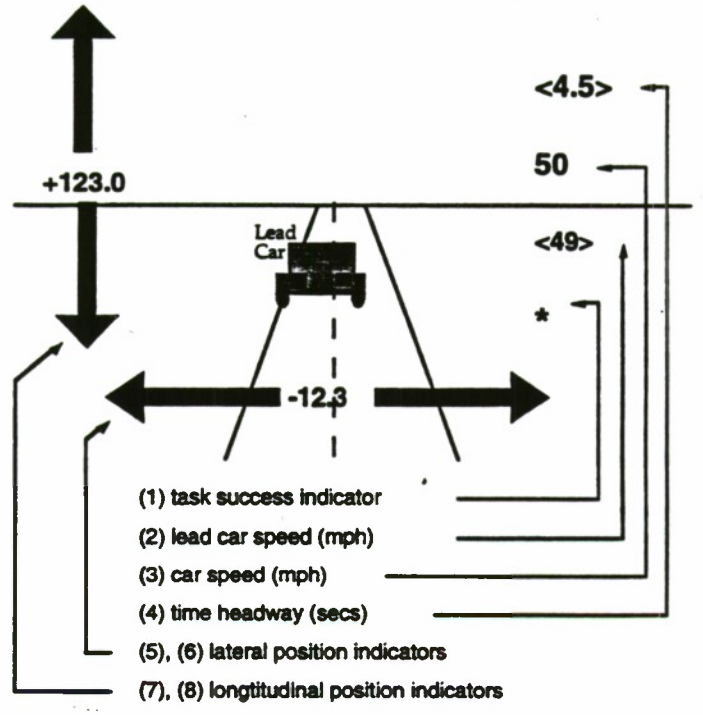
Results

Mental Workload

A total workload score was calculated from the NASA R-TLX by averaging the individual items and analysed as a repeated measure one-way ANOVA between the four types of display (NO HUD, HUD 1, HUD 2, HUD 3). Polynomial contrast analysis indicated a marginally significant linear trend [$F(1,15) = 3.38$, $p < .09$] as shown in Figure 2A. Reported mental workload increased as a function of HUD information level [NO HUD $M = 47$, HUD 1 $M = 48$, HUD 2 $M = 51$, HUD 3 $M = 54$].

[1] Note that the 'NO HUD' condition contained a binary indicator in a HUD format (an asterisk) of task success/ failure. The car following task was incomprehensible without this binary indicator as a default. The NO HUD classification is still justified because this indicator is nominal.

(A)



(B)

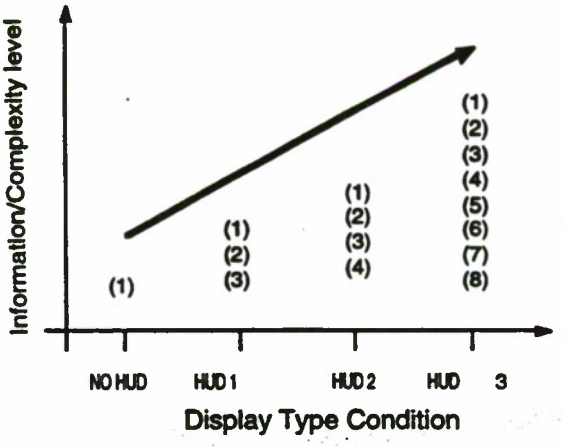


Figure 1. (A) Illustration of task relevant HUD information, and (B) Definition of display types.

Ranking of Distraction

During debriefing, subject were required to rank all the display conditions in terms of the distraction imposed from performing the primary task. The rankings by subjects were statistically reliable [$W = 25.2, p < .05$]. The results for each subject were summed to give an overall rank for each display type. There was a linear trend indicative of greater distraction with increasing levels of HUD information as shown below Friedman Test (3) = 24.52, $p < .001$:

HUD Type	NO HUD	HUD 1	HUD 2	HUD 3
Total Rank	54	49	36	21
Overall Rank	4th	3rd	2nd	1st

Lateral and Longitudinal Control Performance

The percentage of trial time that the subject was inside the defined headway zone was calculated and analysed as a repeated measure one-way ANOVA between the four types of display (NO HUD, HUD 1, HUD 2, HUD 3). Polynomial contrast analysis indicated a significant quadratic trend [$F(1,15) = 7.89, p < .01$] as shown in Figure 2B. Performance was optimised with the two moderate information level HUDs [HUD 1 $M = 44.0$ %, HUD 2 $M = 41.9$ %] and deteriorated comparably with NO HUD [$M = 38.6$ %] and the most complex HUD 3 [$M = 38.6$ %].

Event Reaction Time

The reaction time data was analysed with a mixed design 2 x 2 x 2 ANOVA. The average reaction time for the deceleration events without brake lights [$M = 2.59$ sec.] was significantly slower [$F(1,12) = 45.40, p < .0001$] than for the brake light events [$M = 1.32$ sec.]. Overall, the average reaction time with a HUD [$M = 2.14$ sec.] was significantly slower [$F(1,12) = 6.44, p < .05$] than in the NO HUD condition [$M = 1.77$ sec.]. Eleven of the sixteen subjects had a slower reaction time when using a HUD.

To examine the relationship between perceptual style (field dependency) and response to the traffic events, subject EFT scores were correlated with reaction times in the HUD and NO HUD conditions. The partial correlation was computed to remove the confounding relationship between subject age and field dependency [$r(14) = 0.66, p < .01$]. It was apparent that greater field dependency was associated with significantly slower reaction times to the critical events when a HUD was present [$r(14) = 0.55, p < .05$], but not in absence of a HUD [$r(14) = 0.22$, ns.].

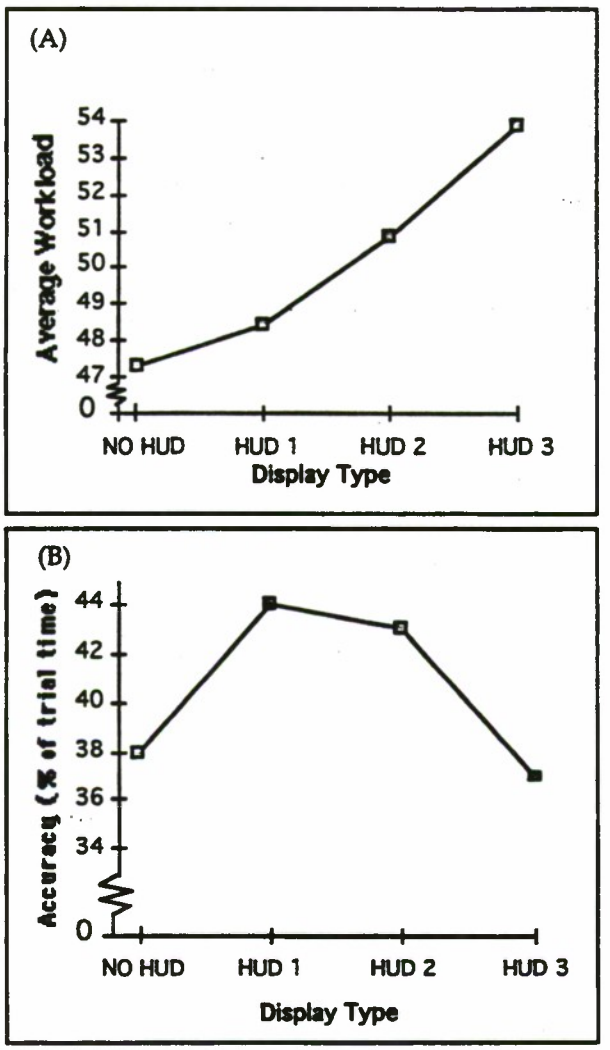


Figure 2. (A) average reported mental workload (NASA R-TLX), and (B) percentage of time within target region of lateral and longitudinal control in each display condition.

Discussion

This study has provided an initial demonstration of the potential cognitive capture effect of automotive HUDs within the confines of a simulator-based study. The HUDs were consistently reported by subjects to be a source of distraction. The presence of a HUD also increased the subjective experience of task workload. The behavioural manifestation of the presence of a HUD

was evident by deteriorated longitudinal and lateral control accuracy with high levels of HUD information, and significantly slower reactions times to safety critical events in the driving context.

Overall, reaction times with a HUD present were 0.38 seconds slower than without a HUD. This represents a 21% reduction in response time. The consequence of this reduction can only be inferred. For example, Schweitzer, et al. (1995) provide tables of reaction times from drivers in response to braking event of a lead vehicle on real roads. These values are tabulated for headways of 6m and 12m and for travel speeds of 40 mph and 50 mph under various degrees of driver expectancy. The most applicable set of values in relation to the current study is for drivers with partial expectation travelling at 50 mph at 12 meters. If the same proportional reduction in the current study is applied to the 50th percentile reaction time from this data set (0.61 sec.), then an average driver would shift toward the 90th percentile. For drivers already prone to higher percentile reactions time such as older persons or those with a field dependent perceptual style, the effect of a HUD may lead to extreme delays which may exceed safety margins.

Conclusion

Any cognitive capture effect imposed by the deployment of HUDs into road vehicles for use by the general population can have serious consequences for safety. It is clear that the use of HUDs in automotive applications should be supported by careful research into the design and evaluation of these displays with particular reference to older drivers and those with a field dependent perceptual style. The apparent exacerbation of cognitive capture effects by increased level of HUD information necessitates a systematic investigation of the amount of information that is optimal for task performance and minimisation of distraction.

This study simultaneously increased information load and HUD presentation. In order to distinguish the effect of increased information from the presentation characteristics of the HUD, future research must manipulate both information level and display type (e.g. conventional head-down, mid-head, HUD) as independent factors (Wickens & Lung, 1995). The tentative evidence of cognitive capture observed in this simulator study is of sufficient significance to merit confirmation in a naturalistic field study with a functional HUD.

Reference

Bossi, L.L.M., Ward, N.J., & Parkes, A.M. (1994). The effect of enhanced image Head-Up Displays on driver peripheral visual performance. *12th Congress of the International Ergonomics Association*, Toronto, Canada (August 15-19).

Fischer,E., Haines, R.F., & Price,T.A.(1980). *Cognitive Issues in head-up displays.* NASA(Ames) Technical Paper 1711. Moffett Field, CA: NASA Ames Research Centre.

Haines,R.F. (1980). A breakdown in simultaneous information processing. In: Obrecht,G., & Stark,L.W. (Eds.) *Presbyopia Research: From Molecular Biology to Visual Adaptation*. Plenum Press.

Larish, I., & Wickens, C.D. (1995). *Divided attention with superimposed and separated imagery: Implications for Head-Up Displays.* Technical Report No. ARL, 91-4/NASA HUD-91-1. NASA Ames Research Center, Moffett Field, CA.

Melancon, J. G., & Thompson, B. (1989). Measurement characteristics of the embedded figures test. *Psychology in the Schools, 26*, 69-78.

Schweitzer, N., Apter, Y., Ben-David, G., Liebermann, D.G., & Parush, A. (1995). A field study on braking responses during driving: II. Minimum driver braking times. *Ergonomics, 38,* 1903-1910.

Ward, N.J., Parkes, A.M., Crone, P. (1994). The legibility of Head-Up Displays within the driving environment: The effect of background scene complexity. *Vehicle and Navigation and Information Systems International Conference*, Yokohama, Japan (August 31 - September 2).

Ward, N.J. & Parkes, A.M. (1994). Head-Up Displays and their automotive application: An overview of the Human Factors issues. *Accident Analysis and Prevention, 26*, 703-718.

Weintraub, D.J., & Ensing,M. (1992). *Human Factors in head-up display design: The book of HUD*. CSERIAC, MI: University of Michigan.

Wickens, C.D., & Long, J. (1995). Object versus spaced-based models of visual attention: Implications for the design of head-up displays. *Journal of Experimental Psychology, 1*(3), pp 179-193.

Modeling Situation Awareness: Using the Influence Diagram

Joseph Sferrazza and Marc B. Wilson

Dowling College

Abstract

Accidents and incidents are man-made disasters and are investigated usually to place blame or ascertain cause. The investigation process consists of a finding of facts or review of the sequential and non-sequential events and conditions which lead up to the accident. This paper uses an influence diagram for assessing situation awareness. An influence diagram establishes a knowledge map of conditions, decisions, events and errors based on historical and surrogate data. The modeling technique is a new method for understanding situation awareness.

Introduction

The safe and successful operation of any complex system based upon the human-machine interface is directly related to situation awareness (SA). SA is the knowledge and perception of important components in an environment, along with an acute understanding of those components. Focusing not only on the internal space in which an operator is functioning within a machine, situation awareness also concentrates on the external environment of a particular circumstance. When properly executed, situation awareness is highly effective in determining a safe and successful outcome of a condition or an event. This is especially true when the awareness is applied to operations involving time-sensitive decisions with a low tolerance for error, as in the case of aviation or maritime maneuvers. Aviation is real-time and maritime is lag-time. Increasing situation awareness greatly reduces the risk factor involved in a circumstance, regardless of complexity.

There are five broad dimensions in the total concept of situation awareness. In order to achieve maximum perception of a given event, one must be alert to all five conditions and how they affect the outcome for the desired level of success and safety. The first dimension of situation awareness is a spatial awareness. This awareness is the physical orientation of the operator in reference to a normal position. For example, is the operator right side up, upside down, leaning left or right? A pilot engaged in making a left bank turn knows that the positioning to the left is not a normal condition and must return to a normal situation upon completion of that turn. A loss of spatial awareness is spatial disorientation. Vertigo is a symptom where the operator has a false notion of his or her actual position. Spatial disorientation can cause a loss of spatial awareness.

Identity awareness constitutes whether another entity in the situation is or is not a problem. Using another example, a sailor sees an aircraft in their line of sight and must determine if that aircraft represents a threat to his safety such as the Vincennes. This simple concept of threat/no threat condition allows for the immediate assignment of a risk factor to that element to clarify the identity awareness.

Temporal awareness is based on the concept of how things change in a situation with respect to time. How fast a condition changes and awareness of that change is temporal alertness. A person traveling to a destination with no particular time limit, has little temporal awareness, as opposed to the person running late, traveling in dense traffic. A combination of temporal and spatial awareness with specific reference to this research paper is introduced later.

In any situation, someone is responsible for each task and this is called responsibility awareness. In certain operations, like driving a car or a motorman on a subway with single controls, only one person can be in control of the machine. There are certain circumstances, however, where two or more people have marrying degrees of responsibility for the combined consequence, as would be the case in a large aircraft. The pilot, co-pilot, flight attendant, etc., would have specific duties and responsibilities and the knowledge of these is also a critical part of situation awareness. This is called Crew Resource Management (CRM), which is only part of situation awareness.

The final dimension of situation awareness is the expectancy element. This classification deals with the behavior of the system in response to the various inputs. When the pilot making the left bank turn in the spatial awareness example anticipated a return to the normal position after the turn, he or she had expectancy awareness. The expected behavior or reaction of an element is based upon acute knowledge of the system in which the operator is functioning. Expectancy awareness is only relevant for dynamic events or conditions when temporal awareness is introduced. The reason for the association of elements, expectancy and temporal, is that without time to define the limits of an expected occurrence, the occurrence is meaningless. For example, the driver of a vehicle expects that vehicle to slow down when the brakes are applied, but that expectancy is anticipated within a specific time frame, otherwise there is no purpose of slowing down. This is very important in lag-time events such as ships or boats maneuvering.

In this paper, the model used is analyzed using the integration of spatial and temporal, known as navigational awareness (Andre, Wickens, Moorman and Boschelli, 1991). Being under a more descriptive awareness, the navigational dimension incorporates a sense of location for the operator which includes the concept of time. For example, a driver of a vehicle traveling on a road under normal conditions knows that he has spatial awareness due to his correct position for the safe operation of that vehicle, sitting upright, facing forward. Add to that dimension his knowledge of geographic positioning within the environment with reference to time, and one has the navigational awareness dimension, which is the orientation of a system within space and time.

Decision Trees

In order to determine if situation awareness has been achieved in a specific circumstance, certain decision making tools can be utilized. The process of decision making is traditionally characterized by a large number of uncertain quantities and interrelated alternatives. This complexity can be deciphered by the use of decision trees. Decision trees are graphic representatives of probability logic applied to decision alternatives (Thierauf & Klekamp, 1970, p. 82). Its branches begin at the first chance event. Each chance event produces two or more possible effects. Some chance events can lead to other chance events and subsequent decision points. The tree's branches are assigned values. The values are based on research that provides probabilities for certain chance events. These probability factors, included on all the branches from the base outward, help determine the best possible course of action in making the decision.

Decision trees have distinct advantages in problem solving because they can make outcomes and alternatives associated with a decision problem explicit. In the examination of a decision tree, one can quickly ascertain the alternative actions under consideration and the possible outcomes that are associated with each uncertain event. Also the sequential order of the decisions and events related to the problem are quite evident, being resolved one after another from branch to trunk or vice-

versa. The solution procedure for solving a decision tree is straightforward but there are also some significant drawbacks to using one. Even simple problems can evolve into a large bushy tree with numerous branches because of the inclusion of new variables along the way. There is no system in a decision tree tool to show how the variables interact with each other, because they follow a sequential order. The sequential order can also hinder the process as many problems can not directly relate to a rigid progression of events.

Influence Diagrams

Another decision making tool that goes beyond the limitations of a decision tree is an influence diagram. This diagram has proven to be a new "tool of thought" (Howard, 1990, p. 3) that can facilitate the formulation, assessment and evaluation of decision problems. Influence diagrams are used to make decisions in any industry, including medicine, business or public policy.

Communicating information more effectively than decision trees, influence diagrams can: 1) define what decision must be made and their sequence; 2) provide specific information that is known at the time of each decision; 3) display how the events are dependent on, or influence, each other; and 4) illustrate which events and decisions in a situation directly impact the overall value to the decision maker (Matheson, 1990, pp. 46-47). A complex problem can be quickly summarized by an influence diagram without losing the important details necessary to fully understand it.

Similar to a decision tree with respect to the highly graphical representation of a problem and also based on probability laws, influence diagrams have the distinct advantage of compressing a majority of the detail in a decision. This feature allows the influence diagram to enhance communications to the decision maker and other agents of the process. The graphical representation of the diagram incorporates various symbols to illustrate the components of the problem. Symbols are called nodes. They are connected in the diagram to each other by arrows or arcs. Five types of nodes are used along with two types of arcs.

A rectangular node, called a decision node, is representative of a decision point, where the decision maker must choose from a set of well-defined alternatives. Probability nodes are oval in form and represent an event that is uncertain and not under the complete control of the decision maker. These nodes would have a probability distribution associated with them. A double oval represents a variable or event that is a function of all the nodes with arcs pointing into it. Since the outcome of the variable can be determined from the inputs, it is called a deterministic node. The value node, a rounded rectangular node, is a type of deterministic node to which the decision maker has assigned a specific numerical value. If the value node has a double border, it indicates that only the expected value associated with the best decision alternative is of interest to the decision maker. These nodes are expected value nodes.

Arcs leading into a decision node indicate that the information coming in is known at the time that decision is made, and therefore are called information arcs. Conditioning arcs lead into a probability node and indicate that the event depends on the outcome at the decision or probability node from which the arc originated.

When combined together in an influence diagram to illustrate a specific problem, the interaction of the various nodes and arcs become readily apparent and the flow of information can be logically followed to determine the best course of action in making the decision.

Determining Situation Awareness Using Influence Diagrams

Applying the concept of the influence diagram to a given circumstance to be evaluated for situation awareness criteria, the problem must first be defined. Because of the unique quality of the influence diagram to show specific interdependency between events and the uncertainty of other events and decisions, the five categories of situation awareness can be independently evaluated. For example, within the responsibility dimension of situation awareness, an influence diagram can show all the factors that have a bearing on the responsibilities of an operator in a specific environment. The diagram can then be reduced to a conclusion based on these factors. It can be shown how the various conditions, decisions, and events, dealing with this one dimension only, are interfaced with one another and combine to yield the basic decision. An influence diagram shows how one condition or event can affect any number of other elements, unlike the sequential order of a decision tree. By assessing the individual dimensions of situation awareness in separate influence diagrams, one achieves the broadest perspective of the problem. It is now necessary to combine these diagrams into one final representation, eliminating duplicate events or decisions. The result is an influence diagram that has the five individual elements of situation awareness interfaced together and leading to the solution. Effectiveness of the diagram is rooted in the concepts that alternatives can be seen at each decision point, possible outcomes are available for each uncertain event and the likelihood of each outcome can be traced.

The Accident Scenario

In this paper, the model used is based on a fictitious accident. Specific references to places, objects and personnel were included to provide clarity. However, the technical aspects of the signal system, the switch position and operation, along with the regulations are based on factual data. Although this accident has never happened, it is probable for it to occur exactly as described.

A freight train (#275) is heading toward the entrance of Chiefton Intermodal Facility (CIF) which is run by Chiefton Transport, located at a busy East Coast port in Virginia. There are several rail entry points to the facility and the block operator has routed the freight train to track #4. Being unfamiliar with the track #4 entrance, the engineer is particularly alert to the surrounding area.

Running directly parallel to track #4 is the main access highway, separated only by a partitioning chain link fence. Heavy truck traffic frequents this road and seamless attachment to the arterial interstate results in vehicle speeds often in excess of the posted limit of 55 mph. An approaching tractor trailer is rapidly entering the area in the right lane of the access road (see Appendix A).

As the Northbound train enters track #4, the engineer notices a visual light signal adjacent to the track is indicating "restricted". Knowing that a restricted signal means proceed at the reduced speed of 15 mph and be prepared to stop, the engineer reduces the train's speed and continues on the track, looking for the cause of the restriction. Restricted signals are generated by any number of factors, including track work, a broken rail, a bad switch, another train, obstructions, etc. Failing to notice the open switch points at the upcoming switch junction on track #4, the engineer is concentrating on a train (#381 see Appendix A) traveling southbound on a parallel track nearby. He or she concludes that this train was the cause of the restricted signal indication. Accelerating the train, the engineer proceeds through the open switch which instantly derails the engine and the freight cars bellow out to the right as a result of inertia, landing directly on the access road (see Appendix B).

Traffic is light in the Southbound direction on the highway and the approaching tractor trailer driver, already behind schedule, takes advantage of the reduced traffic volume by driving at 75

mph. Upon seeing the derailment ahead, the driver cannot react quickly enough to prevent a collision with the wreckage, which has partially blocked the roadway (see Appendix B).

