Linköping Studies in Science and Technology
Dissertation No. 1251

The use of structural equation modeling to describe the effect of operator functional state on air-to-air engagement outcomes

by

Martin Castor

Linköping University
INSTITUTE OF TECHNOLOGY

2009

Graduate School for Human-Machine Interaction
National Graduate School in Cognitive Science
Department of Management and Engineering
Linköping University, SE-581 83 Linköping, Sweden

"The use of structural equation modeling to describe the effect of operator functional state on air-to-air engagement outcomes"

Linköping Studies in Science and Technology Dissertation No. 1251

ISBN: 978-91-7393-657-6
ISSN: 0345-7524

Printed by: LiU-Tryck, Linköping

Distributed by:
Linköping University
Department of Management and Engineering
SE-581 83 Linköping, Sweden
Tel: +46 13 281000

Acknowledgements

Thank you professor Kjell Ohlsson, for supervision and help along the way.

Thank you professor Nils Dahlbäck, for supervision and valuable comments on the thesis draft.

Thank you Erland Svensson, Director of Research, for all your supervision and guidance. It has been an honor to have you as my supervisor and mentor, and I am very grateful for all the “vehicles of thought” and experiences you have instigated over the years.

Thank you Sofia, Jonathan and Joanna, for being my ultimate perspective providers.

Abstract

Computational evidence of the operative usefulness of a new system is crucial in large system development processes concerning billions of euros or dollars. Although it is obvious that the human often is the most important and critical component of many systems, it has often been hard for Human Factors researchers to express human aspects in a computational and strict way. The thesis describes, through data based statistical modeling, how the concepts or constructs of sensor effectiveness, usability of information, mental workload, situation awareness and teamwork relates to each other and to the operative performance in a fourship of fighter pilots. Through the use of structural equation modeling, ad modum LISREL, a statistical model that describes how the operators' functional state mediates the effects between technical system oriented variables, was developed.

The constructs used in the modeling process have received widespread scientific and operational attention. They have also been identified as multi-dimensional. Many different ways of measuring them have been developed in the scientific community, and the thesis focuses on the next step, i.e. how do these higher order constructs relate to each other in something as multi-dimensional as human activity in real situations?

A comprehensive human factors related dataset was collected in a large simulation based acquisition study that examined the requirements and properties of aircraft radar systems. The dataset contains 308 simulated engagements with data from four pilots each, i.e. 1232 cases in a database with 24 variables, generated by 37 pilots. The collected data and the resulting models thus summarizes more than 700 hours of experienced pilots' complex behavior in an operationally valid environment, and in a way that is of theoretical interest as well as of importance in system development processes. The data thus comes from a real world study of complex processes in a dynamic context, although from a simulator. The thesis is a case example of modern ecologically valid experimental psychology. The data collection does not represent a classical experimental setup, but instead demonstrates methodological needs and considerations for human factors practitioners when working in system development studies. The fact that parts of the used data are classified has not affected the models and scientific conclusions, although the practical findings have been partly circumscribed in the presentation.

As a result of the statistical modeling effort, a structural equation model of how the chosen constructs relate to each other, and mediate effects between technical measures by a model of the operator, is proposed. Simplicity of the model was the goal, and based on former experiences and findings, a simplex structure was hypothesized. The final model shows that the covariances between the 24 measures can be explained by a quasi-simplex structure of seven factors.

Sammanfattning

Statistiskt och matematiskt underbyggda slutsatser kring ett systems operativa användbarhet är kritiska i systemutvecklingsprocesser i mångmiljardklassen. Även om det är uppenbart att människan är den viktigaste och mest kritiska komponenten i många system har det ofta varit svårt för forskare inom området Människa System Interaktion (MSI) att uttrycka mänskliga faktorer på ett statistiskt och matematiskt strikt sätt. Den här avhandlingen visar och kvantifierar hur koncepten sensoreffektivitet, användbarhet hos information, mental arbetsbelastning, situationsmedvetande, samarbete och operativ prestation relaterar tilll varandra i en fyrgrupp militära flygförare genom modellering baserad på empiriska data. Genom att använda strukturella ekvationsmodeller, här m.h.a. LISREL, visas en statistisk modell av hur variabler som beskriver operatörernas förmåga att prestera medierar effekter mellan mer systemorienterade variabler.

De koncept eller faktorer som används i modelleringsprocessen har rönt stor vetenskaplig och operativ uppmärksamhet. De har också identifierats som mångdimensionella och inom forskningsområdet har en stor mängd olika mätmetoder utvecklats. Avhandlingen fokuserar på vad som sker i steget efter datainsamling, d.v.s. hur kan dessa faktorer relateras till varandra i något så mångdimensionellt som mänsklig aktivitet i verkliga situationer?

En omfattande beteendevetenskaplig datamängd samlades in under en stor simuleringsbaserad anskaffningsstudie som studerade krav för en ny flygplansradar. Databasen innehåller data från 308 simulerade flygföretag, med data från fyra flygförare per företag, d.v.s. 1232 rader i databasen med 24 variabler på varje rad, vilka genererats av 37 flygförare. Den insamlade datamängden sammanfattar mer än 700 timmar av erfarna flygförares arbete och prestation i en operativt relevant miljö på ett sätt som är både teoretiskt intressant och användbart i en systemutvecklingsprocess. Data kommer från en studie i verkligheten – även om verkligheten var simulerad – som handlade om komplexa processer i en dynamisk kontext. Avhandlingen är i sig ett exempel på en modern experimentalpsykologisk ansats som är giltig och användbar för tillämpade studier. Datainsamlingen representerar inte en klassisk experimentalpsykologisk ansats utan beskriver istället metodologiska behov och avvägningar som en MSI-forskare möter under arbete med systemutveckling. En delmängd av den insamlade datamängden är sekretessbelagd, vilket inte har påverkat modellerna eller de vetenskapliga slutsatserna, men vissa praktiska slutsatser har dock utelämnats.

Resultatet från modellutvecklingen är en strukturell ekvationsmodell som beskriver hur de utvalda koncepten relaterar till varandra och därigenom beskrivs relationen mellan tekniska mått m.h.a. en modell av flygförarna. Enkelhet och överblickbarhet i modellen var en del av målsättningen och baserat på tidigare erfarenheter användes en simplex struktur under modellkonceptualiseringsfasen. Den slutgiltiga modellen visar att kovarianserna mellan de 24 variablerna i databasen kan förklaras m.h.a. en kvasi-simplex struktur med sju faktorer.

Contents

Preface

The author's point of origin to the approach chosen in the thesis is based on a cognitive science background and more than ten years immersion in system development and research in applied settings. The real or simulated operational settings of different military operators, such as fighter pilots and fighter controllers, command and control staff, and tank crews, have been the research environments where the author have been active. The author's experience of applied science and the foundation of the current thesis are thus firmly rooted in what Hutchins (1995) would call a "cognition in the wild" approach. The immersion in these operational settings, and participation in a number of studies providing input to decisions concerning human performance and the functional state of operators, has shaped the methodological approach and frame of mind. During these studies, human performance data have in some ways been very hard to come by, and in other ways often overwhelmingly abundant. This has led to a very articulated need of further developed knowledge of data reduction and modeling procedures.

The constructs and the modeling approach used in this thesis stem from a tradition of very close work with expert practitioners of the military aviation field and support of high level decisions of the Swedish Air Force (Angelborg-Thanderz, 1982, 1989, 1990; Svensson, Angelborg-Thanderz, Sjöberg & Gillberg, 1988; Svensson Angelborg-Thanderz, Olsson & Sjöberg, 1993a; Svensson, Angelborg-Thanderz & Sjöberg, 1993b; Svensson & Angelborg-Thanderz, 1995; Svensson, Angelborg-Thanderz, Sjöberg & Olsson, 1997; Svensson, Angelborg-Thanderz & van Awermaete, 1997; Svensson, Angelborg-Thanderz & Wilson, 1999; Svensson, & Wilson 2002), recently summarized in Svensson, Angelborg-Thanderz, Borgvall & Castor (in press). The author has also been part of several methodology development projects that have compiled human performance measurement method overviews (Castor, et al., 2003; Alfredson, Oskarsson, Castor & Svensson, 2003).

1 Introduction

Woods, Christoffersen & Tinapple (2000) describe Human Factors as a research field and practice that is based on observing people at work. To the degree that one abstracts patterns from this process of observation, one can view Human Factors as the body of work that describes how technology and organizational change transforms work in systems. Wickens (1984) describe the goal of Human Factors to apply knowledge in designing systems that work, accommodating the limits of human performance and exploring the advantages of the human operator in the process.

O'Donnell and Eggemeier (1986) describe the primary concern of Human Factors engineering during system development and evaluation as to assure that the demands imposed by a system do not exceed the human operator's capacity to process information. At this time mental workload was the primary term used when referring to the degree or percentage of the operator's information processing capacity which is expended when meeting system and task demands. Since then, the scientific community has included other concepts or constructs that need to be taken into consideration when considering the human ability to interact with systems and with each other.

For questions concerning the cognitive aspects of Human Factors requirements, solid answers traditionally have been hard to find. When a system designer needs answers concerning the physical ergonomics of a population of pilots, figures and facts are available in databases on human anthropometry. However, when a designer wants to know how much information the pilots can manage, or how the operators' mental models of a computer-based automated system is affected by fatigue, usually only the vaguest of answers can be found. This fact has been observed by many Human Factors researchers and practitioners during the years, but the answers are still very vague. The search for appropriate and quantitative methods of data collection and analysis that can be usefully applied when addressing performance in the cognitive domain continues.

1.1 Rationale for thesis

In their Code of Best Practice for Experimentation, Alberts and Hayes (2002) describe how military experimentation in the past has focused predominantly on the assessment of technology, using traditional measurement approaches from the engineering sciences. Evidence from the past few years suggests that these traditional measurement approaches have yielded little more than anecdotal insight into how technology and human performance combine in either productive or unproductive ways. Such data, while interesting, does not provide decision makers with the quantitative foundation for assessing the return-on investment of various transformation initiatives.

In order to make progress in a system development process, you need some form of more or less explicit and valid feedback or metrics that indicate whether the process is moving along in the desired direction. If you cannot measure the impact of changes in any way, it is hard to make structured advances during a development process. The conceptual models of developers and decision makers of how the world works affect system development to a very large extent. Models of how the human operator interacts with a system or models of human performance are often of interest, but hard to formulate with a high degree of explicitness. And, even while psychologists may have elaborate theories concerning human activity, Hair, Anderson, Tatham & Black (1989) note that "users in the field" often exhibit as elaborate theories. Concepts are introduced and used without scientific tests of their validity and reliability. For example, military officers gladly use concepts such as survivability, lethality and sustainability when analyzing a military unit. These concepts may very well be useful in a development process, but for scientific progress it is very important that they also are evaluated with scientific rigor, and that effort is applied in order to transform the operational concepts into scientific concepts.

Ever since the earliest days of applied Human Factors, researchers and practitioners have used different concepts and approaches to understand and describe human work in order to provide input to the design of new systems and work practices. In parallel, the academic experimental psychology field has developed a large number of measurement techniques and statistical procedures to analyze many phenomena in, for example, human perception, cognition, and sociology, which can be labeled psychometrics.

Criticism that have been raised against these quantitative methods is that they answer questions such as "how long?" or "how often?", but rarely shed any light of the "why?" and "how?". This is probably the reason for the increased interest in ethnographic approaches. According to Carlshamre (2001) case research as a term emerged and was accepted in the human computer interaction field during the mid-1980s. It was then a means of describing a new and wider field of study than the psychological experiments of human information processing aspects that were typical for the 1970s and the early 1980s. Other descriptions of the development of the human computer interaction field and the movement towards more ethnographic and social team oriented approaches over the last decades can be found in Carroll (1997) and Bannon (2001).

There are a number of approaches to case research, where the researcher's amount of "intrusion" into the process studied varies. In one end of this spectrum we find the participant observation approach, which in turn could range from "the complete observer" to "the complete participant". Among the more extreme approaches we find interactive research and action research (e.g. Aagaard Nielsen & Svensson, 2006), where the researcher leaves the traditional role as independent observer and instead takes an active part in the ongoing processes, and contributes intentionally to the outcome of the activities. Here the researcher both participate in, and contribute to, a change in the community under investigation. As a consequence, he or she obtains an in-depth and firsthand understanding of the process.

Ethnomethodologically oriented researchers such as, for example, Lützhöft (2004) argue eloquently that the "engineering view", where knowledge that has not been accumulated through scientific measurement is seen as less useful or acceptable, is problematic. Lützhöft describes how ethnomethodological approaches can provide very valuable design input. She argues that the task is to establish a two way interpretive process where the researcher stands between end-users on the one hand and designers and developers on the other hand, to facilitate translation between what end-users mean and what designers and developers need to know. However, even while a study like, for example, Hutchins (1995) famous study of navigation practice provides a very informative and rich insight into the thinking and acting of the humans in focus, these approaches still do not provide the computational justification sought for in high stakes decisions. Also, the generalizability of this type of descriptive presentations in all its detail, with local and specific knowledge, and resulting usefulness in scientific theory building, can be discussed.

The study of human work and human action in real settings and situations, i.e. outside the laboratory, have received extensive attention during the last 20 years and a number of theoretical frameworks, e.g. naturalistic decision making (Klein, Orasanu, Calderwood & Zsambok, 1993), recognition primed decision making (Klein, 1989), distributed cognition (Hutchins, 1995), activity theory (Engeström, Miettinen & Punamäki, 1999), dynamic decision making (Brehmer, 1992), joint cognitive systems (Hollnagel & Woods, 2005), have been formulated. These frameworks focus, to different degrees, on the individuals cognition, the task/context, interaction between people, collaboration between individuals, and the artifacts and systems that are used to support the individuals thinking and communication.

The present author's personal experience is that the ethnographic and psychometric approaches often are described as pitted against each other as contrasting approaches and that a researcher must choose between them. The answer lies, as ever so often, somewhere in between, as they provide different types of answers and are useful in different phases of system development or scientific understanding. One of the goals of this thesis is to demonstrate that there is great merit in using second generation statistical techniques, which offer interesting possibilities even where classical experiments cannot be performed, as often is the case in "real world" studies.

For the constructs and measures in focus in this thesis there exist a large number of methodological reviews (e.g. Lysaght, et al., 1989; Harris, Hill, Lysagth & Christ, 1992; Alfredson, et al., 2003; Castor, et al., 2003; Wilson, et al., 2004), but the hausse seem to have been twenty years ago even though they are still discussed (e.g. Wickens, 2008; Durso & Sethumadhavan, 2008). Progress is slow, mainly consisting of methodological reviews and measurement refinements along with a debate concerning the scientific justification of the constructs.

This may be the results of a different number of reasons, e.g. a) the whole psychometric/human performance measurement approach is too positivistic, b) human behavior is too complex to capture other than qualitatively, c) the results are only of practical use within each particular system development process, and d) it is too resource demanding/hard to conduct experimental studies in real settings, and e) classical experimental requirements, e.g. full control of all variations in independent variables are hard to meet.

Another reason may be that the methodological approaches used make it hard to integrate experimental results from different human performance measurement studies with each other. Experimental design, data collection and analysis of human work, at least from realistic settings, are resource demanding activities. Given that the "chunk size" of typical Human Factors research projects, at least those the author are aware of, rarely are larger than, as a very gross approximation, a half to one million euros or dollars, the number of experiments that a group of researchers can perform within a project is rather limited.

Salas (2008), in his review of his work as the editor of the Human Factors journal, notes that there is an explicit need for a theory infusion in order for the Human Factors field to make progress, and for many applied contributions to provide additional value. Researchers of the field have often carefully studied one phenomenon or construct and tried to relate it to performance. But, analogous with the line of thought expressed by Newell (1973) in his famous "You can't play 20 questions with nature and win" paper, we need to develop a "general theory" of how different phenomena or constructs relate to each other. Cognitive and social processes interact with each other, and cannot be studied separately.

In order to make more progress it might be the case that we need to further develop methods for meta-analysis and model building to be able to compare and integrate results from the human performance measurement experiments that are being performed. We need to develop models that can be experimentally and statistically tested and compared, and ultimately rejected when our understanding has increased.