Probability Distribution of the Model

Selected probability distributions used in the model presented have been assigned to a few variables based on research that concluded their likelihood in contributing to accidents. In the case of alcohol influence, certain key questions can not be answered reliably when searching for statistics about traffic accidents and alcohol usage. One reason for this difficulty is that the data is scattered among many sources; hospitals, police, courts, coroners, and traffic accident and driver licensing data systems (Donelson, 1985). There are, however, enough studies to substantiate the distribution assignment for alcohol in the model.

In the summary of findings from Jones and Joscelyn, 1979, (as cited in Evans and Schwing, 1985), of fatally injured drivers in the United States, 40-55% of the drivers had a blood alcohol content (BAC) exceeding the legal limit. The Traffic Injury Research Foundation (TIRF) of Canada, often cited as a principal source of statistics on alcohol and road accidents, concludes that the usage rate is between 45% and 51% (Donelson, 1985). Evans (1991) states that 47% of fatalities from traffic crashes are directly attributable to alcohol usage. Alcohol is therefore the largest single factor contributing to traffic crash losses (Evans, 1991, p. 188).

Representing one of the primary human factors, driving fatigue is believed to have a powerful effect on commercial vehicle drivers. Fatigue significantly increases driving errors and decreases driver alertness (Lin, Jovans and Yang, 1994). Although fatigue is a sufficiently vague concept and difficult to precisely define, several important studies indicate a probability distribution. The United States Department of Transport and Communications and the National Transportation Safety Board (NTSB) have determined that driver fatigue is responsible for 29% of fatal traffic accidents (Lin, et al.).

Supporting the difficulty in defining fatigue, Taoka (1993) studied the problem from various perceptions including fatigue from drug ingestion and sleep disorders (insomnia, sleep opnea and narcolepsy). His study also differentiated between drowsiness and actually falling asleep. The conclusion of the Taoka study and of recent studies by the NTSB found that driver sleepiness causes 31% of fatal truck accidents (Taoka).

Driving time has proven to be another factor that directly affects fatigue. A time-dependent regression model formulated by Lin et al. (1994) shows the relationship of driving time to survival probability in motor carrier operations. The model concludes that after five hours of driving, the probability of survival from an accident drops to .8 and after ten hours of driving the probability falls to .5. Time of day and driving experience were also analyzed but their effects were not as pronounced as driving hours.

Lin, T., Jovanis, P.P., and Yang, C. (1993) had also formulated a previous time-dependent model based on accident occurrence and off-duty hours. Again, driving time had the strongest direct effect on accident risk. On the basis of these modeling results, Lin et al. determined that there was a 32% higher accident risk with drivers who had the minimum eight-hour, off-duty time between operations. The effect of rest periods to combat the
fatigue element in motor carrier operations is the subject of ongoing work by the University of California in their research reports concerning motor carrier driving risks (Lin, et al.).

Excessive vehicle speed is considered a main cause of accidents. The difference in speed variance has been linked to an increase in overall accident rates (Hajek, Billings, Hoang and Ugge, 1994). Hajek et al. monitored truck traffic over a two-year period and concludes that during the daytime hours, 16% of trucks had exceeded the posted speed limit. They also established that the percentage is slightly higher when given the operational freedom of light traffic volume.

Case studies cited by Summala (1985) of four European countries prior to and immediately after the initiation of speed limits show a significant decrease in fatalities. The savings in fatalities were enormous with 500 people saved every year in Finland alone. If the speed limit had been in place during the 1960-1970 time period, the savings in lives would have amounted to 4,500 people. Summala concluded therefore that these major changes in fatality trends were due to the general speed limits.

In the regression model of highway fatalities (Michener and Tigne, 1992), it is difficult to find statistically significant results for fatalities from the increase in the speed limit to 65 mph. However, the model did show that a higher speed limit is associated with an increase in vehicle miles traveled (VMT). Increasing VMT shows a significant statistical increase in highway fatalities of approximately 3%. This implies that highway fatalities will also rise by 3% as a result of the higher speed limit (Michener and Tigne).

The conclusions about alcohol, fatigue and speed were derived from general transportation statistics involving motor vehicles. Incorporation of rail transit in the model necessitates probability distributions based on available railroad accident data. The terminology used to describe the data is consistent with that utilized by the United States Department of Transportation/Federal Railroad Administration (U.S. DOT/FRA) 1993 (see Appendix C).

Between 1987 and 1992, the largest single contributor to train accidents has been human factors, accounting for 31 to 35% of the total accidents, followed closely by signal and track defects. Total rail accidents are divided into four types, collisions, derailments, highway-rail impacts and others. Of these types, derailments were responsible for 68-70% of all accidents during that time period. Breaking the statistics down in to accidents by type and track classification, we are able to support the data in the model.

Human factors directly contributed to 67% of the accidents in 1992 that occurred on yard tracks and derailments during that same period amounted to 47% of the accidents, again happening on yard tracks. Train operations that were governed by human factors involving general switching rules and the use of track switches were responsible for 47% of the total accidents. The percentage was identical when a switch's mechanical condition caused an accident. An excessive amount of total injuries, 67%, from train accidents were the result of operating practices (human factors - as cited in U.S. DOT/FRA, 1993, p. 41) involving track switches.

The greatest percentage of total fatalities/injuries involving railroad accidents/incidents during 1992 were those that occurred at public highway-rail crossings (92%). Of interest to this model is the fact that 9% of accidents with fatalities/injuries were caused by something that was struck by or ran into a locomotive or its cars (U.S. DOT/FRA, 1993).

Correlating these statistics into the model supports the assigned probability distributions for human factors, switch operations and derailments that would be likely to occur. The engineer's actions in response to the restricted signal indication and his/her apparent inability to observe all the possible causes for that signal, constitute human factors elements. Derailments are a major contributor to rail accidents, along with switch operations and usage. The model shows that the open switch is responsible for the derailment of the train. Location of the accident inside a train yard is supported by the accident rate on yard track caused by derailments and human factors.

In determining probable cause for the December 1993 derailment of an Amtrak (Silver Meteor) train in Palatka, Florida, the NTSB concluded that the engineer failed to maintain full attention to the train's operation and that a combination of prescription and over-the-counter medications contributed to that deficient attention level (NTSB Report, 1991). Again, human factors, fatigue and drugs played a critical role in causing the accident.

Determining Areas of Investigation and Root Causes

In the process of accident investigation, three elements must be considered to determine the probable cause. As shown in Appendix D (Goldman, 1995, p. 50), the elements of human factors, machine and environment all play a determining role. By isolating each element, exploring all possibilities of the element and moving to the next element, you progress toward a conclusion. The majority of conclusions about accident causation tend to fall into the overlap area of the three elements (see Appendix D).

Root cause determination can be made using several decision tools. Both the influence diagram and the decision tree tools serve to identify cause, but the advantage of the influence diagram is paramount. Decision trees show sequential order but they do not display where the key elements should be applied. For example, Appendix E shows a decision tree for the truck in the model. The truck can proceed at 55 mph or 75 mph but there is no reason shown for that decision. Schedule information, as well as mechanical or road conditions is also omitted. Given that traffic is light and the vehicle is speeding, one may conclude that the speed was probably the result of traffic volume. At higher speeds, the driver's reaction time would be less, decreasing the possibility of avoiding the derailed train.

In the train decision tree (Appendix F), the engineer identifies the restricted signal and slows down. It can not be determined from the tree if the engineer is aware of the track's condition or how much stress the restricted signal placed on his or her concentration. Even the probability of fatigue is absent. Illustrating all the other variables into either decision tree schematic would make the model very difficult, if not impossible, to follow to any reasonable conclusion.

The Model

To construct the influence diagram, it is necessary to separate the four components of SA and evaluate each one independently. Both the truck and train elements have been separated for clarity. Only those factors which play a part in the particular component of SA have been included in the respective diagrams.

In Appendixes G and H, navigational awareness is evaluated. Note that the deterministic node for both the train and the truck is the schedule. A variety of probability nodes affect the schedule. The restricted signal in Appendix I shows a major deterministic node under identity for the train. The truck identity diagram, Appendix J, shows no such node. It is important to remember that in separating SA into its elements, not all nodes can be equally represented for a given situation. Responsibility awareness for the truck situation (see Appendix K) shows the schedule as a determinant again and that is not the case with the train responsibility (see Appendix L). The reasoning is that the driver of the truck would have much greater control over his or her ability to affect the schedule. The last element, expectancy awareness, has no deterministic nodes for either category. Appendix M for the train and Appendix N for the truck simply show a reliance upon concentration, confidence and skill.

In all the separate diagrams, many probability nodes are repeated to illustrate that certain factors like fatigue, drugs/alcohol and stress will have a bearing on the outcome no matter how you segregate the SA elements. This demonstrates the ability of the influence diagram to display integration between probabilities, something that is quite difficult to determine from a decision tree.

In constructing the model, it becomes necessary to extract the nodes that are common to each SA element. Combining the common nodes with the remaining nodes yields one diagram representing the dimension of SA. Certain nodes have been consolidated in the model because their relationship to other nodes is identical. For example, the human error node (see Appendix O) includes the operator's reaction time, confidence, concentration and driving habits. This is

permissible because these separate elements have the same influence on a common node. Confidence, concentration, driving habits and reaction time all influence the speed/control decision box, as the human error node indicates. The same consolidation principle has been applied to the familiarization, mechanical, location and congestion nodes. Elements for both the truck and train diagrams are included (see Appendixes G-N).

The model displays two deterministic nodes, two decision boxes and two value nodes. From any probability node, it can be seen how that factor/event influences other factors/events linked to it. Note that both the information arc and the conditioning arc are utilized in the model. Probability assignments for the major factors of fatigue, drugs/alcohol, and human error are based on the statistics presented. The remaining probabilities were arbitrarily selected. It is important to remember that the situation presented makes certain assumptions because it is fictitious and subject to many different factors. Probability variations would account for a different outcome as anything is probable within the model.

Conclusion

When applied to a given situation, with varying probabilities, the influence diagram proves superior in its ability to derive knowledge about uncertain quantities. The diagram can frame a decision and characterize values that govern that decision. No other decision tool possesses its span of application and power of computation. This paper has demonstrated, with graphic clarity, that a given scenario can be analyzed for root cause by the application of this tool. Even the complication of having two separate entities collide as a result of completely different factors does not impede the analysis. The influence diagram continues to prove its usefulness with growing popularity in decision making.

References

Donelson, A. (1985). Between molecule (alcohol) and mayhem (road crashes): The case for humane intervention and the role of social and behavioral sciences. In L. Evans and R.C. Schwing (Eds.), Human behavior and traffic safety (pp. 421-483). New York: Plenum Press.

Evans, L. and R.C. Schwing (Eds.). (1985). Human behavior and traffic safety (pp. 421-483). New York: Plenum Press.

Evans, L. (1991). Traffic safety and the driver. New York: Van Nostrand Reinhold.

Taoka, G.T. (1993). Driver drowsiness and falling asleep at the wheel. Transportation Quarterly, 47, 583-595.

Thierauf, R. and Klekamp, R. (1970). Operations research models. New York: John Wiley and Sons.

Lin, T., Jovanis, P.P., and Yang, C. (1994). Time of day models of motor carrier accident risk. In Transportation Research Record #1467, Traffic and roadway accident analysis and traffic records research. Washington, D.C.: National Academy Press.

Lin, T., Jovanis, P.P., and Yang, C. (1993). Modeling the safety of truck driver service hours using time-dependent logistic regression. In Transportation Research Record #1407, Large vehicle safety research (pp. 1-10). Washington, D.C.: National Academy Press.

Hajek, J.J., Billing, J., Hoang, P., and Ugge, A.J. (1994). Use of weight in motion scale data for safety related traffic analysis. In Transportation Research Record #1467, Traffic and roadway accident analysis and traffic records research. Washington, D.C.: National Academy Press.

Summala, H. (1985). Modeling driver behavior: A pessimistic prediction? In L. Evans and R.C. Schwing (Eds.), Human behavior and traffic safety (pp 43-65). New York: Plenum Press.

Michener, R. and Tigne, C. (1992). A poison regression model of highway fatalities. The American economic review 82,

Goldman, D. (1995). What really happened to cause the accident of incident. ASTM Standardization News, 23, 50-55.

Andre, A.D., Wickens, C.D., Moorman, L., and Boschelli, M.M. (1991). Display formatting techniques for improving situation awareness in the aircraft cockpit. The International Journal of Aviation Psychology, 3, 205-218.

Matheson, J.E. (1990). Using influence diagrams to value information and control. In R. Oliver and J. Smith (Eds.), Influence diagrams, belief nets and decision analysis (pp. 25-48). New York: John Wiley and Sons.

Howard, R.A. (1990). From influence to relevance to knowledge. In R. Oliver and J. Smith (Eds.), Influence diagrams, belief nets and decision analysis (pp. 25-48). New York: John Wiley and Sons.

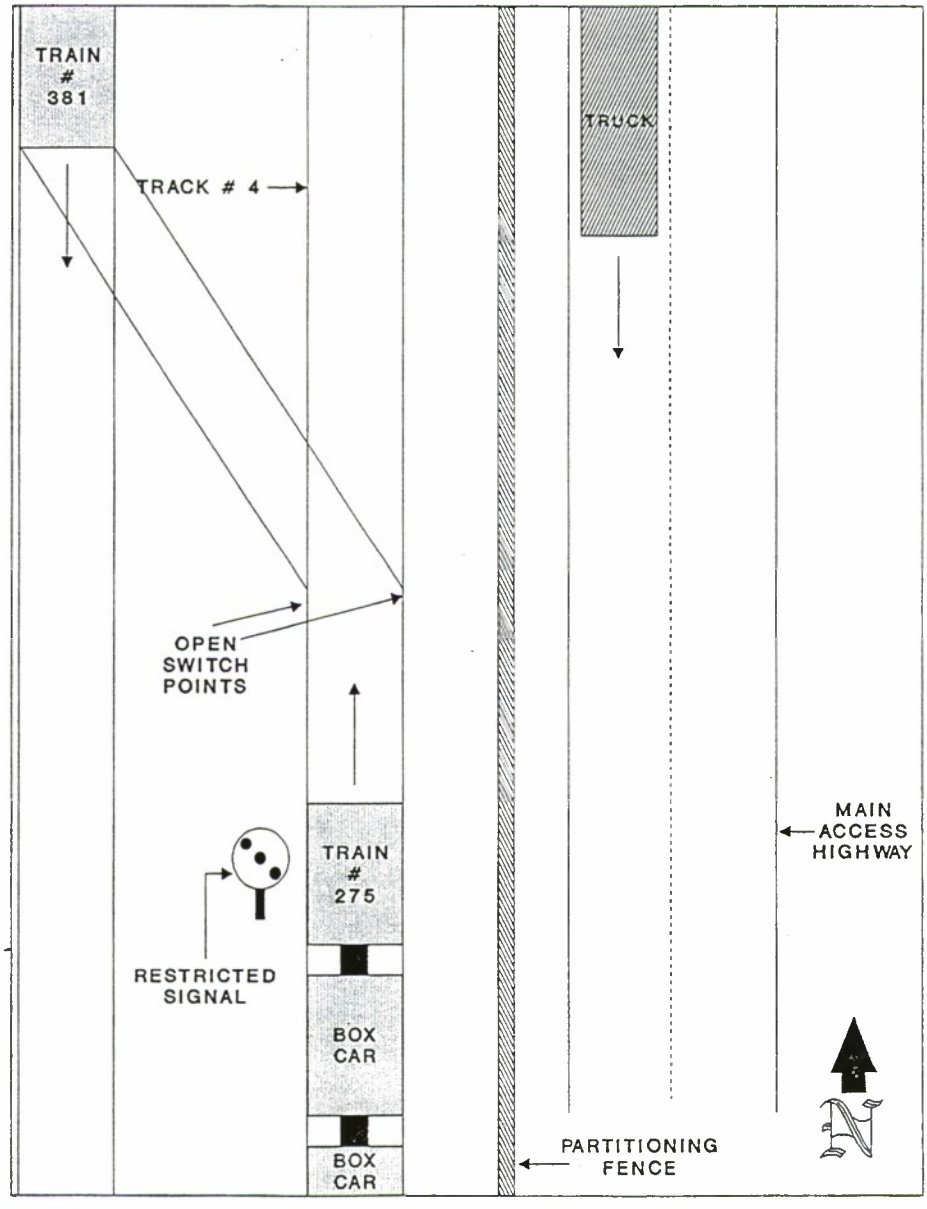
U.S. Department of Transportation/Federal Railroad Administration. (1993).

Accident/Incident bulletin #161, calendar year 1992. Washington, D.C.: Federal Railroad Administration.

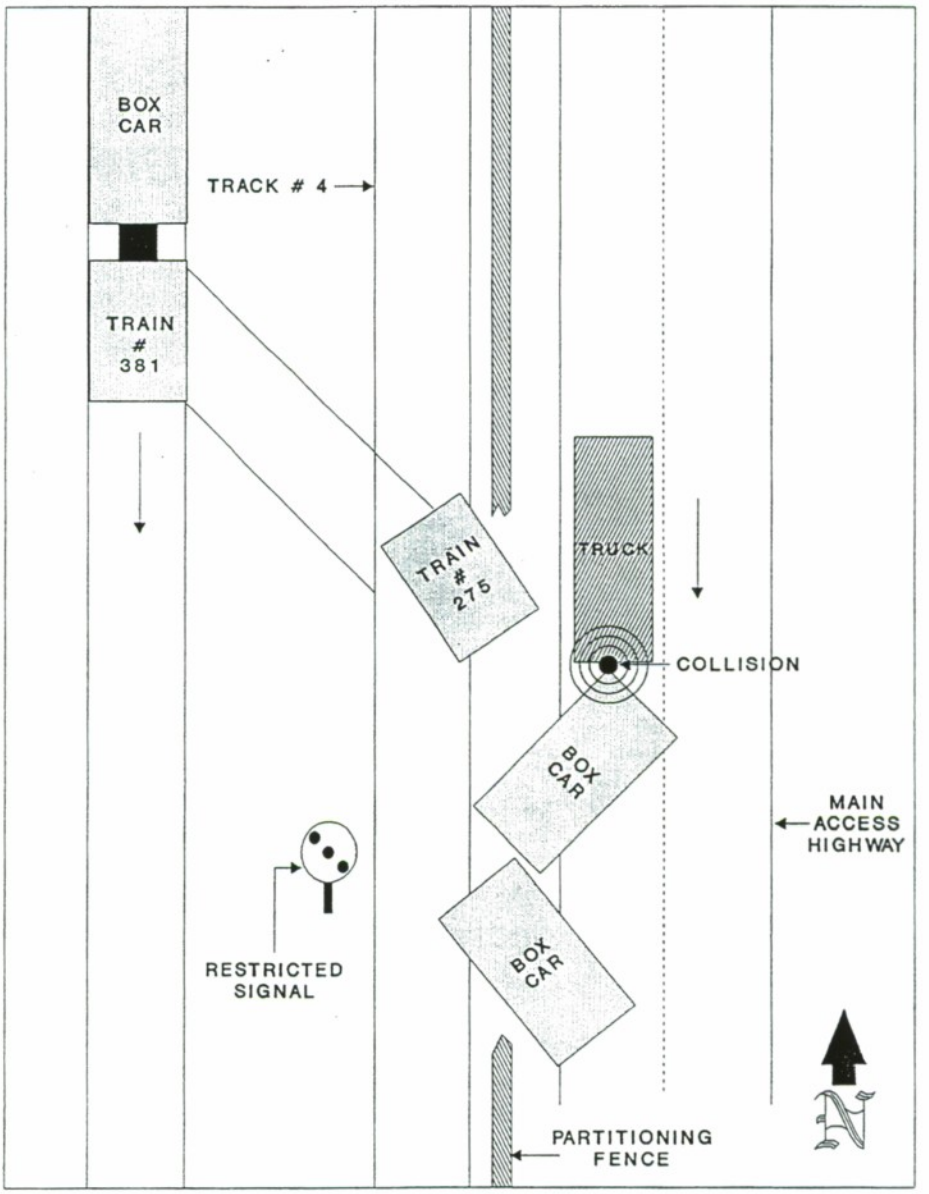
National Transportation Safety Board report. (1991). Derailment of Amtrak train 87, Silver meteor in Palatka, Florida, December 17, 1991. Washington, D.C.: National Technical Information Service.

Jones, R. and Joscelyn, K. (1979). Alcohol and highway safety: A review of the state of knowledge. (National Highway Traffic Safety Administration technical report No. DOT-HS-803-714). Washington, D.C.: National Technical Information Service.

Appendix A


TRAIN
#
381
TRACK # 4
TRUCK
OPEN
SWITCH
POINTS
TRAIN
#
275
RESTRICTED
SIGNAL
BOX
CAR
BOX
CAR
MAIN
ACCESS
HIGHWAY
PARTITIONING
FENCE

Appendix B


BOX CAR
TRACK # 4
TRAIN # 381
TRAIN # 275
TRUCK
COLLISION
BOX CAR
MAIN ACCESS HIGHWAY
RESTRICTED SIGNAL
BOX CAR
PARTITIONING FENCE
N

Appendix C

Classification of Accidents/Incidents

Train accident. A collision, derailment or other event involving the operation of railroad on-track equipment resulting in damages that exceed the reporting threshold.

Train incident. Any event involving the movement of railroad on-track equipment that results in a death, a reportable injury, or a reportable illness, but in which railroad property damage does not exceed the reporting threshold.