As a consequence of the scientific needs, challenges and earlier efforts mentioned briefly above, the primary research goals of this thesis are:

- To empirically show how a model of a human operators functional state, as expressed by the constructs of perceived usability of information, mental workload, situation awareness and teamwork, mediate between sensor effectiveness and operative performance. The goal is to develop a model that quantifies the relation between the constructs and expresses the relation between technological systems, individual's cognition and interaction within a team, and the performance of the total human-machine system. Theoretically, and almost philosophically, it is interesting whether the complex processes studied can be expressed in something as abstracted and simplified as a simplex structure.
- To further develop and demonstrate measurement and modeling methodology that is usable in quasi-experimental or natural settings, where for example the possibility to manipulate independent variables is low or non-existent.

Where many other theses dissect a concept or construct in depth, this thesis uses a "holistic" perspective that tries to integrate previous conceptualization of important human work processes. The basic approach has been to take the most commonly used reference or definition and not delve too deep into the concept. This is most evident in the case of situation awareness, where the conceptualization used is based on the most commonly cited reference (Endsley, 1995b). While introduction of this construct has resulted in a quite substantial amount of scientific debate over the last 20 years (e.g. Dekker & Hollnagel, 2004), and while other researchers have dug deeper and proposed other concepts such as situation assessment, sensemaking (Weick, 1995; Klein, Moon & Hoffman, 2006), or situation management (Alfredson, 2007), the "original" concept is used here.

The constructs used in the thesis and their measurement methods have received widespread attention. Many definitions exist, but still unified and overarching theories or models are lacking. Scientific contributions are still appearing, but interest seems to have tempered a bit. So, although the constructs used in the thesis are "old", they are still relevant, and the thesis is an attempt to stimulate the human performance measurement field.

1.2 Simulation based acquisition

Woods and Decker (2000) ask the question of how we can study a world, and its work practices that does not yet exist, and calls this "the envisioned world problem". This is necessary as new technology often transforms the nature of practice in a domain, with changes in both the cognition and collaboration of the practitioners.

One of the most concrete ways to study a future world is through the use of simulation. For material acquisition studies, the study or requirements definition for the systems of the future, this has been called Simulation Based Acquisition (SBA). The term has been seen in official documents since 1997 (DMSO, 1999; Sanders, 1999) and have been used rather frequently, at least in the military domain. The approach and "revolution in materiel acquisition" that SBA is presumed to represent, is a reaction to the fact that the development and purchase of modern weapon systems is becoming increasingly expensive and time consuming. The average development time of a new weapon system, like a new aircraft or ship is at least 8–10 years, and the time until the system is in operational use is 15–20 years. For example, the development of the Swedish JAS 39 Gripen aircraft was initiated around 1982, and the Gripen was operational, although not in all functions, in 1997.

In short, SBA implies that needs analysis, requirements analysis, development, and evaluation should be done with extensive support of simulation. One important component of SBA is also a new "procurement culture" where all stakeholders can affect the final product to a larger extent. Through SBA, input from all different scientific and engineering disciplines and operational requirements can meet early in a development process. Through simulation, researchers, developers and end-users can get the "touch and feel" of a system even though no physical system yet exists. In order for a development or acquisition project to be successful it is important that the stakeholders have a relatively unified view of requirements for the system, although the stakeholders may have partly different goals. In order to find this unified view, shared conceptual models are needed.

There are typically a large number of stakeholders involved in the kind of material acquisition processes that have been studied for the thesis and they are schematically described below and in Figure 1:

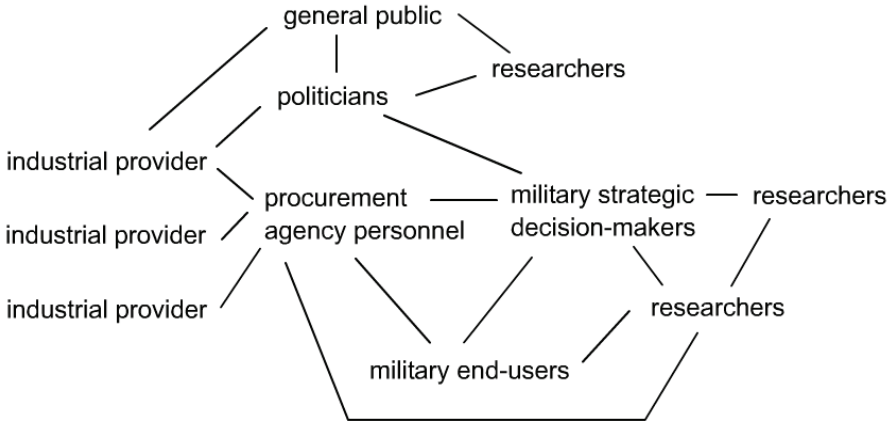


Figure 1. Conceptual sketch of different stakeholders in a military acquisition process.

- **Military end-users,** i.e. the operators that in the end will use the system or operational concept. As their life and mission success could depend on the performance of, for example, a sensor, their requirements are high and they rarely advocate the most inexpensive solution.
- **Military strategic decision makers**, i.e. the persons who look at the problem from the perspective of strategic goals and effects. In Sweden this would be the Supreme Commander and officers at the Armed Forces Headquarters. Their mission and concern are to get the most "bang-for-the-buck" for any investment, and they have the whole Armed Forces budget to consider when choosing a solution.
- **Procurement agency personnel**, i.e. the persons who directly order and manage the procurement process for a system. In Sweden this would be an employee at the Swedish Defense Material Administration, FMV.
- **Industrial providers**. A number of commercial companies typically provide solutions for any given problem and they, of course, want to sell their product or expertise.
- **Researchers** from several scientific disciplines that are supporting the process with methodological or technical expertise. In Sweden these researchers would typically come from the Swedish Defence Research Agency (FOI), or possibly from a university.
- **Politicians** as representatives of the **general public** and tax payers who provide the funding.

Between 2002 and 2005, a simulation based acquisition study concerning a new radar system for the Swedish JAS 39 Gripen aircraft was conducted. The study was conducted in order to define the technical and operational requirements and effects of different solutions. During the study eight different radar alternatives were evaluated. A number of different parameters and characteristics of the radar system were modified, with some alternatives providing radically new tactical capabilities for the pilots. Apart from increased knowledge on which parameters that affected the technical performance the most, a central question was to what extent the pilots could fully utilize the new capabilities.

The empirical data collection that was used for the current thesis was an integral part of this SBA study.

1.3 The nature of a model

Through observation of the world, i.e. the collection of data, science tries to explain how the world works. But, when the data has been collected much work remains and Ahl and Allen (1996, p. 45) describe the difference between data and models:

> *"Although they are a critical part of science, data are not the purpose of science. Science is about predictability, and predictability derives from models. Data are limited to the special case of what happened when the measurements were made. Models, on the other hand, subsume data. Only through models can data be used to say what will happen again, before subsequent measurements are made. Data alone predict nothing."*

The use of different modeling techniques to develop explanatory models is a core activity of most scientific studies. Harré (2002, p. 54.) states that, by definition, a model is a real or imaginary representation of a real system. Thus the basic logic of a model is an analogy in terms of patterns of similarity and differences between the model and whatever system or process that is modeled.

Models, as almost anything else, can be described on different levels of abstraction, and in order to exemplify, Figure 2 describe a representation, or model, of a snowflake on three levels of abstraction. All three of the models in the figure, to different degrees, capture essential properties of a snowflake, even while every real snowflake is said to be different. There are patterns which clearly identify it as a snowflake and these patterns are found in every snowflake, i.e. it is possible to describe models of what snowflakes looks like.

Figure 2. A representation, or model, of a snowflake described on three levels of abstraction.

Good models capture the essential properties of, and facilitate insight into, a system or process. Thereby models can be used as predictive tools, and be the base for important decisions. Regardless of simplicity, the model still needs to contain the essential information in order to be useful. As exemplified in Figure 2, the search for the "one and only" model or level of representation is "wrong", and the abstraction level of choice depends upon the purpose of the model. A model can, as shown, be described on different

levels of abstraction, and any model will face challenges regardless of level of abstraction. The model can be challenged because it fails to provide an idealization about the structure of the system, which approximates the actual behavior of the system good enough, or that it buries the important processes in a mass of "irrelevant" detail.

For a researcher it is a tradeoff where he or she tries to maximize explanatory power without making too much simplifications or violation of reality. It is the researcher who defines the frame of the model and chooses which factors to include, based on experience, good judgment, earlier scientific findings, theory, and model purpose.

A modeling effort is always part of a larger process. The purpose of the model is to be used, as a predictive tool in a practical application, as a "vehicle of thought" before a major decision, or as the current view of a phenomenon within a particular research field. An important step in the model development is thus to decide when a model is fit for purpose, and thereby practically and/or scientifically useful.

For the model(s) described in this thesis the purpose was: a) to describe the relation between the chosen modeling constructs in the simplest possible way, b) to justify them as separate constructs, c) to exemplify a modern approach to experimental psychology and how it is applicable in a SBA study.

1.3.1 Different types of modeling

The concept of a model and the process and purpose of modeling means rather different things to different researchers. During the ten years the author has been working with applied research, he has been involved in several projects that have been labeled as "modeling projects". However, the meaning and purpose of modeling has been quite different.

System development modeling frameworks/methods

In several systems development projects where the present author have been involved, different types of modeling methods and formalism have been used in order to define requirements for different systems. For example, the system development methods FEDEP (Federation Execution and Development Process), MODAF/DODAF (MoD/DoD Architecture Framework), and UML (Unified Modeling Language) have been used in these projects.

The products from this type of system development modeling are requirements documents and conceptual models describing potential users, interactions between systems, hardware and software needs, project risks, and so on from a number of different perspectives. These requirement documents are used in the communication between different stakeholders and the developers of a system.

Computational modeling

Within the field of human behavior representation or computational cognitive modeling, the goal is to develop computational models of human behavior and cognition. Ultimately the goal is to express a Unified Theory of Cognition (Newell, 1990), i.e. to find a

computational framework in which "all" cognitive processes can be expressed. The phenomena of interest for Human Factors and the human behavior representation community range over what Anderson (2002) calls "seven orders of magnitude", from neurological phenomena, best described on a timescale of milliseconds, to social behaviors, where hundreds of hours is a more appropriate timescale. As a consequence, a plethora of computational modeling architectures exists. A recent summary of state of the art is provided by the NATO RTO group HFM 128 Human Behavior Representation in Constructive Simulation (Lotens, et al., in press) and another in, the often cited, seminal work of Pew & Mavor (1998).

An example of computational cognitive modeling of the current air combat domain in the Soar framework is the TacAIR-Soar project (e.g. Coulter, Jones, Kenny, Koss, Laird & Nielsen, 1999; Laird, Coulter, Jones, Kenny, Koss & Nielsen, 1998), in which agents that could fly all types of US Air Force missions, based on a large set of rules, were developed. Computational modeling has many times been used to express phenomena on a cognitive level. Predictive cognitive models of attention and planning relevant for the thesis are relatively commonplace in the computational cognitive modeling literature (e.g. Doune & Sohn, 2000).

Statistical modeling

The focus of this thesis is empirically based statistical modeling and how we can quantify and relate different phenomena to each other with the help of statistics. A large number of statistical methods have been developed, to analyze bivariate and multivariate relations (e.g., Tabachnick & Fidell, 1996). Given that the thesis is a case description of a statistical modeling effort, this type of modeling is not elaborated further in this section.

System development modeling, computational cognitive modeling, and statistical modeling, in theory, should be intertwined, e.g. with decisions and descriptions in system modeling, based on validated human performance data. Similarly, all the if-then rules[1] in a computational cognitive model probably should be based on statistical models of human performance. The present author's experience is that these types of datasets and models are very hard to find when cognitive phenomena and performance are the concern[2].

[1] If a production rule system, e.g. Soar, is used. Other solutions exist.

[2] Apart from data and models of rather isolated phenomena, e.g. sleep deprivation (e.g. Gunzelmann, Gluck, Price, Van Dongen, & Dinges, 2007).

1.4 Structural equation models

The phenomena or concepts of interest to human factors researchers are often not directly measurable. In statistics these abstract phenomena have been called latent variables, factors or constructs. Examples of latent variables in psychology are, e.g. different types of intelligence and personality. The same is true for the constructs used in this thesis and the label construct or latent variable is used from here on.

A clear example and analogy of a latent variable from the physical sciences is provided in Wilson et al. (2004). The temperature can be measured by a number of different scales such as the Kelvin (K), Réaumur (R), Fahrenheit (F) and Celsius (C) scales. However, the manifest and measurable variation in the scales is a consequence of the amount of excitation of nuclear particles, and it is not the movement of the particles that is measured or observed directly. Thus, temperature can be considered as the hypothetical phenomenon affecting and explaining the covariation in the scales, and a latent variable or factor which finds manifest expression on the different scales, as illustrated in Figure 3.

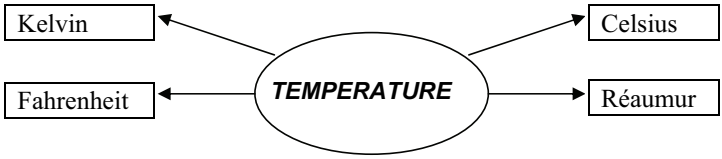


Figure 3. The construct of temperature and (some of) its manifest measures.

The causal relationship between constructs and quantification of the effects between them is of interest to almost all researchers, regardless of discipline. Sewell Wright invented path analysis (1918) as a methodology to analyze systems of structural equations and to describe how a number of interesting constructs relate to each other. Three of the most important aspects of path analysis are the path diagram, the equations relating correlations or covariances to parameters, and the decomposition of effects (Bollen, 1989).

Methodological desires by researchers using path analysis have for example been: a) to be able to measure the latent variables of interest through multiple manifest variables in order to get better measurement, b) to be able to accommodate for measurement error, and c) to be able to statistically compare alternative models.

Structural Equation Modeling (SEM) is a quantitative statistical method that was developed to satisfy these methodological desires. SEM combines the benefits of path analysis, factor analysis and multiple regression analysis (Jöreskog & Sörbom, 1984, 1993; Tabachnick & Fidell, 1996). SEM is based on correlational statistics, i.e. the linear relationships between variables, and the common variance between the variables forms the basis for the analyses. SEM analyses and presents the degree of relationship between

variables in terms of explained variance. A hypothesized model is tested statistically in a simultaneous analysis of the entire system of variables, to determine the extent to which the covariance or correlation matrix stipulated by the model, is consistent with the matrix based on the empirical data. If the statistical goodness of fit between the two compared matrices is adequate the model is a plausible representation of the relations between variables that the model developer has specified.

While most other multivariate procedures essentially are descriptive by nature (e.g. exploratory factor analysis), SEM takes a confirmatory (i.e. hypothesis-testing) approach to data analysis, although exploratory aspects can be addressed. Whereas traditional multivariate procedures are incapable of either assessing or correcting for measurement error, SEM provides explicit estimates of these parameters.

Hoyle (1995) describes three main differences between structural equation modeling and other approaches. First, SEM requires formal specification of a model to be estimated and tested. It forces the model developer to think carefully about their data and to formulate hypotheses regarding each variable. Second, SEM has the capacity to estimate and test relationships between latent variables. Third, SEM is a more comprehensive and flexible approach to research design and data analysis than any other single statistical model in standard use by social and behavioral scientists. Hoyle also describes SEM as similar to correlation analysis and multiple regression analysis in four specific ways. First, SEM is based on linear statistical models. Second, requirements such as independence of observations and multivariate normality will have to be met. Third, SEM promises no test of causality. It merely tests relations among different variables. Finally, like any other quantitative analysis, post-hoc adjustments to a SEM model require cross-validations.

The development of a structural equation model is supported by special software packages. The first and most widely spread software is LISREL (Jöreskog & Sörbom, 1984; 1993; SSI, 2008) which is an acronym for Linjära Strukturella Relationer (LInear Structural RELations). LISREL was originally developed by the two Swedish professors Karl Gustaf Jöreskog and Dag Sörbom. One of the earliest references to LISREL methodology is Jöreskog (1973), and since then LISREL has been developed in several generations. Currently version 8.80 is the latest version. Several other software packages have been developed, where AMOS (SPSS, 2008) and EQS (MVSOFT, 2008) probably are the most widely spread, apart from LISREL.