Definitions

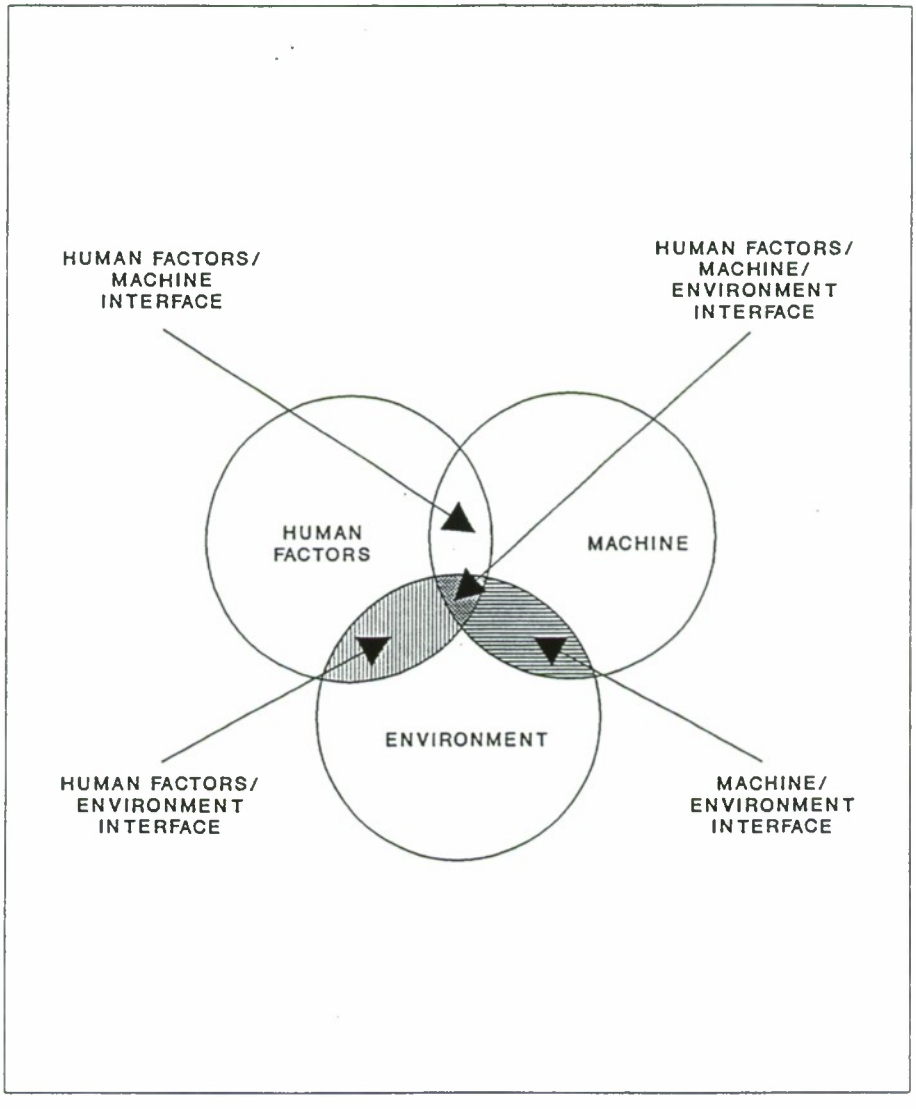
Derailment. A derailment occurs when one or more than one unit of rolling stock equipment leaves the rails during train operations for a cause other than collision, explosion or fire.

Human Factors. Behavior affecting elements of railroad employee job performance.

Reporting Threshold. The level of railroad property damage, resulting from a train accident involving on-track equipment, over which a railroad company must report the accident to the Federal Railroad Administration.

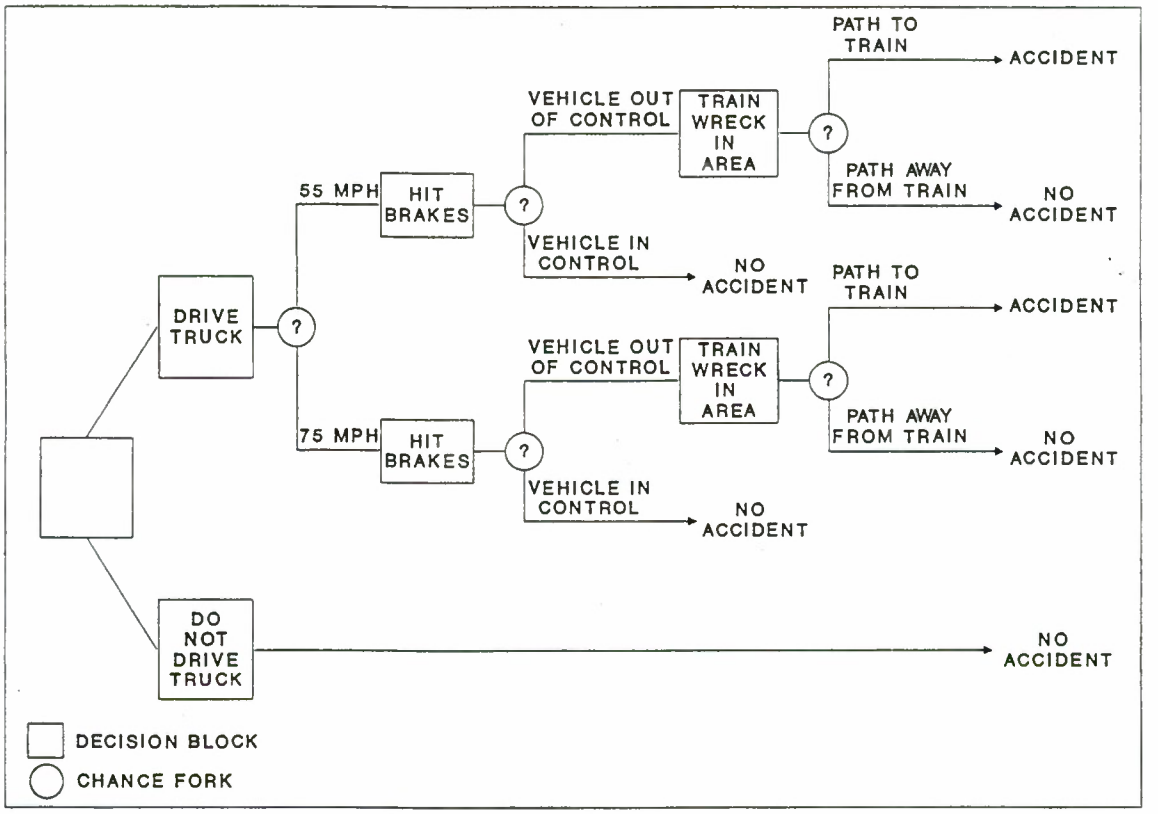
Note: The authors disagree that human factors is behavior (U.S. DOT/FRA terminology) since that implies behavior should be adjusted for a better outcome. It is our belief that human factors is performance, defined as the knowledge, skill and ability to do the job. The machine needs to be readjusted to provide increased performance.

Appendix D. Accident investigation areas.



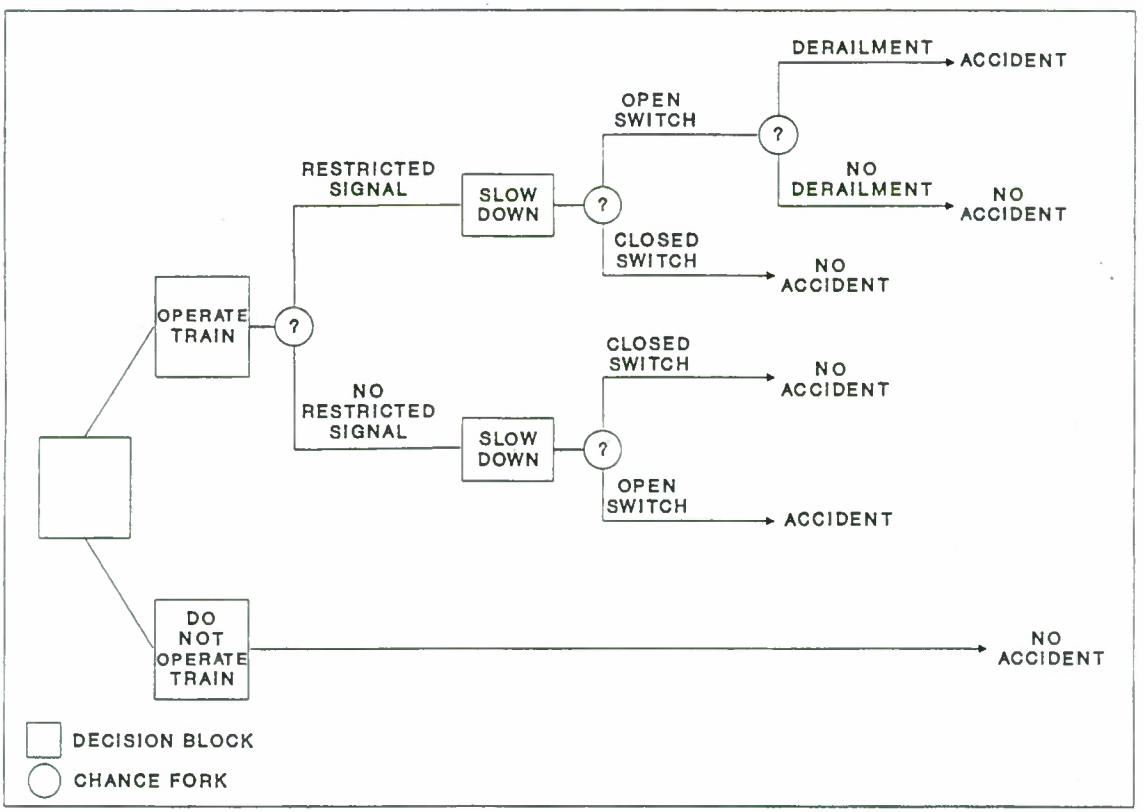
NOTE: CHART REFERENCE, GOLDMAN, 1995, p50.

Appendix E. Decision Tree Model - Truck.



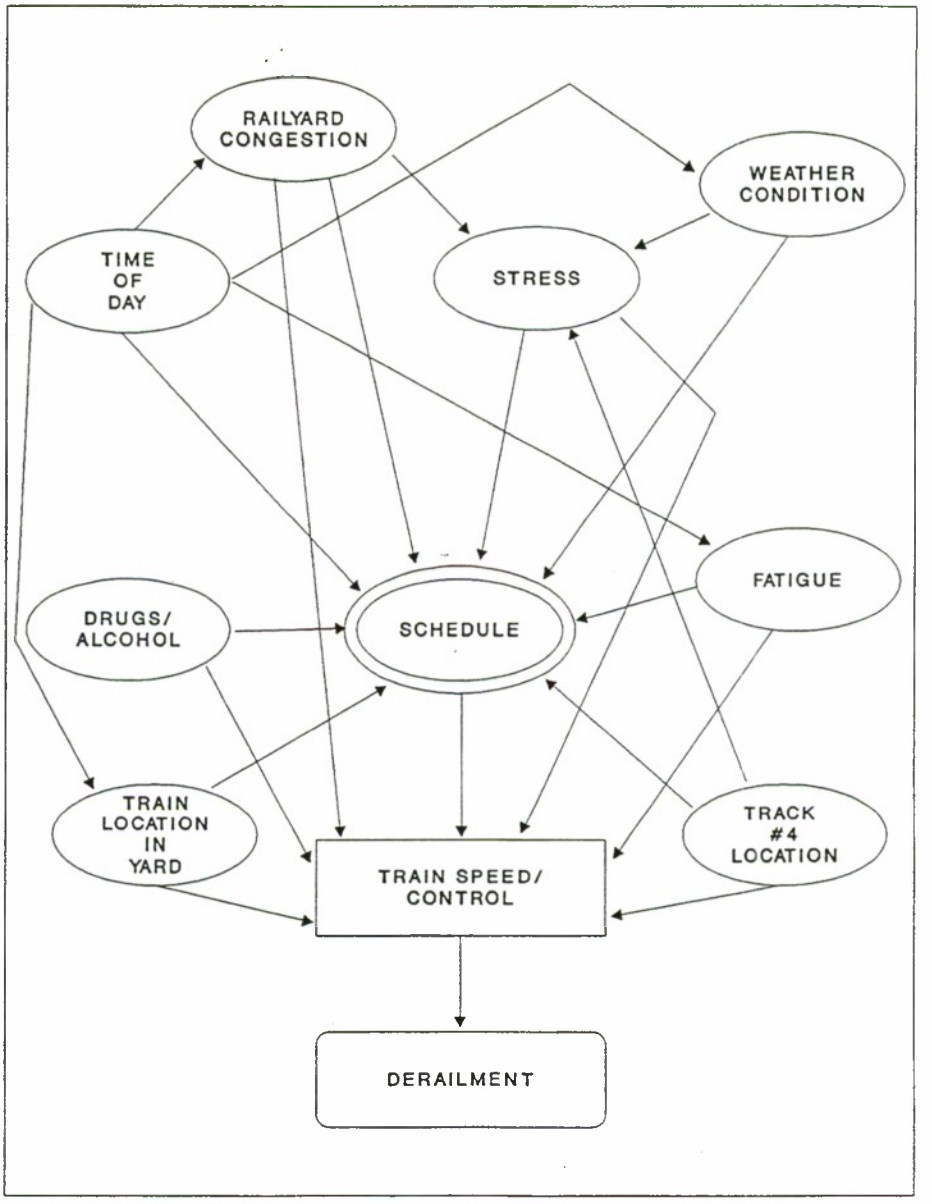
NOTE: MODEL REF, GOLDMAN, 1995, p. 54.

Appendix F. Decision Tree Model - Train.

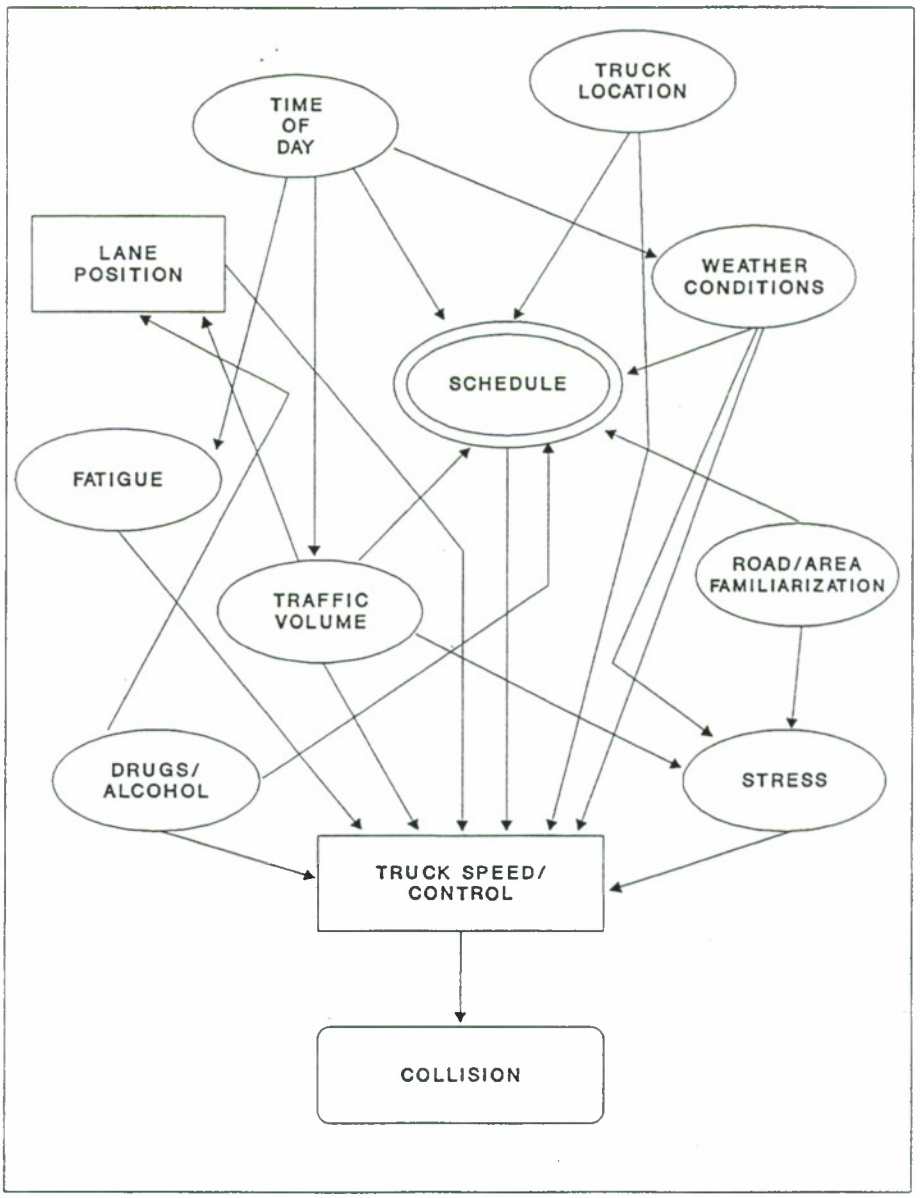


NOTE: MODEL REF, GOLDMAN, 1995, p. 54.

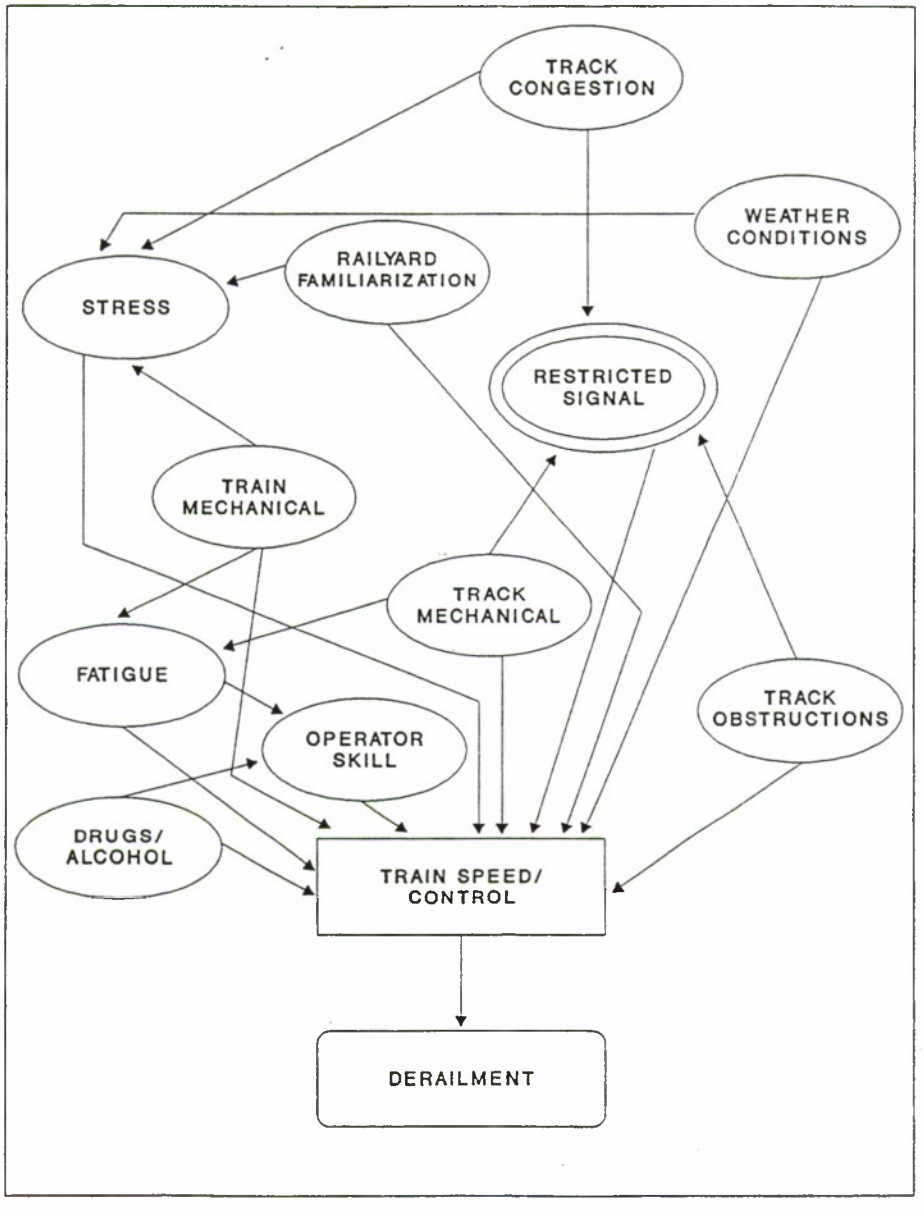
Appendix G. Navigational Awareness - Train.



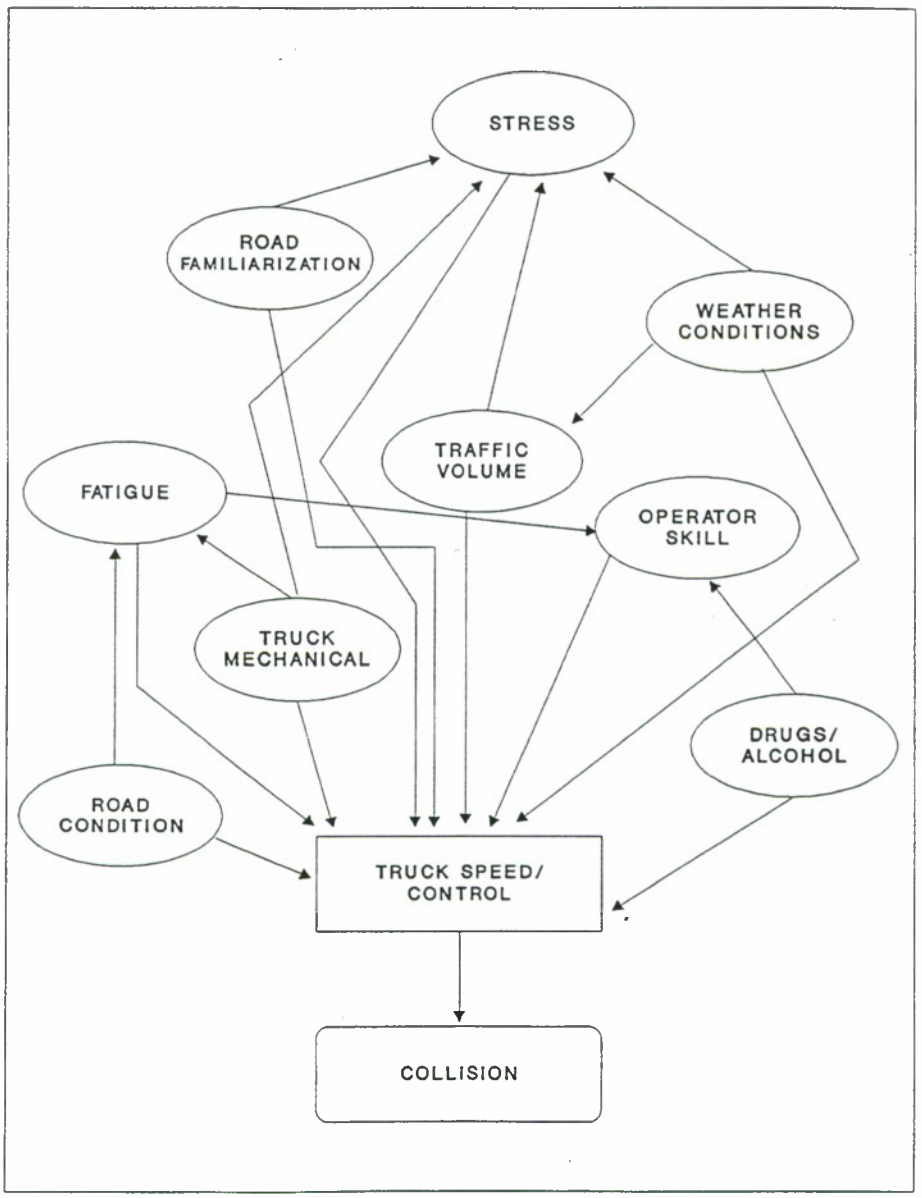
Appendix H. Navigational Awareness - Truck.