Structural equation modeling have been used for many years and is a popular methodology for non-experimental research, where methods for testing theories are not well developed, and ethical or practical considerations make traditional experimental designs unfeasible. Human factors researchers at the Swedish Defence Research Agency (FOI)[3], Maud Angelborg-Thanderz and Erland Svensson, have used LISREL since 1984.

A structural equation model may have one or more components. One component that is present in all structural equation models is the measurement model that defines the latent constructs through different manifest variables. Another important component is the

[3] Formerly the Swedish Defence Research Establishment (FOA).

structural model. The structural model tests relationships between the different latent variables.

Measurement model

The measurement model is the part of a SEM model which defines relations between the latent variables or constructs and their manifest variables. The manifest variables are often the items/questions of a questionnaire, but can be any type of measured data. In order to provide a well rounded measurement of the construct the manifest variables should be chosen or designed so that they assess different aspects of the construct, i.e. the manifest variables should not be too similar. A pure measurement model represents a confirmatory factor analysis (CFA) model in which there is undetermined covariance between each possible pair of latent variables. The pure measurement model is frequently used as the "null model", where all covariances in the covariance matrix for the latent variables are all assumed to be zero, i.e. the constructs are totally unrelated to each other. In order for the proposed structural model, i.e. the part where relations between the constructs are hypothesized, to be investigated further, differences from the null model must be significant.

Structural model

The structural model describes how the researcher has defined the relationships between the latent factors. It consists of a set of exogenous and endogenous latent variables in the model, together with the direct effects connecting them, and the error variance for these variables. The error variance reflects the effects of unmeasured variables and error in measurement. The exogenous latent variables are those that are conceptualized as to cause variance in the values of other latent variables in the model. Changes in the values of exogenous variables are not explained by the model and they are considered to be influenced by factors external to the model. Endogenous latent variables, those that are influenced by the exogenous variables in the model, either directly, or indirectly affect each other. Variance in the values of endogenous variables is considered to be explained by the model because all latent variables that influence them are included in the model specification. Diamantopoulos & Siguaw (2000) state that models with five to six latent variables, each measured by three to four manifest variables can be considered an appropriate upper level of complexity. Many models found in the literature are not as complex and consist of two or three latent variables. Increases in model size typically results in increasing difficulty to meet the recommended thresholds for model fit.

Residual and error terms

For the majority of variables that are of interest within Human Factors it is very difficult to design measures that will measure a phenomenon perfectly. Thus, error in measurement is assumed, and in structural equation modeling this is addressed by the inclusion of error terms for each variable. Residual error terms reflect the unexplained variance in latent endogenous variables due to all unmeasured causes.

1.5 Structural equation model development process

Structural equation modeling (SEM) is almost a research field in itself and therefore only a brief introduction to the model development process is provided here. Introductionary texts concerning the SEM development process accessible for non-experts, are, for example, provided in Diamantopoulos & Siguaw (2000), Byrne (1998, 2001), and on the Internet (Garson, 2008).

Jöreskog (1993) distinguishes between three scenarios of SEM use that he termed Strictly Confirmatory, Alternative Models, and Model Generating. In the Strictly Confirmatory scenario the researcher formulate a single model based on theory, collects the appropriate data, and then test the fit of the model to the collected data. The researcher does no modifications to the model and either accepts or rejects the model. However, as other unexamined or nested models may fit the data as well or better, an accepted model is only a model that has not been rejected.

In the Alternative Models scenario the researcher proposes several alternative competing theory-driven models. Based on the analysis of the collected data, the most appropriate model is chosen. Although this approach is desirable in principle, a problem is that in many specific research topic areas, the researcher does not find two or more well-developed alternative models to test.

In the Model Generating scenario, the researcher proceeds in a more exploratory fashion, often after first having had to reject an initial model after assessment of its poor fit. Jöreskog notes that although respecification may be either driven by theory or data, the goal is to find a model that is meaningful and statistically well fitting. The problem with the model development approach is that models developed in this way are post-hoc models, which may not be stable and may not fit new datasets. By the use of a cross-validation strategy, where the initial model is developed using one data sample and then tested against an independent sample, some of this concern can be addressed. For the model(s) presented in this thesis, the approach, as in many cases, most closely matches the Model Generating scenario.

Regardless of which of these three approaches that have been chosen, SEM does not in itself provide clues concerning causality in a model, i.e. in what directions the effects go (and specifically in the modeling software, in which directions the arrows point). The causality has to be justified by theory and the good judgment by the researcher.

In a description of the SEM development process, Jöreskog & Sörbom (1993) describe the validation of the measurement model and the fitting of the structural model as the two main steps. The validation of the measurement model is accomplished primarily through confirmatory factor analysis, while the fitting of the structural model is accomplished primarily through path analysis with latent variables. The model that is being developed is specified on the basis of available theory. Constructs are chosen and operationalized by multiple manifest variables and tested through confirmatory factor analysis to establish

that indicators seem to measure the corresponding constructs. The researcher proceeds to development of the structural model only when the measurement model has been validated. Two or more alternative models (one of which may be the null model) are then compared in terms of model fit, which measures the extent to which the covariances predicted by the model correspond to the observed covariances in the data. Modification indexes, suggested by the analysis software, may be used by the researcher to alter one or more model specifications to improve fit, but only if supported by theory.

A solid theoretical foundation is thus needed before a structural equation model is developed, as theory warns us of potential problems such as, for example, excluded variables. Theoretical support is also necessary in order to distinguish between statistically equivalent models. Good definitions are also helpful when identifying appropriate manifest variables/measures.

In another description of the SEM development process Diamantopoulos & Siguaw (2000) describes eight relatively distinct but related steps that a researcher goes through when developing a structural equation model:

1. Model conceptualization
2. Path diagram construction
3. Model specification
4. Model identification
5. Parameter estimation
6. Assessment of model fit
7. Model modification
8. Model cross validation

Brief descriptions of the basic outline and considerations of each of Diamantopoulos' & Siguaw's steps will be provided below.

1.5.1 Model conceptualization

In this initial step the researcher define his or her conceptual model, which translates theoretical assumptions into a conceptual framework. This conceptual model needs to be identified based on existing literature and theory. In this step, the researcher decides which latent variables or constructs that will need to be included, and how they are to be operationalized through manifest variables. During this stage, it is crucial to make every effort to include any important factors that can affect the variables that are included in the model. An omission of important factors represents a specification error and the result can be that the proposed model in the end does not represent the "whole" truth.

Successful development of a structural equation model is to a large extent based on a sound model conceptualization. It is rare that a modeling process that does not start from well established concepts or constructs and tested measures, result in a useful model.

1.5.2 Path diagram construction

In this second step of the modeling process the model developer can describe his or her model graphically as a path diagram. This is not a mandatory step, but it is helpful in order to make the model more explicit for the model developer.

1.5.3 Model specification

The third step is model specification, where the researcher specifies which effects that are null, which are fixed to a constant and which ones that vary through the specification of a command file for the analysis software. The researcher now needs to be very explicit on which variables that will be included and how they shall relate. The specification of a command file can be either through a text or a graphical format.

Effects are represented by an arrow in a path diagram, while null effects result in the absence of an arrow. Note that the existence or absence of an arrow represents a rather strong theoretical assumption. A model where no effect is constrained to zero will always fit the data, and the closer one is to this most complex model, the better the fit of the model to the data. Thus, for a model where many effects are included in the specification, the fit indices reported (see section 1.5.6) are better, but the model is also more complex and harder to grasp for the researcher.

1.5.4 Model identification

The fourth step in the process is model identification, which is performed by the analysis program, e.g. LISREL or AMOS. In this step the empirical data is investigated to see whether there is enough information in the data to do the parameter estimation that is performed in the next step, i.e. that a unique value can be identified for each parameter in the model. If there is a lack of information, i.e. the number of parameters estimated is less than the number of variances and covariances, the model, becomes under-identified and the analysis is cancelled. The model can also become just-identified or over-identified. If the number of parameters estimated are greater than the number of variances and covariances then the model is over-identified.

To exemplify what is done during the model identification the following simple example can be used: Is there enough information to uniquely identify the values of A and B in the equation A * B = 100? The answer is no, as there are several different possible solutions and this would equal to when a model is unidentified. However, if A is fixed to 10 you know that B has to be 10 and the equation can be identified.

1.5.5 Parameter estimation

If the model can be identified, the parameter estimation step can be executed. During the parameter estimation the analysis software create a covariance matrix based on the specified model. If there is no relation between two variables specified during the model specification the covariance is set to zero. The covariance matrix that is proposed by the model is then compared to the matrix produced by the data.

The selection of method of estimation is also an important component of the model specification. Several methods of estimation can be used and ordinarily one will get

similar estimates by any of the methods (Garson, 2008). Maximum Likelihood estimation is by far the most common method and Garson (2008) recommends that it is used, unless the researcher has good reason or counterarguments. Unlike some of the other estimation methods, Maximum Likelihood does not assume uncorrelated error terms. Key assumptions are large samples, manifest variables with multivariate normal distribution, valid specification of the model, and manifest variables on an interval or ratio scale, although ordinal variables are widely used in practice. If ordinal data are used, they should have at least five categories and not be strongly skewed.

1.5.6 Assessment of model fit

Once a model converges and parameter estimates are presented, the question is to what extent the empirical data fit the proposed model. In other words, how well the correlation or covariance matrix produced by the data matches the matrix that is implied by the model. Assessment of model fit is one of the more complex tasks of a SEM analysis. Model fit is related to data, model, and estimation methodology and a plethora of fit indices has been developed over the years.

Jaccard and Wan (1996) describe three classes of fit indices (absolute, parsimonious, and relative) that should be considered when evaluating the fit of a structural equation model. Absolute fit compares the predicted and observed covariance matrices. The chi-square (χ^2), goodness of fit index (GFI), and standardized root mean square residual (Standardized RMR) are indicators of absolute fit.

Large values of chi-square reflect a discrepancy between the observed and predicted matrices. The chi-square is reported with the number of degrees of freedom associated with the model, and a significance test. The degrees of freedom are a function of the number of covariances provided and the number of paths specified and a statistically significant model suggests that the specified paths do not provide a perfect fit to the data. Hence a non-significant value ($p > 0.05$) is desired, but Hair et al. (1995) note that the chi-square is sensitive to sample size and that it is rare to find a non-significant value when sample size is over 500 cases.

The GFI is a function of the absolute discrepancies between the observed and predicted covariance matrices. The recommended threshold for the GFI is 0.90. GFI is sensitive to sample size.

The Root Mean square Residuals (RMR) are the coefficients which result from taking the square root of the mean of the squared residuals, which are the amounts by which the sample variances and covariances differ from the corresponding estimated variances and covariances. The standardized RMR (S RMR) is the average difference between the predicted and observed variances and covariances in the model, based on standardized residuals. The recommended threshold for the standardized RMR is 0.05.

The second category also considers absolute fit, but penalizes model complexity. The more paths specified, the lower the models' parsimony. The Root Mean Square Error of Approximation (RMSEA) is the common choice for measure of parsimony. The RMSEA

fit index is by default reported by LISREL and values approaching zero are desired. Many recommendations state that it should be less than 0.05 in order to represent a good model fit, but for example Bollen (1989) and Browne and Cudeck (1993) state that a value of 0.08 or less could be considered acceptable. RMSEA is sensitive to sample size.

The third category of fit scales compares the absolute fit to an alternative model. The relative goodness of fit measures compares the evaluated model to the fit of another model. When none is specified, the analysis software packages usually default to comparing the model with the independence model, or even allow this as the only option. The Comparative Fit Index (CFI) is a commonly used fit index and Byrne (1998, p. 270) suggest that the CFI should be a fit statistic of choice. The value for the CFI indicates the fit of the model compared to the null model and the recommended threshold is 0.90.

A number of measures based on information theory have also been developed. These measures are appropriate when comparing models which have been estimated using maximum likelihood estimation. They do not have thresholds, like 0.90, and rather they are used when comparing models, with a lower value representing a better fit. AIC is the Akaike Information Criterion and is a goodness-of-fit measure which, adjusts model chi-square to penalize for model complexity. CAIC is the Consistent AIC, which penalizes for sample size as well as model complexity.

Most important when considering different fit indices, and expressed by Byrne (1998, p. 199), is that model adequacy should be based on theoretical, statistical as well as practical considerations. Thus, the causal logic and good judgment of the model developer can never be underestimated. This has also been emphasized from the beginning by Jöreskog and Sörbom, the LISREL developers.

1.5.7 Model modification

When a model have been evaluated with respect to its fit, the modeler can decide whether the model is acceptable or that it needs to be modified in order to fit the empirical data better. LISREL presents suggestions for model improvement, so called modification indices. These modifications are entirely data driven and careful deliberation and theoretical support must substantiate any changes to the model based of the modification indices.

1.5.8 Model cross-validation

The last step of the modeling process is to do cross-validation of the proposed model against a new dataset, or a part of the dataset that have been kept aside for cross-validation purposes. This step is extra important if major changes have been done to the model as a result of the model modification phase.

1.5.9 Guidelines for model development

There are a number of issues to consider when developing a model, which hopefully is evident from the above description of the development process. Thompson (2000, p. 231-232) has suggested the following 10 guidelines when developing and reporting structural equation models:

- Do not conclude that a model is the only model to fit the data.
- Test respecified models with split-halves data or new data.
- Test multiple rival models.
- Use a two-step approach of testing the measurement model first, then the structural model.
- Evaluate models by theory as well as statistical fit.
- Report multiple fit indices.
- Show that you meet the assumption of multivariate normality.
- Seek parsimonious models.
- Consider the level of measurement and distribution of variables in the model.
- Do not use small samples.

These guidelines have been followed in the model development of the thesis.

1.6 The fighter aircraft operations domain

Human work and work domains can be described in many different ways. In Annett & Stanton (2000), many different types of task analysis formalisms are described. Cognitive Work Analysis (Vicente, 1999), activity theory (e.g. Engeström, Miettinen & Punamaki, 1999), descriptions of joint cognitive systems (Hollnagel & Woods, 2005) represent other approaches that can be used when describing human work.

Westlander (1999, p. 123) presents useful terminology that can be used to describe the range of the context and how dividing lines may be drawn in relation to the chosen unit of analysis: a) job task, b) job content, c) work situation, d) organizational activities, e) specific environment, and f) general environment. The description of the domain in this section attempts to capture essential properties of the fighter aircraft operations domain on several of Westlander's levels.

Doctrinally the missions used in the current study would represent the "fighter escort" and "air defense" mission types under the Offensive Counter Air (OCA) and Defensive Counter Air (DCA) roles (Swedish Armed Forces Headquarters, 2005). The primary task of the fighter aircraft in OCA or DCA roles are to gain and maintain air superiority (i.e. deny enemy aircraft the possibility to operate) in order to protect the own airspace from enemy actions (reconnaissance or attack from enemy aircraft and missiles), and to protect the airspace for own missions (attack or reconnaissance).

With the modern medium range missiles, (e.g. AIM 120 AMRAAM), and modern sensors, (e.g. aircraft radar, currently the PS05 pulsedoppler radar in JAS 39 Gripen), and tactical datalinks, the main engagement scenario during these types of missions probably would be a Beyond Visual Range engagement.

Beyond Visual Range (BVR) means a scenario where the enemy is engaged before they can be seen visually. This is possible due to the performance of the sensors available to the formation, either their own sensors (primarily the aircraft radars), and the sensors and information available to the fighter controller. Political implications of downed aircraft and the resulting Rules of Engagement (RoE) might force a Within Visual Range (WVR) engagement, but for the missions used in the SBA radar study, BVR engagements were in focus. Regardless of engagement scenario the pure tactical goal would be to shoot down as many as possible of the enemy aircraft, while not getting shot down yourself. To describe and explain something of the tasks of the pilots, excerpts from the public Wikipedia articles on BVR and the AIM120 missile are provided[4]:

[4] The present author refrains from providing a detailed description of current Swedish BVR tactics, but a non-classified description, sufficient for the current purpose, of BVR engagements and the AIM120 missile can be found on Wikipedia.