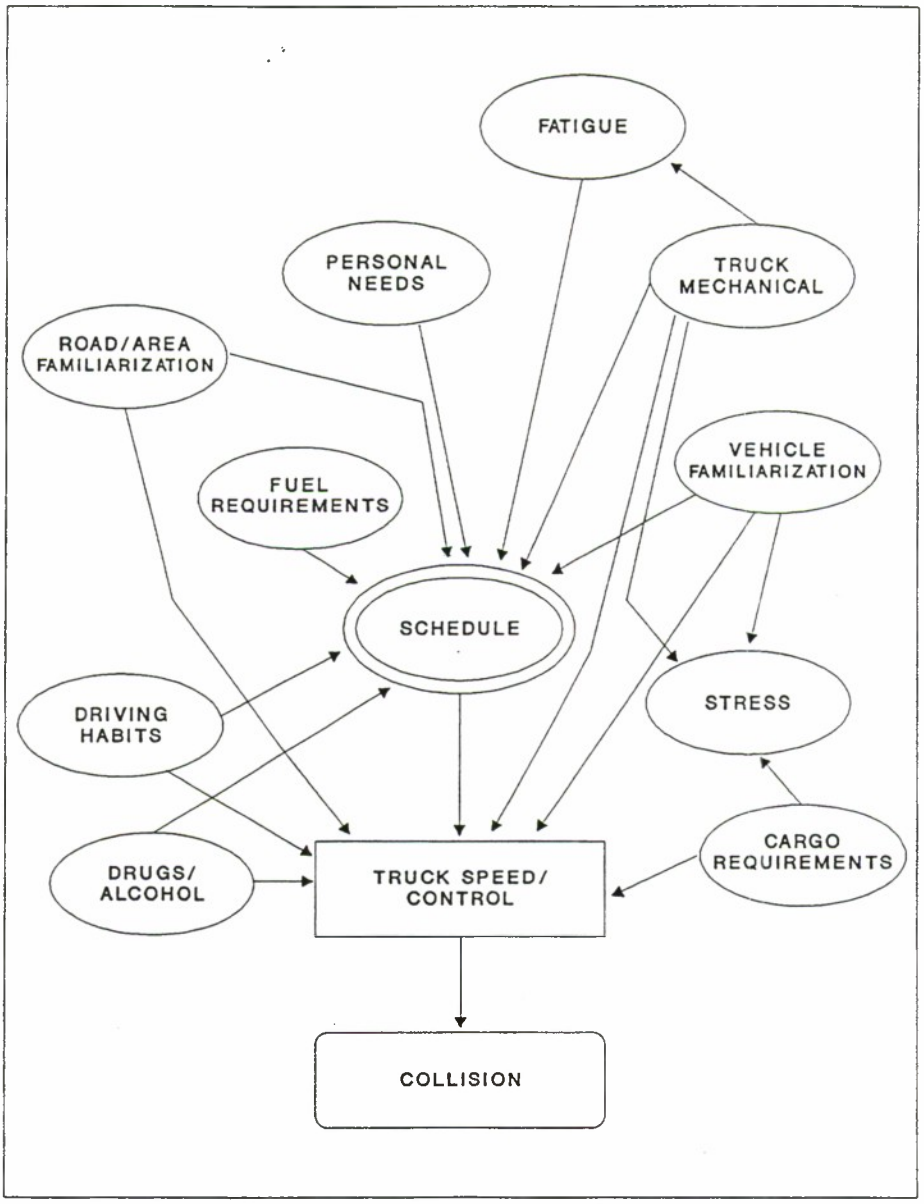
Appendix I. Identity Awareness - Train.



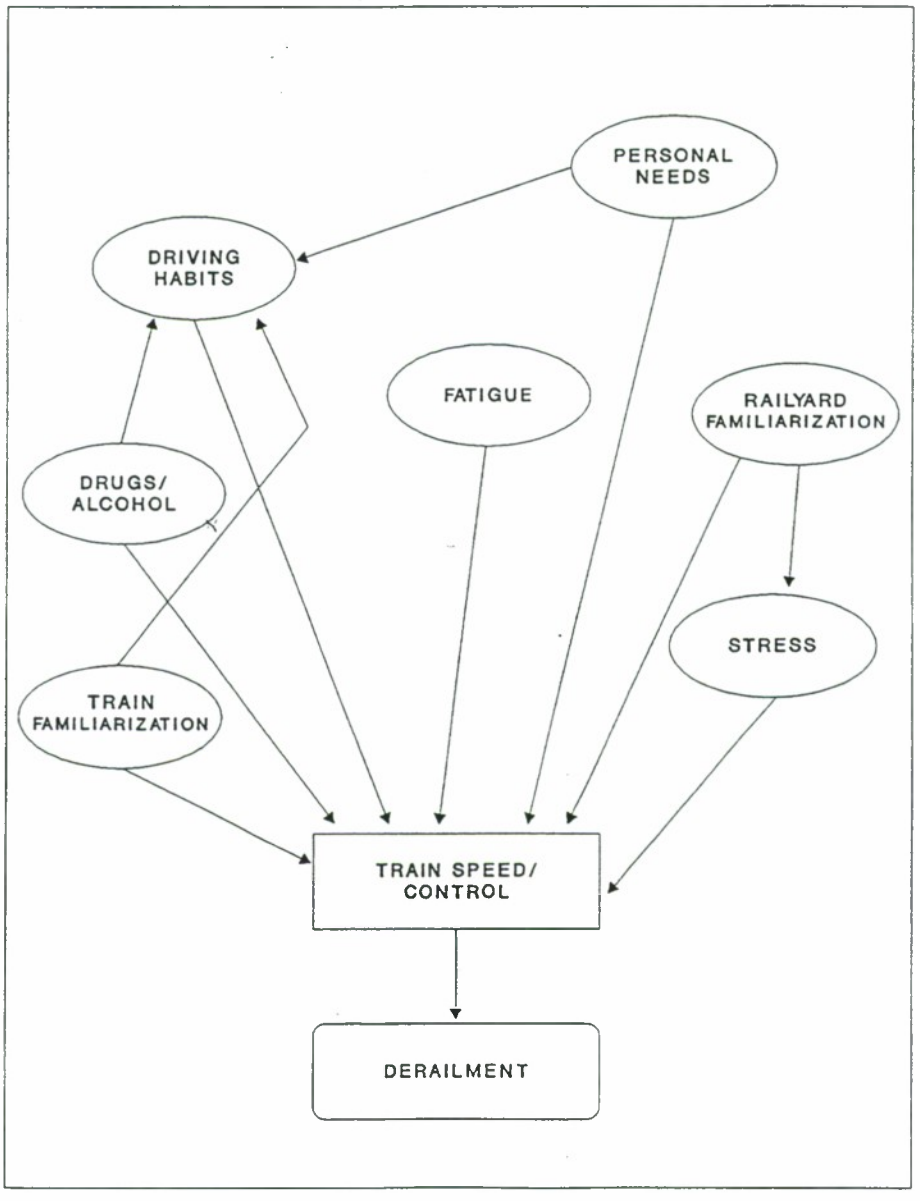
Appendix J. Identity Awareness - Truck.



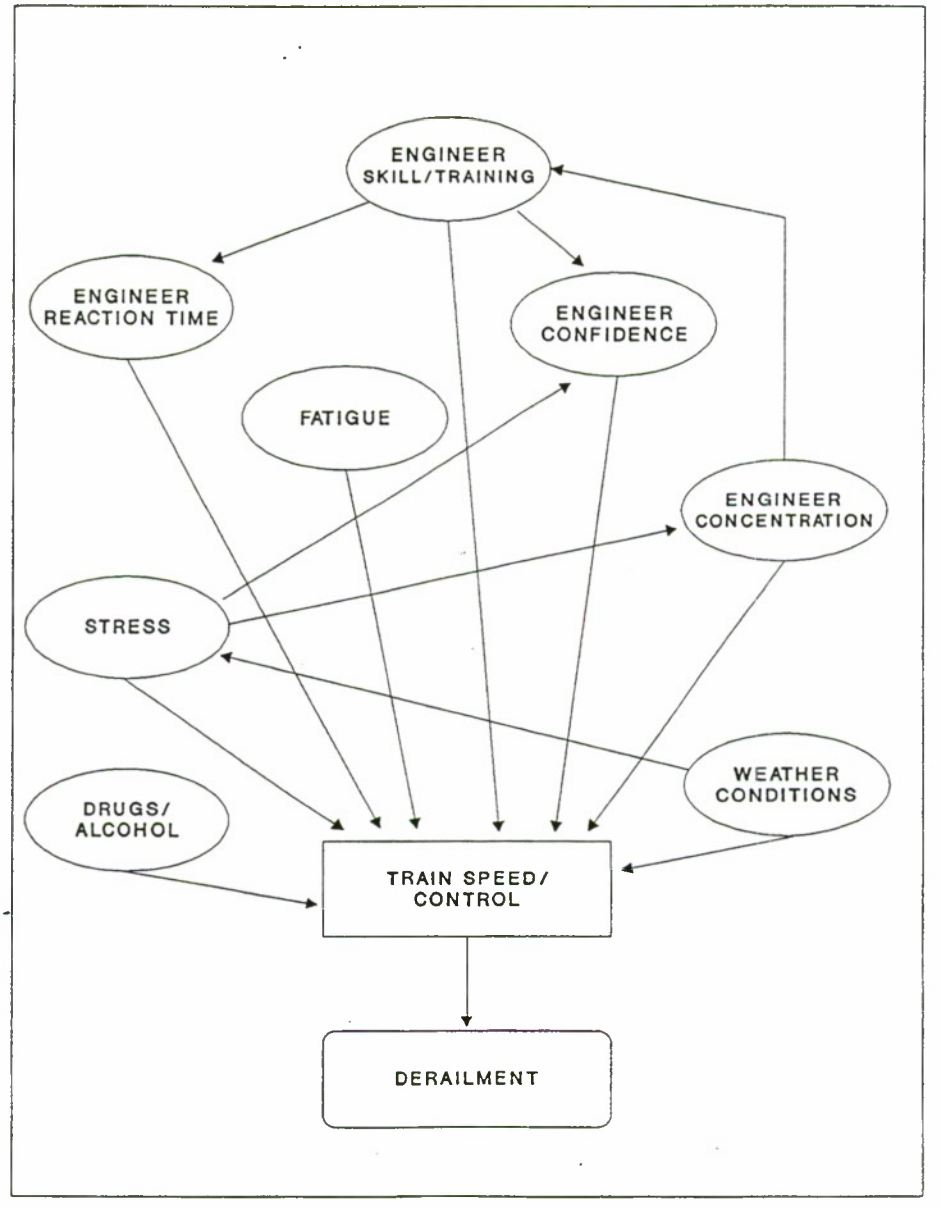
Appendix K. Responsibility Awareness - Truck.



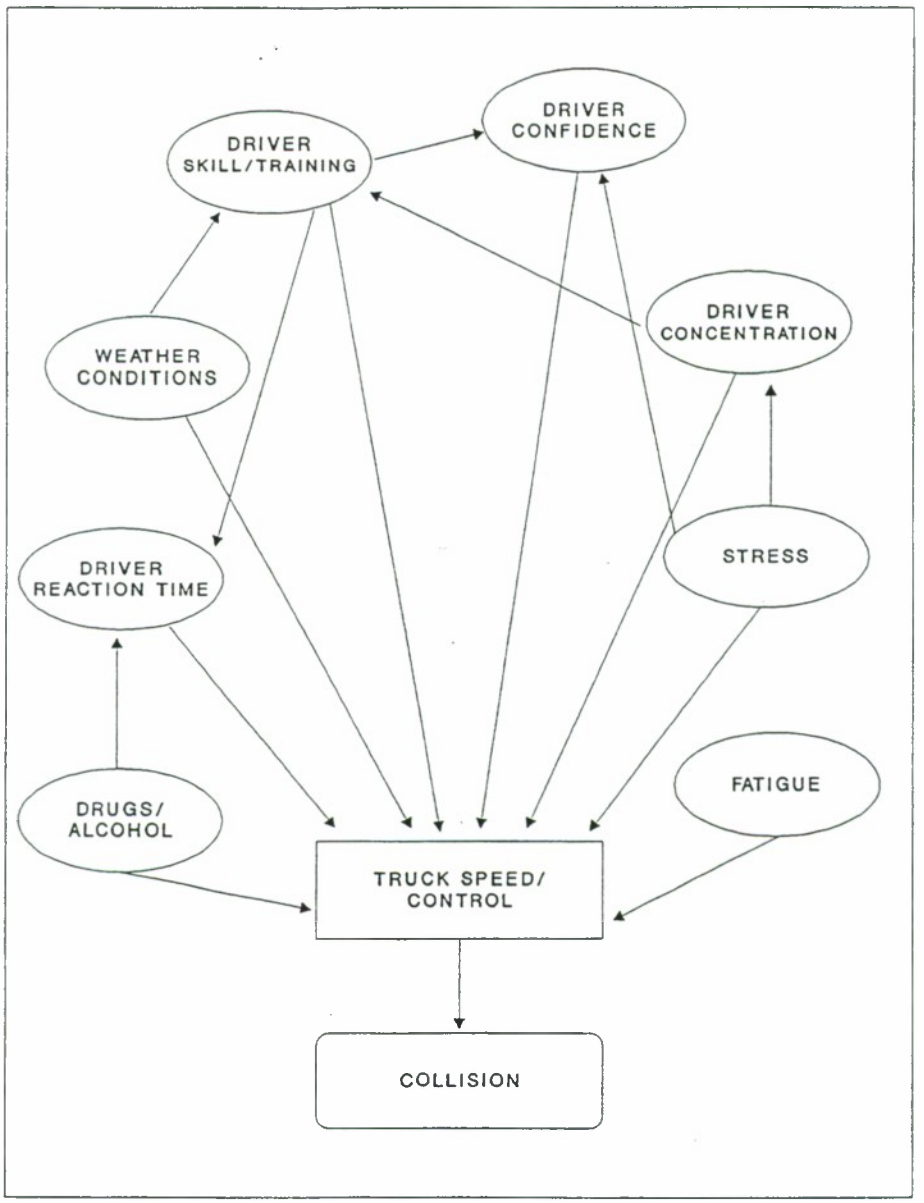
Appendix L. Responsibility Awareness - Train.



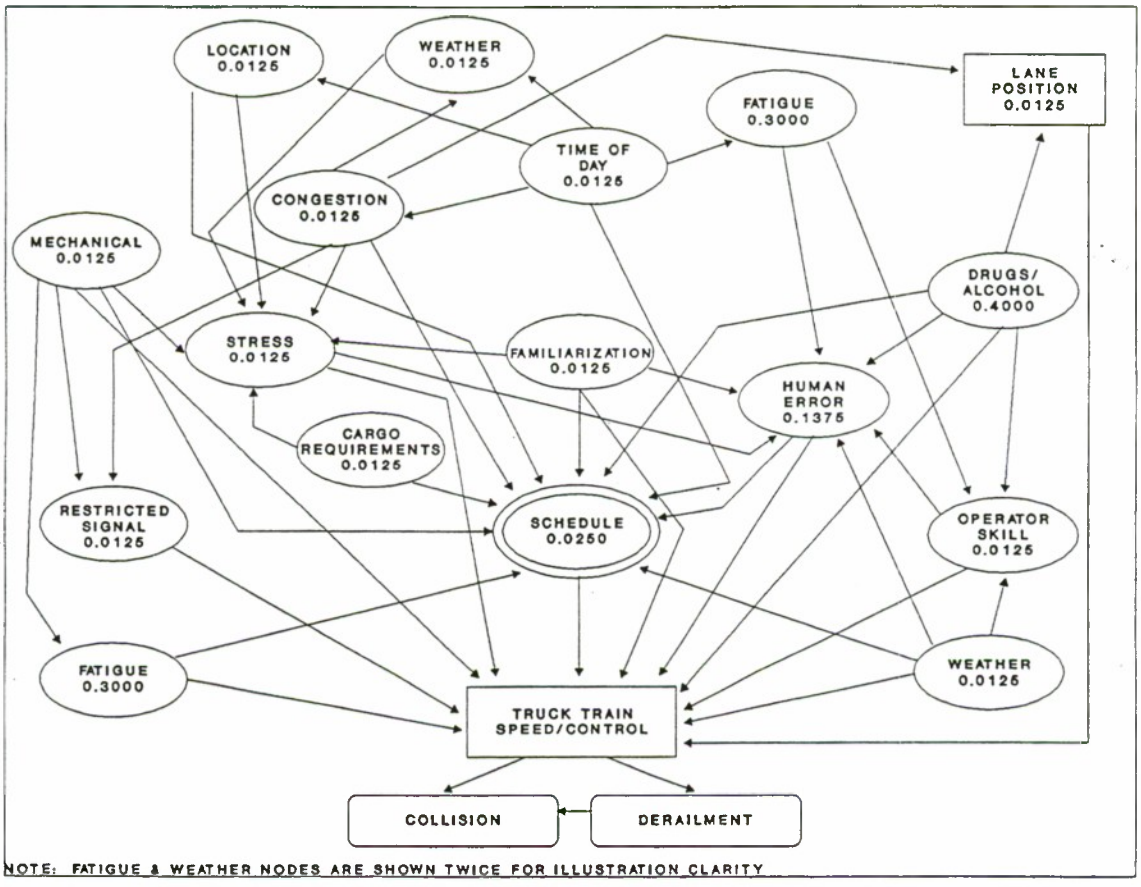
Appendix M. Expectancy Awareness - Train.



Appendix N. Expectacy Awareness - Truck.



Appendix O. Combined influence diagram.



Give Me Situation Awareness, or Give Me Death

Michael Eidelkind[1], Raymond Moffett[2], Don Arendt[3], and Charles McKee[1]

[1] BDM
[2] PRC Inc.
[3] Hi-Tec Systems

Abstract

Combat aviation is an extremely dangerous discipline. In order to survive, aircrews must remain situationally aware, relying heavily on their cockpit displays. Even a moment's distraction could prove fatal. This paper presents situation awareness findings collected through a methodology that blended performance results with subjective workload and priority assessments during an operational test of the recently developed Longbow Apache Attack helicopter. The Mission Assessment Technique, a new means for measuring aircrew workload and priority schemes, is introduced.

Introduction

In the combat aviation domain, aircrews must continuously react to numerous task and environmental factors in order to successfully fulfill their mission goals. To survive on today's battlefield, crews require as much environmental and mission information as possible, without being overburdened by the means of obtaining it. Additionally, they must be able to communicate with each other and coordinate strategy with team members. Studying situation awareness is only meaningful when correlated with improved system performance. To this end, a methodology was developed that combines quantitative and qualitative data from several sources in an effort to determine situational factors that affect performance.

Salas, Prince, Baker, and Shrestha (1995) considered situation awareness a cyclical perceptual process. They stated that "situation awareness occurs as a consequence of an interaction of an individual's preexisting, relevant knowledge and expectations; the information available from the environment; and cognitive processing skills that include attention, allocation, perception, data extraction, comprehension, and projection. This results in an increase in the individual's knowledge, a change in expectations, and another cycle of information extraction."

Team situation awareness, applicable to the combat aviation domain under investigation (Cannon, Bowers, and Salas, 1990) is more difficult to define than individual situation awareness because it not only embraces the many factors presented above, but is also further complicated by the need for coordination and information sharing (Schwartz, 1990; Endsley, 1995). Salas, et al (1995) suggested that to help understand team situation awareness, quantifiable indicators and a measurement scheme need to be devised.

Endsley (1995) stated that the "safe operation of the aircraft in a manner consistent with the pilot's goals is highly dependent on a current assessment of the changing situation, including details of the aircraft's operational parameters, extemal conditions, navigational information, other aircraft, and other hostile factors. Without this awareness (which needs to be both accurate and

complete), the aircrew will be unable to effectively perform their functions. Indeed, ...even small lapses in situation awareness can have catastrophic repercussions." The "other hostile factors" that Army aviators must continually face are enemy forces bent on shooting them out of the sky. Situation awareness keeps the crews alive.

Test Aircraft

The Apache has been the state-of-the-art attack helicopter for the U.S. Army since its incorporation into the fighting force a decade ago. The Longbow Apache was recently developed to dramatically improve the Apache's performance which included replacing a myriad of analog and digital readouts with "glass cockpit" multi-function monochrome displays. The advanced millimeter wave radar system and cockpit interfaces were designed to provide the crews with improved terrain awareness and precise enemy and friendly vehicle position knowledge in a variety of battlefield and weather conditions. Information displayed to the two crew members, together with intra- and inter- helicopter communication and battle coordination capabilities, is intended to facilitate cockpit and team resource management to fulfill the mission goals of neutralizing the enemy, eliminating friendly casualties due to enemy fire and fratricide, and increasing the chances of crew and system survivability.

Test Conditions

During March, 1995, a trained U.S. Army Longbow company was tested against a similar company of Apaches in separate back-to-back trials against an operationally realistic battalion of threat forces. The crews were selected from the 2-229th Attack Helicopter Regiment, Fort Rucker, Alabama. Sixteen members of A Company, 2-229th made up the Longbow Apache company, and sixteen members of B Company made up the control Apache company.