BVR

A main engagement scenario is against aircraft also armed with fire-and-forget missiles. In this case engagement is very much down to teamwork and could be described as "a game of chicken." Both flights of aircraft can fire their missiles at each other beyond visual range (BVR), but then face the problem that if they continue to track the target aircraft in order to provide mid-course updates for the missile's flight, they are also flying into their opponents' missiles. If the enemy fires missiles at maximum range, you will be able to defeat them easily without having surrendered valuable ordnance yourself.

AIM120

Once in its terminal mode, the missile's advanced electronic counter countermeasures (ECCM) support and good maneuverability mean that the chance of it hitting or exploding close to the target is high (on the order of 90%), as long as it has enough remaining energy to maneuver with the target if it is evasive. The kill probability (Pk) is determined by several factors, including aspect (head-on interception, side-on or tail-chase), altitude, the speed of the missile and the target, and how hard the target can turn. Typically, if the missile has sufficient energy during the terminal phase, which comes from being launched close enough to the target from an aircraft flying high and fast enough, it will have an excellent chance of success. This chance drops as the missile is fired at longer ranges as it runs out of overtake speed at long ranges, and if the target can force the missile to turn it might bleed off enough speed that it can no longer chase the target.

The launch distance depends upon whether the target is heading towards or away from the firing aircraft. In a head-on engagement, the missile can be launched at longer range, since the range will be closing fast. In this situation, even if the target turns around, it is unlikely it can speed up and fly away fast enough to avoid being overtaken and hit by the missile (as long as the missile is not released too early). It is also unlikely the enemy can outmaneuver the missile since the closure rate will be so great. In a tail-on engagement, the firing aircraft might have to close to between one-half and one-quarter maximum range (or maybe even closer for a very fast target) in order to give the missile sufficient energy to overtake the targets.

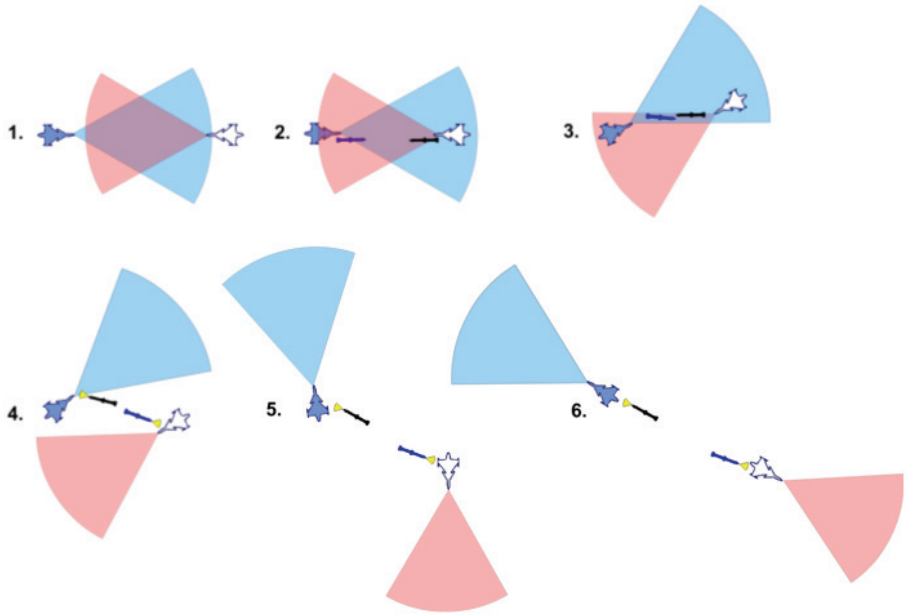


Figure 4. Schematic “steps” in a BVR duel between two aircraft.

In Figure 4 the “steps” of a schematic, but typical, BVR duel with modern “fire and forget” medium range missiles are depicted. At 1) the two aircraft are just about to detect each other through the aircraft radar (the cone in front of the aircraft symbol represents the radar coverage). At 2) the two aircraft have detected each other and have launched a missile each. At 3) the two aircraft maneuver in order to keep radar contact with their target until the missile opens the missile seeker, while avoiding to fly right into the missile that the other aircraft presumably have launched (the pilot has no way of knowing whether a missile really has been launched toward him). The pilots are now doing so called gimbal maneuvers, waiting for the missile to open its seeker. When it opens they no longer need to support the missile with their radar. At 4) both missile seekers open (the small cone in front of the missile symbol). The pilots now get warnings through their radar warning receivers that a missile seeker has opened close to them, meaning that they probably will get hit if they do not do anything. At 5) both pilots do evasive maneuvers to get away from the missiles. At 6) they are both being chased by the missiles, until the missiles terminate (due to a number of possible abort criteria) or hit the aircraft. The decision to launch, to start evasive maneuvering, and to turn back into the duel are three critical decision points.

The outcomes of air-to-air fighter aircraft engagements are influenced by many factors of both technical and Human Factors nature. A non-classified description of some relevant technical and tactical parameters of air combat is provided by Johansson (1999):

Jamming, electronic countermeasures and radar warning receiver

- Range of own jamming versus enemy radar
- Radar warning receiver range and accuracy

Radar

- Radar modes (e.g. Track While Scan, Continuous Wave)
- Aircraft radar range and angle of coverage

Weapon characteristics

- Mean speed
- Maximum range
- Minimum range
- Missile seeker opening distance

Aircraft behavior and tactics

- Absolute and relative altitude of aircraft in engagement
- Absolute and relative speed of aircraft in engagement
- Thrust
- Geometry, i.e. the relative positions and ranges of the aircraft within the twoship or fourship to the enemy, and to other friendly aircraft, e.g. attack aircraft that are being escorted
- Aggressive or defensive stance and risk taking
- Intentions and Rules of Engagement for both sides
- Active versus passive use of radar

Numerical superiority

- Number of own aircraft
- Number of enemy aircraft

Command and Control

- Radar coverage for air surveillance radars (i.e. other friendly radars on the ground or in the air)
- Fighter controllers' threats classification capabilities

Loadout on aircraft

- Fuel
- Weapons loadout

For a long time, the Swedish Armed Forces have during mission execution been practicing "uppdragstaktik"[5]. When "uppdragstaktik" is used, the commander of a mission is given very extensive authority with regards to the specifics of mission execution, and essentially only the mission goal and resources are provided by the higher command levels. The basic idea supporting "uppdragstaktik" is that the mission commander usually has the best overview of the local situation, and that distribution of command authority down to the executing level leads to better results and a higher tempo in operations.

During mission execution and the tactics development in the Swedish Air Force the characteristics of "uppdragstaktik" have been very prominent, and have been the basis of both the doctrine and the practice of Swedish Air Force operations for a long time. This fact is, to a large extent, due to the use of a technical system called tactical data links. Tactical data links have been used for many years to send information between aircraft and between aircraft and fighter controllers on the ground. In the field of tactical data links, Sweden has been leading and was the first country in the world to introduce them. Other countries have only recently started to introduce them in their military aircraft. Supported by the technical possibilities of the tactical data links, Swedish pilots have developed what they sometimes informally refer to as a "floating decision method" during the execution of tactical missions. The practice of this "floating decision method" is very evident in BVR scenarios. The pilot with the best grasp of the situation, or best situation awareness, and the best possibility to engage, acts regardless of his seniority in the formation, and the other pilots adapt their behavior.

The environment, the technical systems and the job tasks results in a work environment that is very dynamic and fast-paced. However, at the same time, the basic schema or outline, with plausible and potential outcomes of a BVR engagement are known to the pilots. One important characteristic is that the pilots do not have the possibility to stop the aircraft to "gather their thoughts" or analyze the situation. The decisions the pilots are made in real time and there is no "no-action alternative" and thus the pilots are constantly acting in some way. Usually a series of interconnected decisions are needed to manage a situation. Many of the decisions concern the temporal dimension, i.e. whether it is optimal to act now or to wait, rather than being a choice between which actions to execute. The most important characteristic is that during military operations there is always a thinking enemy who actively tries to counter the effects of any actions that the pilots execute.