This test represented the last data collection point prior to the acquisition decision for the Longbow system, having already completed controlled technical tests. The U.S. Army performed this test primarily to calculate the effectiveness and suitability of the improved helicopter with trained crew members to fulfill the mission goals mentioned above. The authors of this paper were part of the team responsible for analyzing the operational effectiveness of the AH-64D Longbow Apache helicopter during the operational test. The thesis for this paper was conceived during the test, because the tactics, techniques, and procedures used by the Longbow crews were shaped by the advanced battlefield visualization capabilities. The addition of the new radar system and cockpit displays enhanced their performance and greatly aided in survivability. Additional evidence was noted due to the unique situation of also testing Longbows with the upgraded weapon systems, but lacking the radar system that initiates most of the improved situation awareness.

Experimental Procedure

The Apache (control group) and Longbow (experimental group) teams performed fifteen comparative missions (12 night and 3 day trials) against similar enemy arrays. The test separately pitted four to six Longbow and Apache helicopters against a battalion of enemy ground vehicles. The enemy arrays, as well as whatever friendly ground vehicles included in the trials, were all actual military equipment. Whenever possible, Soviet-built equipment and other vehicles were used. Unavailable vehicles were replaced with similar-looking American-made vehicles. The initial formations and movement and firing strategies were all consistent with understood former Soviet doctrine. The only aspect of the battles which was simulated was the weapon delivery. Missiles and other weapon engagements (specific to the weapons present on each type of player) were simulated by laser pulses for line of sight, direct fire weapons and geometric engagement

system for non-line of sight, indirect fire weapons. Both lasers hitting receivers mounted on all players and geometric engagement were assessed in real-time by the Real-time Casualty Assessment (RTCA) System. Assessments, including total kill, hit but survive, and miss, among others, were based on mathematical models of each weapon's capabilities, survival capabilities of receiving vehicles, ranges, and other factors.

The missions were observed in real-time by the experimenters on the Video Information Processing System (VIPS). The VIPS is a computerized tool, receiving inputs from the RTCA system, and presenting the data on a large screen viewer. Additionally, radio traffic on the helicopter and threat radio nets was monitored while observing the missions. Following all trials, the Longbow and Apache crews (pilots and copilot/gunners) participated in After Action Reviews (mission interviews) to discuss factors contributing to the mission's success or failure. The questions for the interviews focused on specific events that occurred during the mission.

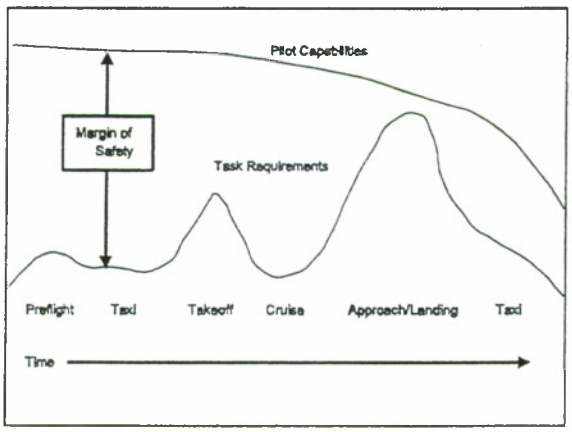


Figure 1. FAA Task Demand/Capability Curve

Mission Assessment Technique

The FAA, in a series of studies performed on aeronautical decision making, showed a notional relationship between pilot capabilities and task demands across the phases of flight (General Aviation Safety Report, 1994). While it is difficult to quantify these on a continuous scale, the idea is useful. Accidents, pilot deviations, and other errors are discrete events in which the requirements of the aviation task exceed pilot capabilities. This is graphically represented in Figure 1.

In order to measure this relationship during the operational test the Mission Assessment Technique was developed by the authors. The Mission Assessment Technique provides a method of defining the situation faced by the aircrew and a means of collecting information on individual and team workload, crew interaction in terms of focus and mutual support, and a comparison between different aircraft or equipment in the same or similar situations.

In order to define the situation, we first specified what we thought were the two major components of situation - environment and function. The environment was further broken down into mission phases - takeoff through landing. The test controls provided like environmental

factors such as light level, battle activity, and weather. Specifically, similar environments were faced by aircrews of both types of aircraft. Since it was desired to collect data within a range of high to low activity and tempo, a generic set of mission phases was used (see phases in Table 1). For the purposes of this assessment, takeoff and landing phases were deleted as being outside the scope of interest.

Functions were defined as sets of tasks required of the aircraft and aircrew. Tasks were defined as a set of requirements and behaviors which are cognitively represented as a unit and have a logical beginning, middle, and end. Most flight-related tasks are composed of subtasks which have a meaningful sequential relationship. Task priority may change depending on the situation and the importance of other tasks that must be completed during the same interval. The basic function set was derived from the set of Mission Essential Functions developed for system reliability assessment. This also allowed for a tighter relationship between human factors, system reliability, and system performance analyses. The function element helped to define the situation in terms of task demands.

The matrix of tasks by mission phase was thought to best define the situation faced by the aircrew (see Table 1). A form was completed for every mission flown to further limit the environmental and task loading factors reflected in the data. The assessment technique had two major components:

- Priority
- Perceived absolute workload

The Priority portion of the data helped to establish the focus of the aircrew both individually and collectively. It produced data for analyzing operators' over- or under-focus on a particular task in light of other events, such as mission performance and survivability. It also allowed some insight into task distribution among crew members and mutual support in light of occurrences during the mission. Aircrews ranked the tasks in priority from 1 (highest priority) to 7 (lowest possible priority) for each mission phase. Ties were not allowed, but crews could mark tasks not applicable if not performed during the respective phase (in which case the lowest ranking would be less than 7).

The Workload portion of the form provided a unidimensional scale for rating perceived workload (from 1 to 7, with 1 meaning minimal workload). A unidimensional scale was used since the primary purpose of this study was to observe the contribution of varying levels of workload to situation awareness and not to provide a rigorous study of workload factors. Similar to the priority rankings, crews could rate tasks not applicable if not performed during the respective phase.

Table 1. Mission Assessment Matrix

Task	Phase			
	Ingress	Release Point to Battle Positions	Battle Positions	Egress
Fly				
Navigate				
Crew Coordinate				
Team Coordinate				
Target Detect/Identify				
Target Engage				
Survive				

Method

The methodology utilizes four different measurement techniques, including subjective workload and priority scheme assessments, performance measurements, and performance comparisons to control groups. As this study was derived from an operational test of the Longbow Apache helicopter, workload levels were only important where correlation to performance results could be established.

Subjective data were collected using operator post-training, usability, and post-test questionnaires. This information was integrated with performance data and analyzed in the context of their operational impact. Operators also filled out the workload and priority scheme matrices. Individual ratings and comparisons of ratings of the same task between pilot and copilot/gunner were used to analyze differences in cockpit resource management and situation awareness between the Apache and the Longbow.

Traditional statistical analyses could not be performed on the results since non-applicable ratings were just as important findings, and could not be dismissed as not part of the total sample. Mean workload ratings for both crewmembers and aircraft were determined, and non-statistical comparisons were made. A function priority scheme was established by looking at the mode of each ranking (from 1 to 7, and not applicable). The Copilot/Gunner and Pilot priority schemes were compared to determine if cockpit resource management took place.

Analysis

Based on the performance data used to determine hits, misses, and kills, the Longbow was determined to be superior to the Apache. The Longbow significantly outperformed the Apache, increasing hits on the enemy and improving aircraft survivability. Crews stated that their performance as a team improved when they implemented the situational information provided by the cockpit displays. They eventually determined how to set up helicopters in head-to-head positions, surrounding the Engagement Area without fear of accidental fratricide.

> "...we were able to place teams in positions that were head to head without a danger of teams shooting each other."
> -Longbow Apache Crewmember

The new cockpit multi-function displays, although giving the crews a great deal of information to look at, actually contributed to reducing their workload. Longbow crews consistently rated their workload lower than the Apache crews on similar functions, throughout the four phases (see Table 2). Flying by the Copilot/Gunners was the only function rated higher by the crewmembers. However, this task was only rated by two Longbow Copilot/Gunners during the first two phases, and just one Copilot/Gunner during the Battle Position (where the mean rating increased to 6.0) the rest rated this task N/A.

Table 2a. Copilot/Gunner Mean Workload Ratings by Phase and Function

Function	Ingress		Release Point to Battle Position		Battle Position		Egress	
	Longbow	Apache	Longbow	Apache	Longbow	Apache	Longbow	Apache
Fly	2.50	2.24	3.00	2.24	6.00	1.58	2.11	2.57
Navigate	1.47	2.97	1.78	3.75	1.17	2.25	1.60	2.82
Crew Coordinate	1.71	2.84	1.81	3.39	2.01	3.52	1.78	2.76
Team Coordinate	1.81	2.73	2.14	3.40	2.58	3.78	2.16	3.08
Target Detect/ Identify	1.80	2.10	2.49	3.07	2.69	4.16	2.03	2.44
Target Engage	1.36	2.47	2.08	3.09	2.56	4.22	1.46	2.47
Survive	1.94	2.44	2.43	3.05	2.69	4.27	2.00	3.29

Shaded cells indicate higher workload rating by phase and function

Table 2b. Pilot Mean Workload Ratings by Phase and Function

Function	Ingress		Release Point to Battle Position		Battle Position		Egress	
	Longbow	Apache	Longbow	Apache	Longbow	Apache	Longbow	Apache
Fly	1.70	2.91	2.23	3.52	1.98	3.71	2.00	3.29
Navigate	1.30	2.15	1.48	2.74	1.36	2.03	1.44	2.53
Crew Coordinate	1.76	2.51	2.00	2.73	2.04	2.95	1.90	2.50
Team Coordinate	1.61	2.62	1.96	2.91	1.97	3.17	1.78	2.56
Target Detect/ Identify	1.55	1.68	1.98	3.34	2.30	3.75	1.73	2.13
Target Engage	1.13	1.67	1.86	3.12	1.88	4.15	1.29	2.37
Survive	1.87	2.29	2.21	3.47	2.61	4.45	2.17	3.02

Shaded cells indicate higher workload rating by phase and function

The mission function priority rankings showed a trend towards crew task sharing. As can be seen in Table 3, flying was most often ranked first by Pilots, but was most often ranked not applicable by the Copilot/Gunners. Crew and team coordination and navigation were the only functions ranked as priorities by both crewmembers during all four phases. Interestingly, target detection, engagement, and survival were all prioritized by both crewmembers during the battle position phase, contributing evidence to the task-sharing trend.

Table 3. Function Priority Ranking Mode by Phase (Longbow Only)[1,2]

Ranking	Ingress		Release Point to Battle Position		Battle Position		Egress	
	CPG	PLT	CPG	PLT	CPG	PLT	CPG	PLT
1	Navigate	Fly	Navigate	Fly	Survive	Fly	Navigate	Fly
2	Crew Coordinate	Navigate	Crew Coordinate	Navigate	Target Engage	Target Detect/ Identify	Crew Coordinate	Navigate
3	Team Coordinate	Crew Coordinate	Team Coordinate		Target Detect/ Identify	Target Engagement	Team Coordinate	Crew Coordinate
4		Team Coordinate		Crew Coordinate	Crew Coordinate	[3]Survive		Team Coordinate
5				Team Coordinate Survive	Team Coordinate	Crew Coord		Survive
6					Navigate	Team Coord		
7						[3]Navigate		

1) Ties of mode ranks are denoted by joined cells; ranking is the higher of the two.
2) Functions with modes of not applicable are not shown on the table.
3) Ranking is one value higher than location on the table.

Conclusion

Although the significant improvement in performance results cannot be attributed directly to increased situation awareness, aircrew performance assessments and interview comments indicated the great impact it brought to their success. Specifically, their situation awareness provided by the sensors and displays kept them alive, in the face of an extremely dangerous enemy.

This situation awareness measuring methodology, because it is based upon a blend of traditional human factors measurement techniques and correlation to system performance measurements, greatly enhanced the Army's ability to understand the impacts of emerging technology. This methodology could be easily applied to other civil and military aviation, as well as non-aviation, applications. Future experiments in more controlled environments should be devised that test the results and conclusions presented in this paper.

References

Cannon-Bowers, J.A., and Salas, E. (1990, April). *Cognitive psychology and team training: Shared mental models in complex systems.* Paper presented at the Fifth Annual Meeting of the Society for Industrial and Organizational Psychology, Miami, FL.

Endsley, M.R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors, 1, 32-64.

Joseph T. Nall General Aviation Safety Report (1994). 1994 Accident Trends and Factors. Frederick, Md: AOPA Air Safety Foundation.

Salas, E., Prince, C., Baker, D.P., and Shrestha, L. (1995). Situation Awareness in Team Performance: Implications for Measurement and Training. Human Factors, 1, 123-136.

Schwartz, D. (1990). Training for situational awareness. Houston, TX: Flight Safety International.

Situation Awareness

Francis S. Bennett

Embry-Riddle Aeronautical University, Prescott, AZ

Abstract

This paper will discuss an account of situation awareness training in a Human Factors in Psychology course.

Students use an "Awareness Profile" program written for use in the Lotus spreadsheet program. Three times a week the students score the behaviors that they wish to analyze in themselves. The lotus file lists a number of behaviors. The students may change the list to suit themselves. A graph can be made of the scores to help the students measure changes.

Students use a neuro-physiological monitoring system (also called a biofeedback machine) as part to the total awareness program of the course. With the use of the NPMS machine, students, working in collaborative groups, design their own experiments for a formal research paper.

Students use the NPMS system to see graphically, on the computer screen, what happens to them physiologically when they interact with persons, situations, machines. Experiments are a cooperative effort of the participants, are videotaped for further study and analyzed in relation to situation awareness, CRM, critical thinking, and communication skills.

The methodology and the value of this type of situational awareness training in undergraduates flight students will be discussed.

Introduction

This is a study of Situation Awareness in an academic course; HUMAN FACTORS IN PSYCHOLOGY. "Psychology" is taken in the sense of mental, physical, emotional and spiritual aspects of the health welfare and happiness of the person.

Situation Awareness and Crew Resource Management

Human Factors in Psychology, a psychology course, has put these concepts into practice, in a full semester course, for the past three years.

Situation awareness is explained as an understanding of the components in the environment at particular moment in time and space, the comprehension of their meaning and the projection of their status in the near future" (Endsley 1988). Crew Resource Management training is closely related to situation awareness. Elements of CRM include leadership, command, interpersonal

skills, communications, problem solving (critical thinking, in this paper), stress and the management of limited resources (Lauber 1987). These ideas were the motivators of the present discussion and excellent sources of information *The Role of Crew Resource Management (CRM) in achieving team situation awareness in aviation settings (Robertson and Endsley 1995).*

The study of human factors (SA AND CRM) begin with the person. Who is the person, not what is the body-machine-tool. When persons look at themselves (in a human factors engineering perspective) pseudo-scientific manner, being in control of their own investigation, with their own goals, expectations and dreams guiding their exploration of their mind-body existence, they will discover themselves and their capacities to be far superior to any conclusions made through reification of the human-thing.

The person can and should take charge of their own destiny. Abraham Maslow discovered in the characteristics of the self actualized person that the person has a natural drive toward health. The drive is fragile and can easily be missed or overridden especially if the person is trained or brainwashed by the system or society into being the efficient body-machine-tool ready for exploitation. I take the term "health" to mean mental, physical, emotional and spiritual health. Maslow studied well people (instead of the sick) and found that the most healthy people he encountered were "self actualized"; through introspection, they took possession of their own inner resources and became the person they wanted to be. This is a continual process and evolution in the growth of the individual. When this introspective growth takes place in the person, they will no longer be at the mercy of not only their own bodily whims but they will also no longer be at the mercy of the society's whims. They will become their own person.

In this course students learn:

1. critical thinking -- basing conclusions on evidence and valid criteria (problem solving)
2. neuro-linguistic programming -- self regulation, introspection
3. understanding of self and others
4. neuro-physiological monitoring -- feedback from the body
5. planning
6. socialization
7. investigative process
8. responsibility
9. moral-ethical behavior
10. communication skills
11. imagery rehearsal

Students write a Progress Report (self evaluation) on these criteria every 3 weeks.

Expectation of Research

The three year research period is expected to reveal, through anecdotal information, taken from student group-research papers, awareness profiles, two-page position papers and progress reports, that situation awareness and crew resource management can be taught in a class room setting using good educational practices without detrimental structuring and rigid "training" processes.

Critical Thinking, Self Awareness, Responsible Moral Behavior (SA and CRM)

The professor must have a philosophy of education that allows students the freedom to be themselves and, within the curriculum of a particular course, set their own goals, discover and meet their own needs besides meeting the requirements of the university and the professor.

Students must feel in control of their own behavior and their own learning processes. Toward this end students become aware (situation awareness) they are the agents of their own learning.

William G. Perry Jr. titled his book *Forms of Intellectual and Ethical Development in the College Years: A Scheme.* The teacher must model intellectual and ethical behavior. The teacher must be a co-investigator and co-learner, cannot dominate, manipulate or otherwise intimidate students. Education is a dialogue in which the students gather information and come to class ready to present what they have learned to other students for validation. Students then "teach" each other and learn from each other. Teachers become teacher-students, teaching and learning from their students. Students become student-teachers, teaching and learning from the teacher and peers. All become co-investigators. This model used in Human Factors in Psychology meets industry's expectations for training in Crew Resource Management and Situation Awareness.

Methodology

Students meet weekly for a three hour period. They practice Group Dynamics, working in groups of five, discussing assigned chapters on Human Factors. Bloom"s taxonomy of educational objectives is used as a model for helping the student discover an informal method of reasoning and evaluating their own thinking processes.

The student manual includes information and activities using a process of self evaluation while interacting with a computer, a biofeedback machine and working in small groups for support in learning and applying critical thinking skills and demonstrating ethical, moral behavior in a social-learning context. Students include examples, cases, questions, problems, glossaries information, bibliographies, references, appendices, on human factors from sources outside the book including the use of internet for current studies and issues.

Topics

Person-machine systems; human error; reliability; human factors methods; vision; audition; information presentation; visual displays; speech communication; workload geometry; anthropometrics; environmental stressors; human workload; automation; cognitive systems.

Communications

Johnson and Johnson"s *Joining Together: Group Theory and Group Skills* is a very comprehensive book of process. Structure is used as little as possible, emphasizing freedom to be oneself. Students learn honest, sincere and open communications is the basis of mental health. Students learn ethical behavior almost naturally, given the right amount of freedom, encouragement and modeling in these and the following group practices.

Students learn to foresee the consequences of their actions and take responsibility for their own behavior. They learn to speak from their own authority rather than parrot the ideas, biases and prejudices they formerly held due to ties of affection, loyalty, devotion or "training." They learn to back up their statements with reason, critical analysis and valid criteria. They learn the subject matter of the course in a social context, improve their communication skills, develop self confidence and like themselves more than ever before.

Critical thinking is an attitude or stance toward knowledge. It is a way of thinking about what it means to believe something. It is characterized by reflective skepticism. Fact can be questioned. The essence of critical thinking is values, goals, actions and the larger context of ideas for the purpose of enlarging student world views and increase their capacities to live examined lives.

Summary of theoretical approaches by Joanne Kurfiss:

Instruction for critical thinking sometimes incorporates development of critical perspectives, (shared mental models) which are frameworks for thinking about the implications and consequences of actions. Education for critical consciousness (Freire, 1973, 1986), empowers learners to influence their society. Critical perspectives can be cultivated through discipline-based instruction or they may be developed through reflection on life experiences. Democracy requires not only an educated populace, but a critically reflective one as well. (Kurfiss, 1989)

Ethical Behavior

According to Perry during the first year of college, students develop a dualistic position, believing there is an "answer" for everything, right or wrong, true or false, black or white.

During the course of the next four years, however, students discover there is no one answer to a question, that there is a multiplicity of answers. The students feel the urge to make a choice, discover and choose the "best" answer, realizes he/she *can* understand and *must act, ought* to get involved, make a judgment/choice of action then be responsible for it. This is ethical/moral behavior.

Results

Results of experimentation from progress reports and course evaluations: learned from experiments (use of neuro-physiological monitoring)

- be organized
- data does not confirm hypothesis
- critical analysis must "judge" data for conclusion
- scientific method, documenting is a necessity
- isolate the variable being tested
- standardize the process for accuracy
- reliance on other group members
- change my physiology by using my mind
- adjust the physical aspect of my body by adjusting my mind
- houghts can change our physical well being
- changed body outputs by concentration
- use feedback to positively control the body
- learned about myself
- my body reacted differently than I thought it would
- learned to schedule group process
- learned how to work together
- limited knowledge gained from experiments but they are beneficial and should continue
- through understanding self awareness we can achieve the highest understanding of ourselves
- information highway, research on the internet are all part of human factors
- keep up with advancing technology
- the course requires initiative and self determination on the part of the student
- the weekly papers and monthly progress reports provoked my free creative thought incorporating my analytical side
- progress reports forced me to confront questions that haunted my daily life
- new and unique teaching methods, each individual learns in a different way
- weekly evaluations demonstrated my strengths and weaknesses
- learned to watch what I do and say on a regular basis
- learned to care more about my friends and myself
- I quit drinking
- I have improved my social relationships, friends like me better
- I think more before I act
- internet helps to get work done on time
- non-traditional, student centered approach to learning
- I learned beyond the scope of the syllabus
- I learned how to control reaction and achieve higher levels of consciousness
- realized how to concentrate on a single stimuli
- progress report served as "self therapy"
- It helped me to order my thoughts and feelings, write them down
- became aware of some changes in my behavior, attitudes and
- personality that would otherwise have passed unnoticed
- I was surprised that my blood pressure changed according to mood
- depressing thoughts can ruin your physical health
- taste consistently produced the most dramatic physiological changes
- I was able to transfer learning from the npms system to my behavior
- I learned to be aware of the signals my body sends me
- I learned not to become a slave to the system
- relaxation and visualization can enhance my life and career
- I learned to not be manipulated
- group dynamics was vital to the learning process
- Communication skills teach honesty, sincerity and openness
- I improved my self image
- Imagery rehearsal improved my health and my learning across the board
- there is nothing quite as important as critical thinking

- the course makes you evaluate yourself and improve in areas where you are lacking
- I learned moral/ethical behavior through critical thinking

Publication of Research

Students in the Fall 1995 Human Factors courses will present eight formal research papers for peer review by an international panel of experts at Cranfield University in England for publication in the First International Conference on Engineering Psychology and Cognitive Ergonomics, Stratford-upon-Avon as a result of this semester's research on Teaching Situation Awareness in an undergraduate setting. Dr. Don Harris of that institution indicated by e-mail he would be interested in this research for March 1996.

Discussion and Conclusion

The CRM committee at ERAU west has discussed teaching CRM across the curriculum. Students report being "fed up" with CRM this and CRM that. They feel inundated with the terms, the process and the training. It has been the experience of the experimenter that another approach can be used. In the experimenter's mind, students must feel in control of their education. When they do, they are motivated to learn. Students become co-investigators, co-learners with the professor. The curriculum can incorporate CRM and Situation Awareness in the learning process without being oppressive.

Anecdotal information and self-reports are increasingly becoming more acceptable as "evidence" in supporting research conclusions. However, in the case of pilot training, more objective hard data is available and readily accessible in light of the use of more sophisticated training devices in aviation today.

There are many variables in teaching and learning. It is difficult to attribute learning to any one or more of these variables. Reasons for improved situation awareness in any or all of the areas studied might be due to factors outside the classroom. In the experimenter's academic classes the students teach themselves, the professor helps them validate their findings, tries to inspire them and supports their critical thinking and analyses. He then judges their performance and gives them a grade that hardly ever adequately represents what the student has learned. The researcher cannot substantiate with data (other than anecdotal) the sense that the students definitely demonstrated significantly better success in their learning during the human factors in psychology course than in a CRM training course. However, it seems that the data shows promise for further research in this area.

References

Adams, J.A. (1989). *Human Factors Engineering.* MacMillian.

Moray, B., (1988). *From Taylorism to Fordism: A Rational Madness.* London: Free Association Books.

Dilts, R., Grinder, J., Bandler, R., Bandler, L.C., and DeLozier, J., (1980). *Neuro-Linguistic Programming: Volume 1 The Study of the Structure of Subjective Experience.* Cupertino, California: Meta Publications.

Erickson, M.H., Rossi, E.L., and Rossi, S.I., (1976). *Hypnotic Realities.* New York, New York: Irvington Publishers.

Endsley, M.R.,(1989)."A Methodology for the Objective Measurement of Situation Awareness," *Situational Awareness in Aerospace Operations* (AGARD-CP-478. NATO - AGARD: Neuilly Sur Seine, France

Freire, P., (1973). *Education for Critical Consciousness.* New York: Continuum.

Freire, P., (1985). *The Politics of Education.* South Hadley, MA.: Bergin and Garvey.

Freire, P., (197?). *The Pedagogy of the Oppressed.* ____: _____.

Hancock, P.A., (ed.) (1987), *Human Factors Psychology.* Amsterdam: Elsevier Science Publishers.

Levidow, L. and Robins, K., (eds.) (1989). *Cyborg Worlds: The Military Information Society.* London: Free Association Books.

Lauber, J.K. (1987). "Cockpit Resource Management: Background and Overview," in H. W. Orlady and H. C. Foushee (eds.), *Cockpit Resource Management Training: Proceedings of NASA/MAC Workshop.* (NASA Conference Publication #2455). NASA-Ames Research Center, Moffett Field. CA.

Maslow, A.H., (1954). *Motivation and Personality.*_____. Harper.

Maslow, A.H., (1968). *Toward a Psychology of Being.* New York, New York: D. Van Nostrand Company.

Nelson, Craig. (1988). "Teaching Critical Thinking and Values Across the Curriculum: Classroom Applications of Perry's "Forms" and of "Women's Ways". A workshop given at the Twelfth National Institute on Issues in Teaching and Learning. Chicago, Illinois: University of Chicago.

Ornstein, R., (1985). *Psychology: The Study of Human Experience.* New York, New York: Harcourt Brace Javanovich.

Ornstein, R.E., (1972). *The Psychology of Consciousness.* New York, New York: Harcourt Brace, Inc..

Perry, William. (1970). *Forms of Intellectual and Ethical Development in College Years.* New York: Holt, Rinehart and Winston. (2nd ed.).

Robertson, M.M., and Endsley, M.R., (1995). "The Role of Crew Resource Management (CRM) in Achieving Team Situation Awareness in Aviation Settings." *Human Factors in Aviation Operations: Proceedings of the 21st Conference of the European Association for Aviation Psychology (EAAP).* (vol.3). London: Avebury Aviation.

Wiener, E.L. and Nagel, D.C., (eds.), (1988). *Human Factors in Aviation.* San Diego: Academic Press.

Wiener, E.L., Kanki, B.G. and Helmreich, R.L., (eds.), (1993). *Cockpit Resource Management.* San Diego: Academic Press.

Ergodynamics and its Application to the Work Productivity and Cost Effectiveness of Ergonomic Projects Implementation

Valery F. Venda and Ilona V. Venda

University of Manitoba

Abstract

Ergodynamics is being proposed as a part of ergonomics studying dynamics of work in dynamic work environment using transformation dynamics theory. Ergodynamics is based at three laws of: 1) mutual adaptation; 2) plurality of work functional structures; 3) transformations of work functional structures. All three laws are illustrated using simple experiment on typing productivity vs. desk height in a single sitting position (for the first law), for sitting and standing (for the second law), and for changing from sitting to standing (for the third law). Application of ergodynamics to the optimization of cost effectiveness of different ways in implementation of the ergonomic projects is discussed.

Introduction

In the paper dedicated to the fundamental theoretical problems of ergonomics W. Karwowski (1991) recalled that Wojciech Yastrzebowski gave it in 1857 this name combining two Greek words, work and natural laws, hoping that future generations will find the laws of this prospective science. Famous Russian psychologist Vladimir Bekhterev organized the first conference on ergonomics in 1921 (he named it "ergologia") stressed a necessity to study the laws of work and ergonomics.

In the keynote address for the XIIth Congress of IEA three laws were suggested as fundamentals of ergodynamics studying dynamics of work in dynamic industrial environment (Venda, 1994).

The laws of ergodynamics were worded as follows: 1) The First Law of Ergodynamics (The Law of Mutual Adaptation):

> **Work efficiency is a bell-shaped function of the factor of mutual adaptation between work functional structure and work environment. Work efficiency is maximal if work functional structure and work environment are mutually adapted.**

The First Law has a very simple visual image as a bell-shaped curve. There are two ordinates: efficiency (productivity, quality, work satisfaction, occupational safety and health), Q, and the factor of human-environment (human-machine, human-human) mutual adaptation, F. For example F can be a desk height influencing typing productivity. Q(F) is a work functional structure.

Ergodynamics studies work functional structures and their transformations. Other work structures are studied in other sciences. Work physiological structure is studied in work physiology, work body structure is studied in work kinesiology and anthropometry, etc.

F. Taylor in 1908 (see Taylor,1971) discovered work productivity is a bell-shaped function of the work environment (in his case, the shovel weight at a casting facility). He was the first researcher who proved that work productivity (Q) is maximal if an environmental parameter (work factor, F) has an optimal value. In his experiments, when a loaded shovel had a weight of approximately 10 kilograms (21.5 pounds), the average worker reached maximal shift productivity. If the weight was lower or higher than the optimal, the productivity was lower. changed the size and, consequently, the weight of a shovel with material moved by the worker. Thus, he found the optimal shovel weight for American males approximately 21 lbs or 10 kg. At this weight, the highest amount of material shoveled, per shift, was reached (see Figure 1). Of course, a smaller man needs a lighter shovel weight for his maximal productivity, and a larger man could reach higher productivity with a bigger shovel and heavier loads.

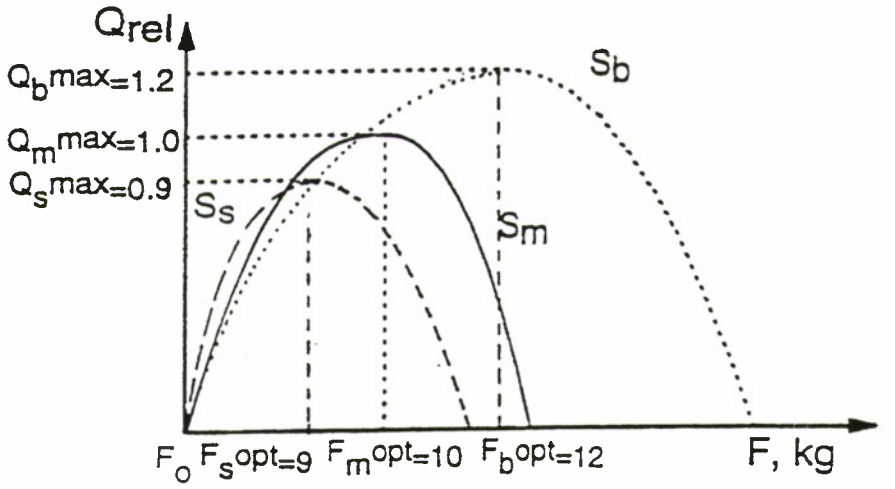


Figure 1. Characteristic curves for shoveling work efficiency for average American men (solid line), small and big. Q - relative work productivity, F - shovel weight with material (F, kg); F_o= the weight of the empty shovel.

Frank and Lillian Gilbreth (see Freivalds, 1987; Konz, 1990) studied workers' micromotions and simple operations (called "therbligs," their last name spelled backwards). They also found the bell-shaped dependence of productivity (Q, a number of bricks laid in an hour) on the work (ergonomic) factor (F, a number of motions on one brick laid), like that shown in Figure 1.

The shovel weight or number of motions for one brick laid is an important work factor (F). It is not an independent parameter of either the working human or of the work environment. It is a common human-environment parameter; it displays human-environment interaction and mutual adaptation during training and work processes. The number of bricks laid per hour *(Q)* is a criterion that one wants to be as large as possible. We call this type of work criteria *functional efficiency,* or simply *efficiency (Q).* Dependence (Q(F)) reflects a certain *work functional structure.*

2) The Second Law of Ergodynamics (*The Law of Work Structures Plurality*):

Every work task can be done with different work structures.

A visual image of the Second Law looks like a family of the bell-shaped curves (Venda, 1994). Every work functional structure has its specific optimal F value and maximal level of efficiency.

3) The Third Law of Ergodynamics (*The Law of Transformations*) was worded by Yuri Venda (Y. Venda and V. Venda, 1991, 1995):

Transformations between different structures of the system and interaction between different systems' structures are maximally effective if they go through a state common and equal for the structures.

Let us illustrate the laws using very simple experiments.

Experiment #1: Typing Productivity vs Desk Height for the same Work Posture.

The experiments were conducted by our students at the University of Manitoba to demonstrate a work functional structure ($Q(F)$), and its mutual adaptation with the environment following changes in (F). We created these experiments under conditions of maximum simplicity to facilitate the condition that everyone participating not only understand the experiments and results, but be able to repeat them at home or in the office (to prove a bell-shaped curve of the functional structure) to allow them to improve work comfort and productivity, thus lowering the risk of repetitive strain injures.

There is a wide literature available on chair and desk adaptation (Bendix, 1986, Bendix and Bloch, 1986), sitting habits (Burandt and Grandjean, 1963), muscle strain and fatigue in sitting and standing positions (Greenberg and Chaffin, 1977), development of work skills to help one to adapt to one's work environment (Salvendy and Seymour, 1973, Salvendy and Pilitsis, 1974).

We will present data only for two our subjects. Indeed everyone should conduct this experiment for her/himself. As a result of the "homemade" experiment, everybody can find the precise individual optimal desk height for him or herself, as well as for family members, friends, colleagues, or for any person who would want to participate in those simple experiments, and effectively work both in a sitting and a standing position. Changing of one's position in lengthy work, with computer for example, aids in the prevention of overexertion of some muscles and organs (S. Kumar, 1992).

Method

We have spent many years looking for this type of experiment on the functional structure curves: theoretically clear, practically useful and easy to understand for everyone. Such an experiment could be reproducible in any circumstances, as well as convenient for the high school and university laboratory.

Subjects were typing text at a Darius laptop computer (similar to a Sharp 6781) with a monochrome screen. Position of the computer and text to be typed is shown at Figure 2. The only variable the experimenters altered was the height of the work surface for the notebook computer. The subjects typed in a sitting position utilizing seventeen different heights in a random sequence. They completed three trials at each level.

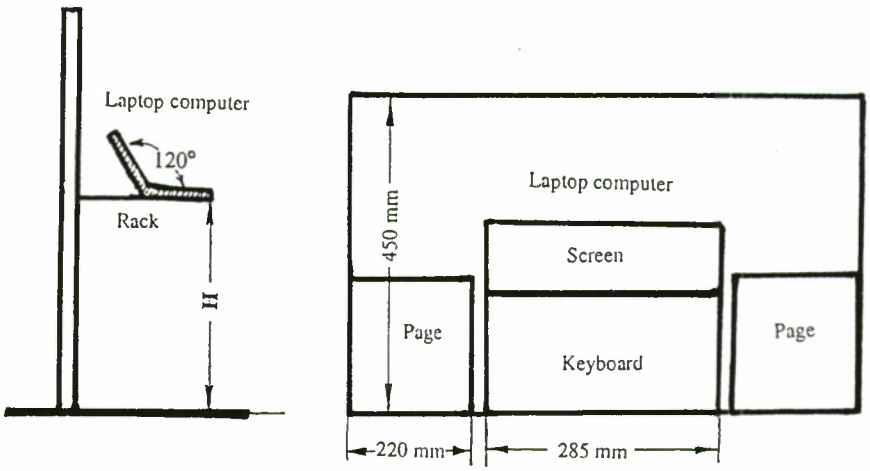


Figure 2. Position of the laptop computer and text in typing experiments.

Results and discussion.

It was found that the lowest and highest margin levels, (where typing is impossible and the number of characters typed was Q=0) as follows: the lowest level was when further lowering required bending to reach the keyboard. The highest level was equal to the eye's sight level.

Aside from the desk height, all other aspects of the work environment were held constant (noise, lighting, position of article to be typed, size and height of chair, computer screen angle).

These results were obtained with two male subjects, similar in their physical characteristics and typing skills, to participate in the experiments:

Subject 1 was 22 years old, 5'11", 155 lbs;
Subject 2 was 22 years old, 6'2", 192 lbs;
The highest height used was 125 cm;
The lowest height used was 45 cm;
The chair was 39 cm from the floor to seat;
No bending at the waist was allowed;
We used 30 different text articles of equal reading level;
The font size was 14 for the articles and computer screen;
The keyboard size was 28.5cm x 11cm.

We plotted the results for the number of characters typed (Q) versus the desk height (F). Figure 3 shows the results of the experiment for typing as the number of the characters typed per minute while the participants were in a sitting position (for 3 and 30 minute trials).

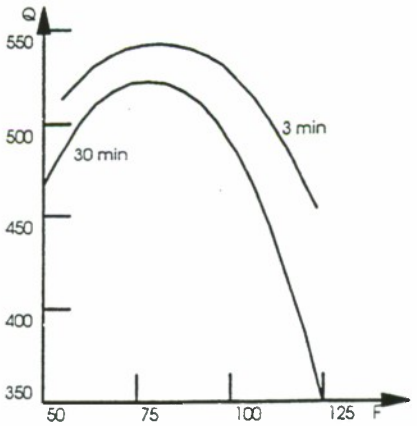


Figure 3. The number of characters typed in a three minutes (Q) as a function of the computer desk height (F, cm) in a sitting position for 3min and 30 min experiments.

The functional characteristic curve for typing in a sitting position has a bell shape, with an optimum at 85.5 cm. The number of typing errors during the three minute period did not have any certain relationship to desk height.

The curves for the 30 minute typing condition were steeper than for the 3 minute typing condition. When height in a sitting position changed from 85.5 cm to 52 cm, the average number of characters typed decreased in the three minute interval from 530 characters per minute to 510 char/min. In the thirty minute test, theses results dropped from 520 to 480 char/min. Thus for longer work, adjusting the work surface height is more important than for short term work.

For the standing position condition, lowering the height influenced output more than increasing the height.

Experiment #2: Studying of Two Functional Structures of the Same Work.

These experiments were conducted in the Human Factors laboratory of the University of Manitoba, using the same methods and work stations as in the previous experiments #1 and #2. Besides typing text on the computer, and peg-board assembly while sitting during one trial, our subjects did the same jobs while standing during another trial. Figure 4 shows the functional structures of subjects' work in both positions for typing.

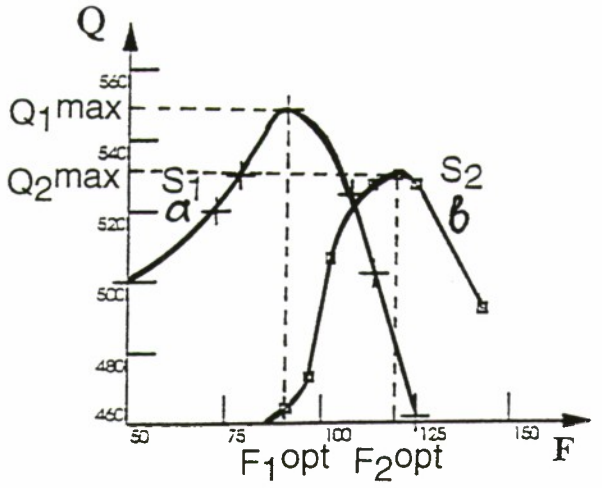


Figure 4. Bell-shaped functional structures of experimental typing efficiency (Q) while sitting (a) and standing (b), vs desk height (F, cm).

The experiments on typing while sitting and standing have shown that one can do the same work task with different work postures, or functional structures. Each structure is modeled with a specific bell-shaped characteristic curve displaying the dependence of work efficiency (productivity of typing) on the environmental factor (desk height).
Comparative analysis of the characteristic curves of a concrete worker and the characteristics of different workstations allows one to reach mutual adaptation between the individual and the appropriate work station characteristic and thus maximal individual work efficiency.

Typing Experiment to Illustrate Smooth and Effective Transformations of the Work Functional Structures.

Data on typing productivity in sitting and standing positions help to model and optimize a process of transformations between the functional structures for better understanding ergodynamics Law 3.

The task is to find optimal desk height so changing the position from sitting to standing and back will give maximal total productivity of typing at all heights. A special condition is that the subject should type at the "transition" height twice: sitting and standing.
A situation could be defined as a following: we change desk heights from F_{min}=50 cm (about 20") to Fmax=150 cm with an interval 5 cm, 21 trials. A subject types 3 min at every height. We may require the subject to type in sitting or standing position at every height. Our task is to maximize a total productivity of the subject at all heights as a sum of the numbers of characters typed at every height. The main question is what height is optimal to change a sitting position on the standing one.

Figure 5 helps to solve this task. Let us assume that the subject started to type at F=50 cm and showed a productivity 500 characters in three minutes. At every next height his or her productivity was growing till an optimal height for typing in sitting position F=85 cm was reached. Then the

productivity in sitting position started to decrease and we were looking a height to change the position to the standing one to reach a maximal total productivity of typing.

Let us compare three possible heights: F=100 cm, F=110 cm and F=115 cm.

At F=100 cm productivity of typing in sitting position is essentially higher than in standing position. Changing the position (that is in our case a functional structure S_1) from sitting S_1 to standing S_2 will lead to the essential dropping of productivity (see Figure 5, left and right sides).

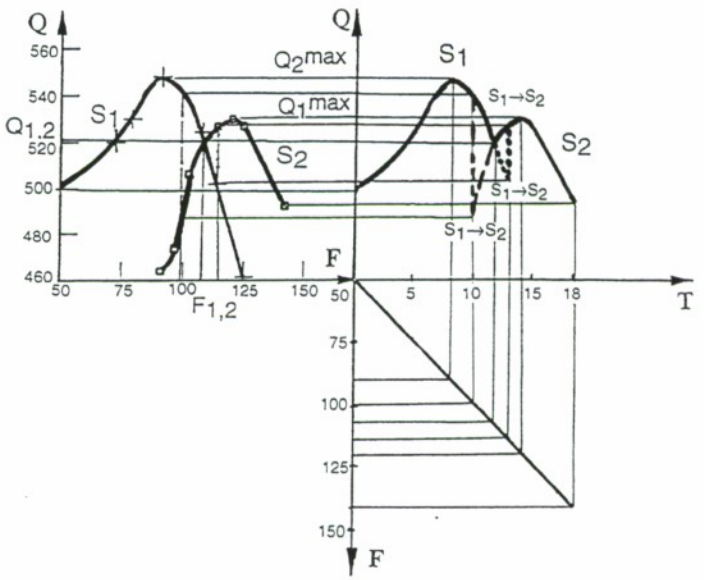


Figure 5. Searching for the optimal height to transform sitting position to standing position if desk height is being increased from 50 cm to 140 cm with increment of 5 cm.

If the subject will change the sitting position to standing one at F= 115 cm, typing in sitting position will be much slower than typing in standing position and that will lead to the essential loss in the total typing productivity.

The maximal total productivity will be reached at F=110 where the productivity in sitting and standing positions are approximately equal each to other, Q_1(F=110)Q_2(F=110)520 char/3 min.

Let us compare the total productivity at three heights: 100, 110 and 115 cm if the position is changed at 1) 100 cm; 2) 110 cm and 3) 115 cm.

1) Q_1(F=100 cm) = 540 char/3 min; Q_2(F=100 cm) = 487 char/3 min; Q_2(F=110 cm) = 520 char/3 min.; Q_2(F=115 cm) = 527 char/3 min. Total productivity at three heights at the first case will be Q_{Tot1}= 540 + 487 + 520 + 527 = 2,074 char/12 min.

2) Q_1(F=100 cm) = 540 char./3 min.; Q_1(F=110 cm) = 520 char./3 min.; Q_2 (F=110 cm) = 520 char./3 min.; Q_2 (F=115 cm) = 527 char./3 min. Q_{Tot2} = 540 + 520 + 520 + 527 = 2,107 char./12 min.

3) Q_1(F=100 cm) = 540 char./3 min.; Q_1(110 cm) = 520 char./3 min.; Q_1(F=115 cm)= 500 char./3 min.; Q_2 (F=115 cm) = 527 char./3 min.Q_{Tot3} = 540 + 520 + 500 + 527 = 2,087 char./12 min.

The total productivity at the second case Q_{Tot2} is bigger than at the first case on 33 char./12 min. and on 20 char./12 min than at the third case. For typing during many hours this difference is very essential. Besides, optimization of desk height and other work environment parameters lower risk of repetitive strain injuries (RSI).

A conclusion is that the total productivity when different work functional structures are used is the biggest if the functional structures are changed in the work condition (in this case at the height) when both structures are equal each to other. We name this condition as a common and equal state for two functional structures that transform each to other.

The third law of ergodynamics may be used to find optimal ways and tactics to transform production technology and processes.

Ergodynamics in Analysis and Increase of Work Productivity

Using laws and principles of ergodynamics we assessed dozens of ergonomic projects. We found ergonomists often make three crucial, yet common, mistakes in practice:

1) They consider the objective environment (machine, display, manual, advice, management decision) characteristics (E) as those actually involved in the workers' performance as a subjective image (F). Thus, they usually assume that F=E. As a result, an ergonomic design often fails because it is irrelevant to the real work.

2) They do not study the qualitative and quantitative characteristics of the work functional structure the workers use in reality, from one side, and the subjects use in the laboratory or in special tests, from other side. As a result, ergonomic recommendations based on laboratory data could easily decrease work efficiency further.

Ergodynamics will help ergonomists and applied psychologists to overcome these mistakes and greatly improve the practical impact of their recommendations and designs.

Weighing the starting characteristic curves of system structures and their modification potential by changing component set-up and activities, one could solve the task of defining system development transformation dynamics. Conversely, one could choose these S_i to obtain a desired (pre-assigned) development curve, typified by the shortest time, least transformation losses, and so on.

An intensive search for a new structure and its development becomes imperative when the former structure-strategy (production process, scientific theory or social system) is about to reach a "ceiling" (plateau). The greater the difference between the old and new functional structures of manufacturing the greater will be the drop in productivity.

Proceeding to the system structural transformation, one should use ergodynamics to be able to predict and organize a complicated dynamic process well in advance to avoid a negative consequences such as following:

1. If one made a structural transformation of the manufacturing company and meeting big financial losses decides to go back to the old structure and thus effects the reversal too fast, before the new structure has taken body and form, the damage from such an abortive attempt will be irreparable. It will prohibit, for a long time, any other try.

2. If one steers the structural transformation process at crash rates without a new structure, but under new conditions (F), the old structure will not cope with the strain; the system may collapse.

In planning a structural transformation, one must chose a new structure, looking for an optimal efficiency increment (DQ), an acceptable factor increment of the factor (DF), an acceptable efficiency drop during transformation (DQ_T) and an acceptable transformation time, DT.

In choosing an acceptable efficiency dip and its transformation duration, one should weigh the negative side effects: a drop in the company gains at the first structural transformation stage is very hard to bear. Here, one should brave the transformation period problems until the new profit occurs. When a new structure has formed, a further change of F (the change that at first ushered in a recession) an up-trend in efficiency will accompany it.

Therefore one must identify a set of dominant interrelated transforming factors and control them in a target-oriented manner. Consummate wisdom here is to be able to endure the drawbacks of the transformation period-- the declining productivity and the inevitable psychological repercussions.

A hasty structural transformations may entail even worse results than relying on the old manufacturing structure. Indeed, if one launches a structural transformation and destroys the old structure but then, as the transformation process is nearly over, if one makes an about-face and tries to revert to the old pattern, the outcome will be really disastrous. The old structure is no more, while the new structure, still at the inchoate stage, is not productive yet.

Cost Effectiveness of Different Transformation Ways in Implementing of Ergonomic Project

There is a very special question could be analyzed using ergodynamics theory: If there are several functional structures available, for example S_a, S_b, and S_c, what transformation tactics is the most economical and cost effective? The objective is to use the ergodynamics to minimize the economic losses of the manufacturing facility during transformations of the work structures while new ergonomic project is being implemented. For example we suggested to change electronic assembly workstations so instead of traditional direct watching assembly operations accompanied with bending neck and body the workers can be seated in a straight position watching the operations on TV screen. Therefore we have a choice to transform traditional, direct vision, assembly processes (S_a) first to the mixed, direct and indirect vision (S_b) and then to the indirect ones (S_c) or to transfer direct vision assembly (S_a) right into indirect one (S_c). What variant of transfer of new workstations will lead to less losses and larger profit? This is the question for ergodynamics.

Figure 6 shows variants of transformations among three structures S_a, $S_b$and $S_c$and corresponding changes in system efficiency. Efficiency (productivity, quality, profit), Q, is a function of different work factors (complexity of operations, software, management structure, level of workers motivation), F. Efficiency may be studied as a static function of F: Q(F) (see left side of the Figure 6) , and dynamic function of F and time, T: Q(F,T) (right side of the Figure 6).

Integral efficiency (for example, company profit) gained between two values of F may be expressed as a square under certain structure characteristic curve. If different structures were used subsequently the square (integral) will be calculated under all respective curves (Venda and Strong, 1994).

Theoretical principles and practical methods of the ergodynamics are described in the book by V. Venda and Y. Venda (1995). Let us explain the main idea of ergodynamics in application to the implementation of ergonomic project using sketches made by Ilona Venda. Let us assume the company with lower productivity than a competitor is a loser in the competition (Figure 7).

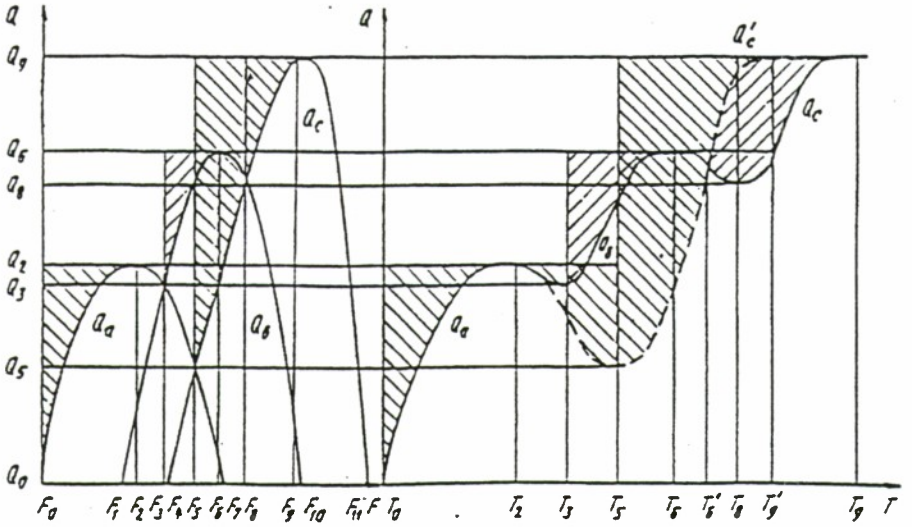


Figure 6. Comparison of the efficiency (or/and profit) losses and gains for different transformation tactics and ways.

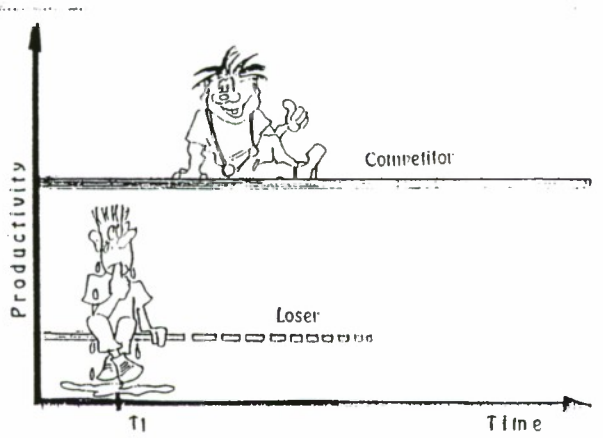


Figure 7. Losses and tears push towards needs in ergonomic innovations and projects

Usually the company which is loosing the competition with others needs help from the ergonomic consultants (Figure 8).

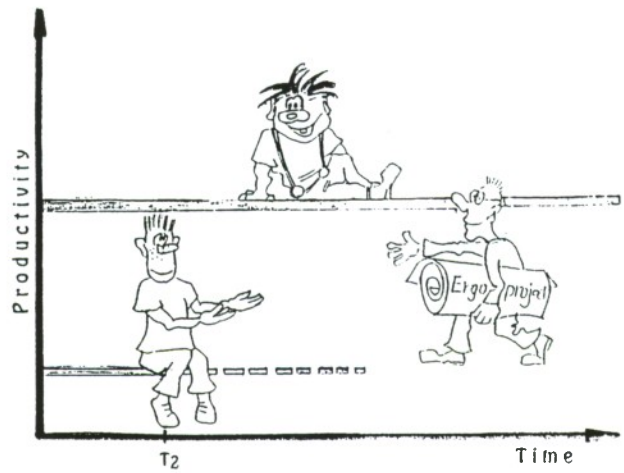


Figure 8. Loser prays for innovative ergonomic miracle

Every consultant, innovator or designer ergonomist always promises quick investment return and success (Figure 9).

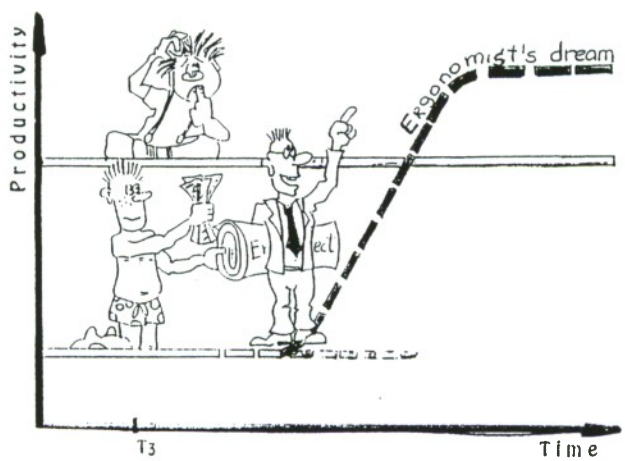


Figure 9. Ergonomic designers and innovators always promise immediate take-off

In reality implementation of any new ergonomics project leads first to even larger losses, decrease in productivity and profit (Figure 10).

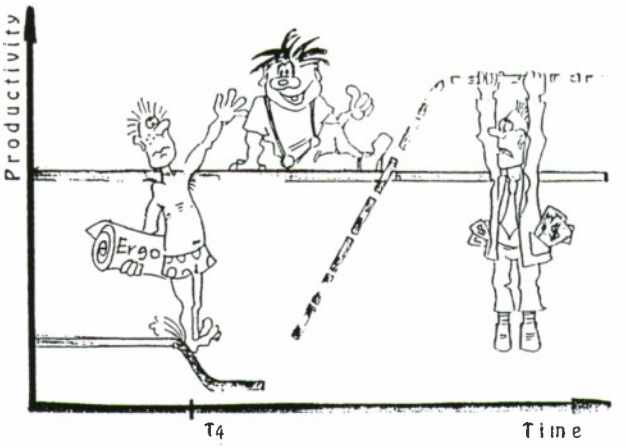


Figure 10. In reality transformation dynamics leads first down.

If the client was not taught ergodynamics and was not prepared to the temopary losses he or she gets very angry and rejects the project (Figure 11).

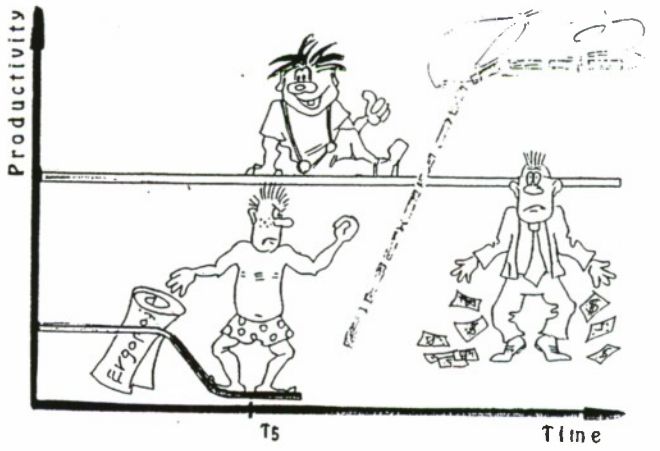


Figure 11. Innovator is perceived as a liar.

Ergodynamics helps to predict, plan and well organize the process of implementation of the new projects and technologies (Figure 12).

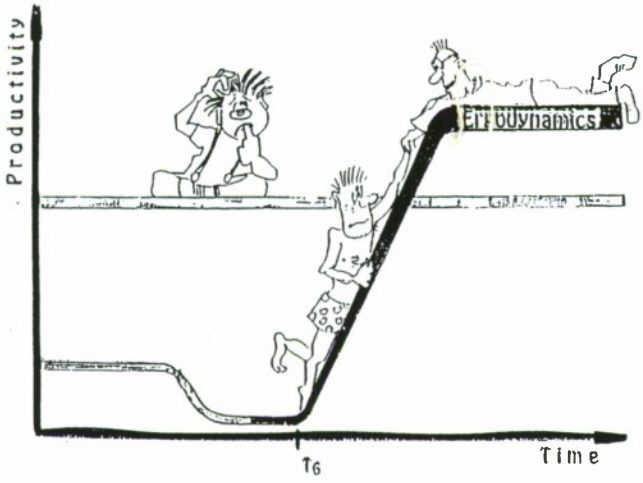


Figure 12. Ergodynamics helps to plan real transformation process while implementing ergonomic project.

As a result of the collaboration with specialist in ergodynamics the client get a great success and previous winner in competition becomes a new loser and therefore a client for the ergodynamics consultant (Figure 13).

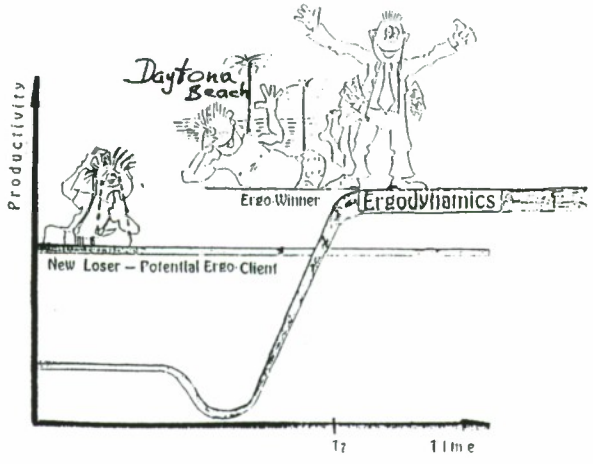


Figure 13. Transformation dynamics: successful implementation of ergonomic projects.

Conclusion

Ergodynamics studies dynamics of individual, team or company work functional structures. It is based at three laws: 1) mutual adaptation; 2) plurality of functional structures; and 3) transformations.

Ergodynamics helps to predict and plan work productivity dynamics in training and re-training processes or when implementing new ergonomic project. The total productivity during implementation of the ergonomic project, when different work functional structures are transformed each to other is the biggest if the functional structures are changed in the work condition when both structures are equal each to other. We name this condition as a common and equal state for two functional structures.

Acknowledgments

These studies were supported by Northern Telecom Canada Ltd., Natural Science and Engineering Research Council of Canada, Bell-Northern Research and the University of Manitoba.

References

Bendix, T. (1986) Chair and table adjustments for seated work. In: *The Ergonomics of Working Postures*. London: Taylor and Francis.

Bendix, T. and Bloch, I. (1986) How should a seated workplace with a tiltable chair be adjusted? *Applied Ergonomics*, Vol. 17 [2], 127-135.

Burandt, U. and Grandjean, E. (1963) Sitting habits of office employees. *Ergonomics*, Vol. 6 [2], 217-228.

Freivalds, A. (1987) The ergonomics of tools. In: International Review in *Ergonomics, v.1*, 12-48.

Greenburg, L. and Chaffin, D. (1977) *Workers and Their Tools*. Midland, Michigan: Pendall Publishing.

Karwowski, W. (1991) Complexity, fuzziness and ergonomic inompatibility issues in the control of dynamic work environments. Ergonomics, 34 (6), 671-686.

Konz, S. (1990). *Work Design: Industrial Ergonomics*, Worthington: Publishing Horizons.

Kumar, S. (1992). Ergonomics in rehabilitation: a conceptual model, In: *Advances in Industrial Ergonomics and Safety*, London: Taylor and Francis, 1157-1164.

Salvendy, G. and Pilitsis, J. (1974) Improvement in physiological performance as a function of practice. *International Journal of Production Research*, Vol.12 [4] 519-531.

Salvendy, G. and Seymour,W. (1973) *Prediction and Development of Industrial Work Performance*. N.Y.: Wiley & Sons.

Taylor, F. W. (1971). *The Principles of Scientific Management*, N. Y.: Harper and Row.

Venda, V. F., and Venda, Yuri V. (1991) Transformation dynamics in complex systems, *Journal of Washington Academy of Science*, #4, December.

Venda, V. F. (1994) Dynamics in Ergonomics: Theory and Tips (Manifesto of Ergodynamics) - Keynote Address for the World Ergonomics Congress IEA'94, Toronto, August, 1994. *Proceedings of the 12th Triennial Congress of International Ergonomics Association IEA'94*, HFAC, Toronto, 1994. 34-36.

Venda, V. F., and Strong, D. R. (1994) Ergodynamics and cost effectiveness of the structural transformations in manufacturing. *Human Factors in Organizational Design and Management-IV,* 175-180.
Venda, V. F. and Venda, Yuri V. (1995) *Dynamics in Ergonomics, Psychology and Decisions: Introduction to Ergodynamics*, Norwood: Ablex Publishing Corporation (in press).

New Workstations for High Productivity and Low Risk of Occupational Injures: Design and Industrial Testing

Valery F. Venda and Ilona V. Venda

University of Manitoba

Abstract

There are a lot of repetitive strain injures of neck, back, shoulders, and wrists at electronic assembly industry and at manual material handling operations in many other industries. Awkward work position with bent neck and body, and lifted arms are the main causes of the injures. Northern Telecom, Bell-Northern Research and the University of Manitoba collaborate on the project Assembly workstations for high productivity and low risk of occupational injures. Using methodology of the ergodynamics, the authors invented and designed new assembly and manual material handling workstations presenting manual operations at TV screen which allow workers to be seated in a straight back position and use a negative tilt of work surface, in addition to the usual positive and zero tilt. A large magnification of printed circuit board or any product handled manually is another important advantage of the new workstations. Assembly operations at traditional workstations and new workstations (the V- workstations) were assessed and compared at Nortel Wireless Systems Calgary Plant using criteria of productivity, tempo, left and right upper trapesius muscle strain measured with EMG and motions of neck and back measured with goniometers made by Premed As and with direct measurement at videotapes. Tests proved advantages of the V- workstations in comparison with traditional ones.

Introduction

Almost every day, new data concerning the epidemic of repetitive strain (stress) injuries (RSI), or cumulative trauma disorders (CTD), is published. On December 28, 1994 the Investors Business Daily published a report of the US Labor Department stating a 10% increase in RSI occuring during 1994 with 302,000 workers claiming injuries versus 34,700 workers claimed RSI's in 1984. There are several detailed surveys on RSI and ergonomic studies of their prevention (Marek, Wos, Karwowski, Hamiga, 1992; Fisher, Andres, Airth, Smith, 1993). Our major goal is to increase work productivity and decrease losses caused by the repetitive strain injuries (RSI) or cumulative trauma disorders (CTD) at electronic assembly and material handling operations. We studied the experience of RSI at the assembly plants of Northern Telecom and other electronics, communication, computer, watch and other companies, and in material handling (diamond processing, etc.). It was found that neck and shoulder RSI are the most widely spread traumas among the workers of slidelines at which printed circuit boards (PCB) are assembled and in manual material handling. The RSI's are caused by bending of the worker's neck, along with a repetitive lifting of the arms during a work shift at assembly slideline and material handling (Fisher, D.L., Andres, R. O., Airth, D., and Smith, S. S. (1993); Marek, T., Wos, H.,

Karwowski, W., and Hamiga, K. (1992); Schuldt, K., Ekholm, J., Harms-Ringdahl, K., Nemeth, G., and Arborelius U. P. (1986).
It was found that workers at the assembly lines bent their neck and body excessively (Figure 1). The surveys have shown that neck flexion, along with lifting of the arms, leads to the development of RSI which may, in turn , lead to decreased work productivity and quality.
The NT Calgary Wireless Systems Plant provided ErgoLab with the traditional slide line workstations which are currently being used to simulate typical assembly operations within the ErgoLab.

Ergodynamics approach to design of highly productive and low RSI risk workstations

We found that all current designs of the assembly workstations which lead to extensive RSI's were based on the erroneous assumption that it is possible to find an optimal solution for both, the worker's arms and eyes. Using our new methodology of ergodynamics (Venda and Venda, 1991, Venda, 1994, Venda and Venda, 1995), we studied work functional structures of the arms and eyes, acceptable angles of the assembly work surface (PCB) for the arms operations and acceptable angles of the visual object (display, screen) for the neck and eyes. Position of the head (angle C), and thus the neck flexion depend on the angle of the visual object. The smaller the angle of the object, the larger flexion of the neck. At the same time, position of the arms depends on the position of the PCB: the higher PCB is better observed by the eyes, but increased lifting of the arms increases the shoulder muscle strain.

Therefore our analysis showed the traditional ergonomic approach to the design of the assembly and material handling workstations is misleading, as no solutions could be identified which was optimal both for the eyes-neck and for hands-shoulders. Thus, we suggested the work surfaces for the arms and eyes should be separated.

We invented new type of assembly and material handling workstations, where the work surfaces for the arms and eyes are separated and optimized.

Our new assembly and material handling workstations allow the workers to be seated in a straight neutral work position thus lowering fatigue and risk of RSI's of the neck, shoulders and back. The workstation also accomodates a negative tilt of the work surface thus lowering the risk of carpal tunnel syndrome. Lowering the fatigue may lead to increased work productivity and quality.

No previous attempts to solve this problem for assembly operations was successful. We realized the problem reqires a new theoretical approach.

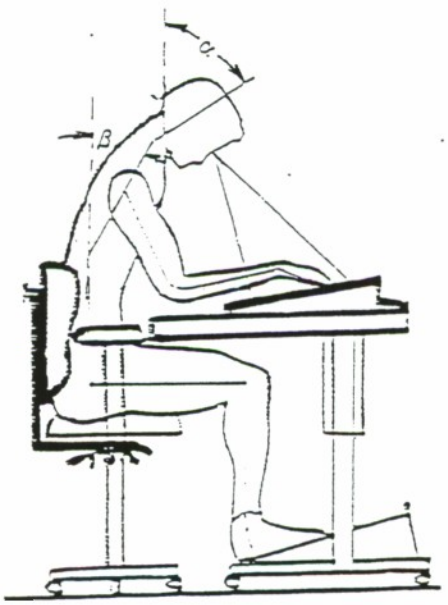
This study is the first attempt to use the laws of ergodynamics for practical purposes. The Third Law of Ergodynamics (The Law of Transformations) by Yuri Venda (Y. Venda and V. Venda, 1992, Venda, 1994) states:

> Transformations between different work functional structures, and interaction between the different structures are maximally effective if they go through a state common and equal for the structures."

Figure 2 displays the work functional structures for the worker's eyes (visual performance), Qe(F), and for the arms, Qa(F). F represents the angle of the respective work surface. Intersect point of the curves Qa(F) and Qe(F) models the common and equal state for the eyes' and arms' work functional structures. That means using common work surface for the arms and eyes leads to a relative maximum work efficiency only at Qrel=0.43 when $F=22^{o}$. All other angles decrease the efficiency of visual performance or the arms, and therefore decrease the work productivity at



a)



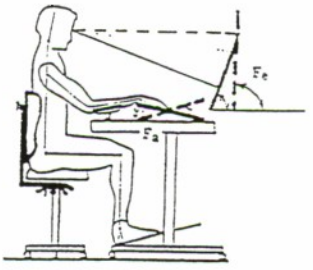
b)

Figure 1. Awkward worker position at the traditional workstation.
Bending neck, body, lifting arms, flexion of the wrists inevitably lead to RSIs.

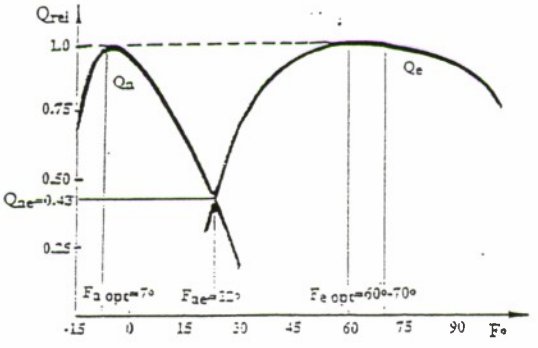
assembly operations. Thus the traditional approach of ergonomic design to assembly workstations based on coordinating the eyes and arms in assembly operations should be changed: a comfortable, neutral position of the arms leads to neck flexion, whereas a straight, neutral position of the neck leads to an increased lifting of the arms.

The ergodynamics design principle used in this Project is such that a solution to the problem must be found by attempting to maintain an independence between the arms and eyes. On the basis of ergodynamics principles new types of workstations were designed (Figure 3).

The worker watches the assembly operations on a TV screen, which displays a magnification of the assembled parts image. The neck and body are in a straight, neutral position. The arms are also in a comfortable, neutral position accompanied by arm supports, which lowers the muscle strain. The positions, motions, and associated muscle strain of the head, back, and arms are registered, measured, and analyzed using the advanced, portable system Physiometer-400 (Aaras, 1994). It also allows for four channels of electromyograpgy signals, with recording by a miniature computer by Hewlett-Packard, carried by the worker on a belt, or installed at the workstation. Assembly operations are simulated at the Lab using standard NT equipment: workstations, Kan-Ban boxes, lights, foot rests, etc. to match the industrial environment as close as possible in order to perform the same experiments at different NT plants. Magnification of the parts assembled can be varied to a large degree. These workstations allow the workers to use a direct vision (traditional way) and indirect, or TV vision, changing work posture and further decreasing both static and dynamic muscle strain.



a)



b)

Figure 2. Optimal angles of work surface for arms (Fa) and eyes visual performance (Fe) are very different. We suggested eyes and arms should work separately.

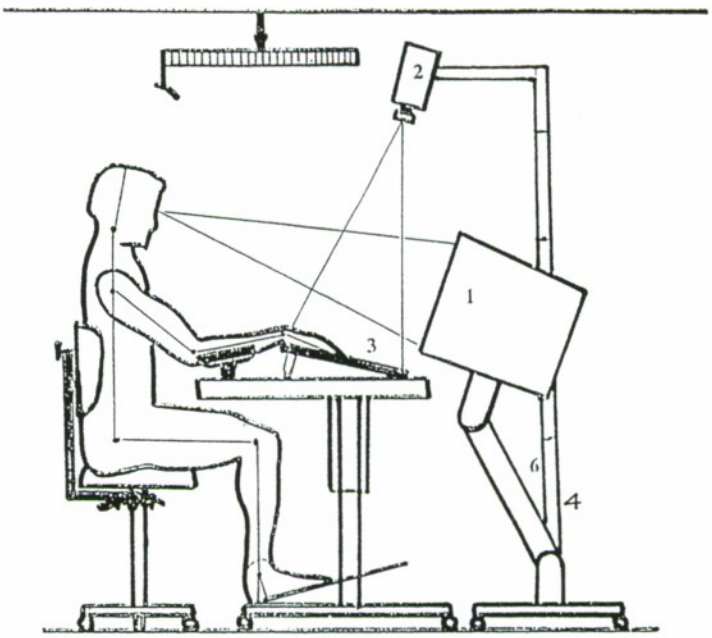


Figure 3. Assembly and material handling workstation (The V-workstations), presenting manual operations at TV screen, allowing negative tilt of work surface and straight work position

1- screen displaying manual assembly or material handling operations; 2 - TV camera, optical projector or mirror; 3 - adjustable tilt, the workstation allows a positive, zero and negative tilt of PCB, watch, diamond or any other object or material assembled or handled; 4 - autonomous adjustable module supporting screen and TV camera as possible solution.

Ergodynamics and justification for industrial ergonomic experiments

R&D activities on ergonomic design and experimental studies of prospective RSI-free workstations, reorganizing work sensory-motor activities, professional retraining are based at the laws and principles of ergodynamics (Venda, 1994). Ergodynamics helps to understand why many traditional research experimental methods have a limited application in ergonomic practice. Transformation dynamics theory (ergodynamics) describes every work functional structure as a bell-shaped curve or more often a family of bell-shaped curves $Q_i(F_j)$, where Q_i is a criterion of work efficiency, F_j is a factor of work efficiency (V. Venda and Y. Venda, 1991, 1992, V. Venda, 1994). Let us assume that a real operator (worker) uses in practice the work functional structure S_{op} (Figure 4). To assess ergonomic design mockup (workstation, control board, display) in the lab a group of the laboratory subjects using a work functional structure S_{sub} is invited. One can organize a full range experimental studies by changing the work efficiency factor

F from F_{sub}^{min} to F_{op}^{max}. Performance of the subjects with the factor F in the interval F_{sub}^{min} - F_{op}^{min} is irrelevant because operators (workers) never work in this interval. In the interval F_{op}^{min} - F_{sub}^{opt} increasing F leads to an increase in Q_{sub} and Q_{op}. This means Q_{sub} and Q_{op} are positively correlated each to other, and the better results demonstrated in the lab by the subjects using new workstation indicate better efficiency of the operators in practice. But the difference between Q_{sub} and Q_{op} in this interval of very simple tasks is very large: professionals do not work with this low load and low complexity of the tasks.

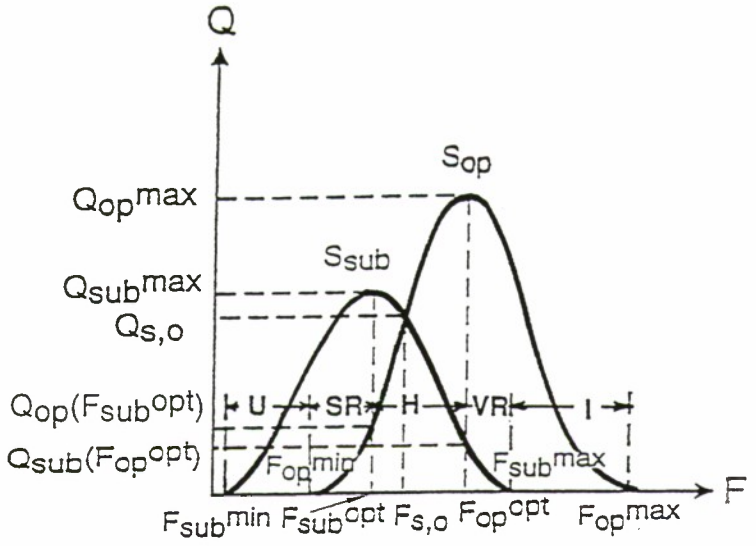


Figure 4. Analysis of the low practical efficiency of the experiments with laboratory subject, S_{sub}, for improvement of the work functional structure of the real operator, S_{op}.

This means the experimental results obtained in this interval are useless. If we conduct laboratory experiments in the interval F_{sub}^{opt} - F_{op}^{opt}, then practical implementation of the research results will be shocking: the better performance efficiency experienced in laboratory testing, the worse work efficiency results in practice: Q_{sub} and Q_{op} have a negative correlation in this interval. Thus common sense here is much better than research recommendation. In the interval F_{op}^{opt} - F_{sub}^{max} the subjects show such a low efficiency that their results are useless in practice. It is obvious that the subjects cannot work with a task complexity higher than F_{sub}^{max}. Thus the ergodynamic analysis shows that identity between work functional structures of the test subjects and operators (workers) is required to get practically valid experimental ergonomic results which could be implemented to the practice.

Laboratory subjects usually have work functional structures essentially different from those used by the real workers. Therefore, in the design and improvement of assembly workstations we used laboratory experiments with students as subjects only for a preliminary evaluation and to improve the test methods and techniques of the new workstations. Our main interest was in the industrial testing of the workstations.

Industrial comparative testing of traditional and V-workstations

Twelve workers of the Northern Telecom Wireless Systems Calgary Assembly Plant participated in the experiments. Each of them worked, on first day, at the Traditional workstation (made by GWS) or Cut-out workstation (advanced traditional). The traditional workstation used is one of the most spread slide-line workstations with narrow front. It was coded as T-workstation or simply T". Cut-out workstation is a traditional workstation which uses a wide adjustable table with cut-out which allows the worker to be proximally closer to the printed circuit board (PCB) and the boxes which contain the components to be inserted (populated) into the PCB. After this stage each worker then worked seven days at the Venda workstation based on either Traditional one as combination Venda-Traditional workstation (coded as VT) or Cut-out one as combination Venda-Cut-out workstation (coded as VC). The same module including TV camera and monitor on the portable stand with adjustable height was used with T-workstation and C-workstation. Therefore we had respectively VT and VC workstations. Design of the Venda workstations and easy transformation of traditional assembly and manual material handling workstations into the Venda-workstations are discussed in the paper by V. Venda and I. Venda in these Proceedings.

The experiments lasted one month, from February 9 - March 11, 1995. They were conducted in tree shifts: 1) 7 am - 3.30 pm; 2) 3.30 pm - midnight; 3) midnight - 7 am.Assembly operations at traditional workstations and new workstations (the Venda workstations) were assessed and compared using criteria of productivity (number of PCBs assembled in a shift), tempo of work (relative number of PCBs assembled using traditional workstations was considered as 1.00, all other numbers - for VT and VC, were compared with the tempo at the traditional workstation - T or C), left and right upper trapesius muscle strain (EMG as % of maximal voluntary contraction, MVC (Aaras, 1994)), angles of motions of neck and back from a straight position (EMG and angles of motions from straight neutral position were measured using the portable Physiometer PHY-400 manufactured by Premed, Norway, and widely used for assessment work postures and motions (Aaras, 1994).

Table 1 shows the comparison of productivity of assembly operations for traditional (T) workstation - at first day, and for the Venda workstation based on traditional one (VT) during second-eight days. Productivity at VT increases quickly and becomes higher than at T by the second training day. There is a dip in productivity on seventh day, when workers tried to do all operations, including soldering, exclusively using the indirect (TV) vision. They managed to trasform all their skills by the eighth day, when productivity proved to be higher than at traditional workstation (day 1). VT leads to a higher stability of productivity of different workers: this was evaluated by comparing the standard deviations SD for day 1 (T) to days 4 and 5. Table 2 shows the results on the comparison of productivity of work at advanced traditional cut-out workstations (day 1 at C) and Venda workstation based on this cut-out workstation (VC). Advanced traditional workstations (C) allow higher productivity than ordinary traditional workstations (T): 53.17 vs. 33.80. The productivity 53.17 was overcome using VC by the sixth day. The productivity 33.80 was overcome using VT by the second training day. But productivity 53.17 shown at C was not reached at VT even by the seventh training day. This means new principle of organizing assembly workstations has more advantages in combination with advanced traditional workstations.

Table 1. COMPARISON OF PRODUCTIVITY OF ASSEMBLY WORK WITH TRADITIONAL WORKSTATION ('T') AND V-WORKSTATION BASED ON THE TRADITIONAL ONE ('VT')

	Number of PCBs assembled in a shift							
WORKER #		DAYS at VT						
1	Day 1 at T	2	3	4	5	6	7	8
2	31.00	-	23.00	-	46.00	-	17.00	50.00
4	19.00	13.00	35.00	40.00	40.00	42.00	-	37.00
6	35.00	36.00	38.00	33.00	46.00	43.00	32.00	-
10	24.00	35.00	53.00	52.00	39.00	51.00	37.00	-
12	60.00	-	-	-	32.00	-	43.00	51.00
MEAN	33.80	28.00	37.25	41.67	40.60	45.33	32.25	46.00
S.D.	15.90	13.00	12.34	9.61	5.81	4.93	11.12	7.81

Table 2. COMPARISON OF PRODUCTIVITY OF ASSEMBLY WORK WITH CUT-OUT TRADITIONAL WORKSTATION ('C') AND V-WORKSTATION BASED ON THE CUT-OUT TRADITIONAL ONE ('VC')

	Number of PCBs assembled in a shift							
		DAYS at VC						
WORKER #	Day 1 at C	2	3	4	5	6	7	8
1	39.00	7.00	-	35.00	35.00	-	49.00	-
3	59.00	21.00	40.00	50.00	28.00	-	42.00	-
5	60.00	21.00	55.00	63.00	67.00	60.00	77.00	69.00
7	45.00		16.00	23.00	-	-	-	-
9	57.00	41.00	55.00	54.00	58.00	52.00	49.00	-
11	59.00	-	46.00	-	-	50.00	-	-
MEAN	53.17	22.50	42.40	45.00	47.00	54.00	54.25	69.00
S.D.	8.91	13.99	16.07	15.92	18.49	5.29	15.52	

Comparison of the traditional workstations (T and C) with V-workstations (VT and VC) showed a general advantage of VT and VC in most criteria: EMG of Right Trapesius, head and back movements (positions) (see Tables 3, 4 and 5). Individual differences are very essential, and additional experiments are needed to collect more statistics.

Table 3. COMPARISON "T" AND "VT" ON RIGHT TRAPESIUS (RT. TRAP) AND LEFT TRAPESIUS (LT. TRAP) MUSCLE STRAIN AND HEAD AND BACK FLEXION AND SIDE MOTIONS IN RELATIVE UNITS: (T-VT)/100

WORKER #	RT. TRAP.	LT. TRAP.	H. FLEX	H. SIDE	B. FLEX	B. SIDE
2	0.18	0.84	-0.01	-0.16	-0.21	0.23
4	0.90	-1.00	0.36	0.33	-0.05	0.48
6	1.03	-0.17	0.18	0.34	0.16	0.02
10	0.00	0.65	0.98	-0.05	0.24	-0.11
12	-0.08	0.87	1.45	0.80	0.49	0.40

Table 4. COMPARISON "T" AND "VT" ON RIGHT TRAPESIUS (RT. TRAP) AND LEFT TRAPESIUS (LT. TRAP) MUSCLE STRAIN AND HEAD AND BACK FLEXION AND SIDE MOTIONS IN RELATIVE UNITS: (C-VC)/100

WORKER #	RT. TRAP.	LT. TRAP.	H. FLEX	H. SIDE	B. FLEX	B. SIDE
1	0.47	0.04	0.00	0.00	0.00	0.00
3	-0.10	0.32	0.15	0.19	-0.70	0.31
5	0.27	0.90	0.43	0.33	0.33	0.27
9	-0.17	0.30	1.09	0.27	0.29	-0.18
11	-0.27	0.51	0.74	0.23	0.10	0.02
MEAN	n/a	0.41	0.48	0.20	n/a	n/a

Table 5. COMPARISON "T" AND "C" vs "VT" AND "VC" ON RIGHT TRAPESIUS (RT. TRAP) AND LEFT TRAPESIUS (LT. TRAP) MUSCLE STRAIN AND HEAD AND BACK FLEXION AND SIDE MOTIONS IN RELATIVE UNITS: T(or C) - VT (or VC)/100

WORKER #	RT. TRAP.	LT. TRAP.	H. FLEX	H. SIDE	B. FLEX	B. SIDE
1	0.47	0.04	0.00	0.00	0.00	0.00
2	0.18	0.84	-0.01	-0.16	-0.21	0.23
3	-0.10	0.32	0.15	0.19	-0.70	0.31
4	0.90	-1.00	0.36	0.33	-0.05	0.48
5	0.27	0.90	0.43	0.33	0.33	0.27
6	1.03	-0.17	0.18	0.34	0.16	0.02
9	-0.17	0.30	1.09	0.27	0.29	-0.18
10	0.00	0.65	0.98	-0.05	0.24	-0.11
11	-0.27	0.51	0.74	0.23	0.10	0.02
12	-0.08	0.87	1.45	0.80	0.49	0.40
MEAN	0.22	0.33	0.54	0.23	0.07	0.14

Conclusion

1. New type of assembly and manual material handling workstations was designed and tested. This invention helps to prevent RSI's (or CTD's) at the assembly and material handling plants and productions. The Venda-workstations allow the worker to be seated in a straight, neutral position. Industrial testing of the invention have proved its advantages in comparison with the existing traditional workstations. Potential users: companies producing assembly and material handling workstations; telecommunication, computer, electronic, audio-visual assembly companies; insurance companies; Workers' Compensation Boards (to return injured workers instead of paying them compensation for long time); manual material handling productions and plants. Patent Application on Assembly workstations with a negative tilt was officially filed at International Patent Office, London, England.

2. The new type of workstations significantly decreases muscle strain and, consequently, risk of RSI injuries.

3. The Venda-workstations allow a large magnifications of the product assembled or handled thus decreasing visual strain and increasing work productivity.

4. The Venda-workstations allow those workers who already had developed neck, wrist, back repetitive strain injuries and who cannot work at the traditional workstations to return back to work.

5. Ergodynamic principle of the testing of ergonomic design products reads as follows: identity between work functional structures of the test subjects and operators (workers) is required to get practically valid experimental ergonomic results which could be implemented to the practice.

6. Traditional (T and C) and the newly invented (the Venda, VT and VC) assembly workstations were tested and compared at Calgary Northern Telecom Wireles Systems plant for one month, three shifts a day. Twelve workers participated in testing. They were regular assembly slide-line workers and performed normal technological operations. Comparison of the workstations T vs. VT and C vs.VC was done using parameters of productivity of work (number of PCBs assembled in a shift), tempo of work as a number of PCB's assembled in one minute expressed in a relative values (data on workstations T and C were coded as 1.00); EMG-Right trapesius, EMG-Left trapesius expressed in %MVC, Neck Flexion, Neck Side Motions, Back Flexion and Back Side Motions expressed in the angles from the straight, neutral position. All measurement are presented in relative values: T(or C) - VT(orVC)/100. Most values at Tables 4-6 are positive. Therefore the new workstations required less muscle strain and deviations from the straight neutral position of neck (head) and back.

7. Industrial testing of the newly invented workstations have proved their advantages in comparison with existing traditional workstations (including advanced ones). All assembly operations including the insertion of electronic components, bending, cutting and soldering their leads were successfully done by all participating workers. Training period was very short and effective; it took 4-6 shifts. Training was done with minimal initial losses of productivity.

8. Productivity of work (as a number of PCB assembled during the shift) was higher when the Venda workstations were used: 46 for VT vs. 34 for T and 57 for VC vs. 53 for C.

9. All parameters registered using the PHY-400, including muscle strain, neck and back flexion, and side motions are better when the Venda workstations were used. Improvement is especially significant for the neck flexion (important factor causing neck injuries), and for EMG of the left trapesius.

10. The experiments have demonstrated that the new type of the workstations significantly increases work productivity and decreases muscle strain and consequently the risk of RSI's.

Acknowledgments

This studies are being supported by Northern Telecom Canada, Bell-Northern Research and Natural Science and Engineering Research Council of Canada. Our personal thanks go to Dr. James Laidlaw, Dr. Jack C. Dyment, Deborah Stokes, Bill Kukla, and Shona Anderson.

Ashish Kaushik, David Kuss and S. Marinov participated in conducting experiments, registration and processing data.

References

Aaras, A. 1994, The impact of ergonomic intervention on individual health and corporate prosperity in a telecommunications environment. *Ergonomics*, 37, NO 10, 1679-1696.

Fisher, D.L., Andres, R. O., Airth, D., and Smith, S. S. (1993) Repetitive Motion Disorders: The Design of Optimal Rate-Rest Profiles, *Human Factors*, 35(2), 283-304.

Marek, T., Wos, H., Karwowski, W., and Hamiga, K. (1992) Muscular Loading and Subjective Ratings of Muscular Tension by Novices when Typing with Standard and Split-Design Keyboards, *International Journal of Human-Computer Interaction*, v. 4, #4,387-394.

McAtamney, L. and Corlett, E. N. 1993, RULA: a survey method for the investigation of work-related upper limb disorders. *Applied Ergonomics*, 24(2), 91-99.

Schuldt, K., Ekholm, J., Harms-Ringdahl, K., Nemeth, G., and Arborelius U. P. (1986) Effects of changes in sitting work posture on static neck and shoulder muscle activity. *Ergonomics*, Vol. 29, No 12, 1525-1537.

Venda, V. F. (1994) Ergonomics: theory and tips. Manifesto of ergodynamics - Keynote Address for the World Ergonomics Congress IEA'94, Toronto, August, 1994. *Proceedings of the 12th Triennial Congress of International Ergonomics Association IEA'94,* HFAC, Toronto, 34-36.

Venda, Valery F. and Venda, Yuri V. (1995) *Dynamics In Ergonomics, Psychology and Decisions: Introduction to Ergodynamics*, Ablex, Norwood, N.J.

Venda, V. F., and Venda, Yuri V. (1991) Transformation dynamics in complex systems, *Journal of Washington Academy of Science*, #4, December.

Venda, Yuri V., and Venda, V.F. (1992) Introduction to the Transformation Dynamics, In: *Advances in Industrial Ergonomics and Safety-IV*, London: Francis and Taylor.