The fourship[6] or the twoship formations are the basic building blocks used in a mission. Only rarely does a pilot operate by himself. In addition to the pilots, a fighter controller is an important fifth or third member of the group. The fighter controllers sit in a bunker underground or in a flying command and control aircraft, and have more powerful long-range sensors than the fighter aircraft sensors at their disposal.

[5] "Uppdragstaktik" have here been kept in its original Swedish form instead of being translated into "mission tactics", as this would imply a too broad meaning in English.
[6] A fourship could for example be four pilots with four JAS 39 Gripen aircraft.

In an engagement, superiority in numbers is highly desirable. For example, in a fight between two fourship formations, the fight quickly collapses for the side that first loses one aircraft. This is due to the fact that the fight essentially can be described as a number of one-on-one duels, see Figure 4. If one side has one aircraft more than the other side there is a “free radical” that can outmaneuver aircraft on the other side.

The financial cost of four JAS 39 Gripen aircraft departing for a mission, including weapons, fuel, and training costs for four pilots would be at least 1500 million Swedish crowns. The implications of erroneous action, in terms of human tragedy for pilots or civilians can, of course, be very severe, by any means of counting.

The members of a formation are interdependent with regards to what Matheiu, Marks and Zaccaro (2001, p. 295) refer to as “input interdependence”, which identifies to what extent teams must share inputs such as people, facilities, environmental constraints, equipment, and information related to collective goal achievement. The weapons of the fourship formation can be considered one such shared resource. Every JAS 39 Gripen has four main pylons underneath the aircraft, and thus can carry, for example, four medium range missiles, apart from two short range missiles on the wing-tips. A fourship formation can thus carry a maximum of 16 medium range missiles and every missile launch decreases the strength of the unit rather considerably as the combat strength of one aircraft can be considered to decrease by 25% after each one of the missiles has been fired. The decision to launch is therefore one of the main decision points during a BVR engagement.

As hopefully evident from the description above, a number of relationships in both the physical domain (e.g. the relative positions of the aircraft), the information domain (e.g. how much and how information is handled and presented in the aircraft) and the cognitive domain (e.g. the functional state and performance of the pilot), are of vital importance to understand the outcome of air-to-air engagements.

1.7 Relevant modeling constructs

The theoretical constructs used in this thesis are founded on a long tradition of close co-operation with the Swedish Air Force (Angelborg-Thanderz, 1982, 1989, 1990; Svensson et al., 1988; Svensson et al., 1993a; Svensson, et al., 1993b; Svensson & Angelborg-Thanderz, 1995; Svensson et al., 1997, 1999; Svensson & Wilson, 1999, Magnusson, 2002). The respective constructs have been studied both nationally in Sweden and internationally by a large number of researchers. For example, Rehman (1995) state that workload, situation awareness, performance (and vigilance[7]) are the main human performance measurement concepts that have received widespread attention.

The author's view is that a very important motivation for using these constructs in the modeling process is that the pilots talk about them among themselves. Even though, for example Hollnagel (1998) describes mental workload as an example of a construct that derives from a "folk model in psychology", the present author rather sees this allusion as a validity check. While many researchers see mental workload as an observable and measureable phenomena, Hollnagel's position is, that it is not an actual stage of, for example, pilots' information processing that would fit in a structural model of their cognition. This position could be taken for all constructs used, even for something as manifest as performance – in this thesis measured by "objective" variables – often quickly become problematic when analyzed further.

A point is that, for further reading and evaluation of the proposed model, it is important to remember that all the constructs used in the modeling process summarize and abstract a number of performance critical behavioral processes. It is also important to note that the constructs not are orthogonal to each other, and depending on perspective and the level of decomposition, they share content and meaning with each other. However, one of the goals of the thesis is to show that the constructs can be distinguished from each other with practically useful operationalization.

Vidulich (2003, p. 118) presses the point, that perhaps more important than the exact definition for, e.g. situation awareness or mental workload, is to which category of theoretical concept the construct is assigned. Vidulich uses Lachman, Lachman and Butterfields (1979) categorization with "black box theories" versus "structural theories". In this categorization a black box theoretical concept relates inputs to the cognitive system to its outputs, and provides a name for observed or hypothesized relationships between inputs and outputs, but it does not attempt to explain any causal mechanisms that account for the relationship. Correspondingly, the structural theory or model describes, at least at a logical level, the "inner workings" of the box whose output it seeks to explain.

[7] Given the short duration of the missions of the study where data comes from vigilance is not relevant here.

The present author adopts the view of Gopher and Donchin (1986) who take the following position:

> *"We imply that we are describing some entity or some property of entities that is not given entirely by the relationship by our empirical observations. At the same time, we assume that this excess meaning can be captured, studied, and measured in ways that would advance our understanding of the system and make it possible to use the concept for practical activities."*

1.7.1 Earlier models & results

In earlier Swedish and Swedish-US studies (e.g. Svensson, et al., 1993; Svensson & Angelborg-Thanderz, 1995; Svensson et al., 1997, 1999) a similar approach to human performance measurement and modeling as the one used in this thesis has been used. This earlier work emanated from a need of cost-effective training of pilots of the Swedish Air Force.

The purpose of the study reported in Svensson, Angelborg-Thanderz & Wilson (1999), was to analyze the effects of mission complexity and information load on mental workload, situation awareness and operative performance. In the first phase of this study 20 fighter pilots performed 140 missions in the air. In the second phase, 15 pilots performed 40 missions in a flight simulator. The pilots answered questionnaires tapping mission complexity, mental workload, mental capacity, situation awareness, and operative performance.

The resulting model, see Figure 5, has its starting point in the difficulty of the missions, as rated by the pilots before the mission. Increasing mission difficulty is followed by an increased general mental workload and an increase in the perceived complexity of the information on the Tactical Indicator (called TSD, Tactical Situation Display, in the figure) and the Target Indicator. The model tells us that there is a strong connection between the information load on the displays and "mental overloading" with "mental tunnel vision" as a consequence. The model show that increases in general workload and information complexity on both the Tactical Situation Display and Target Indicator decrease the situation awareness of the pilots. It was also found that there is a strong relationship between situation awareness and pilot performance. About 40 percent of the variance in performance can be explained by the variance in situation awareness.

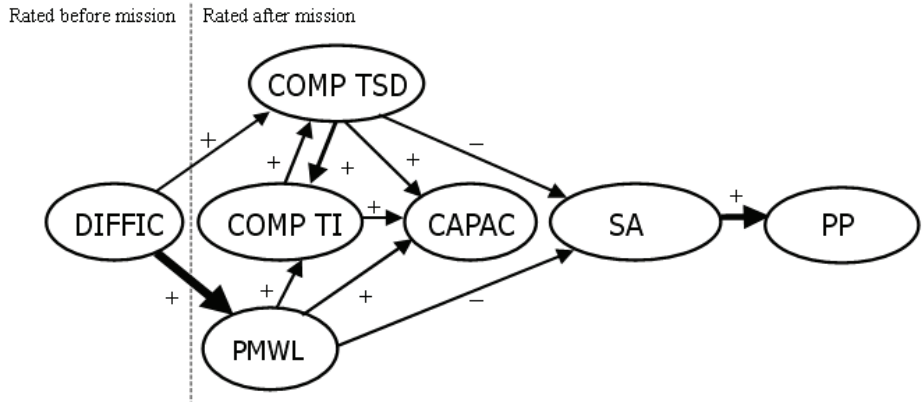


Figure 5. The final structural model presented in Svensson et al., 1999. + indicates positive effects, and – negative effects. Thick arrows indicate strong effects. All effects are significant ($p < 0.05$).

The LISREL-model in Figure 5 conceptually shows the relationships between the constructs of difficulty (DIFFIC), complexity tactical situation display (COMP TSD), complexity target indicator (COMP TI), mental reserve capacity (CAPAC), pilot mental workload (PMWL), situation awareness (SA), and pilot performance (PP). The manifest variables are not shown in the model.

As described in Nählinder, Berggren & Svensson (2004), there is an overall pattern reoccurring when using LISREL modeling on subjective ratings regarding mental workload, situation awareness, and performance. This pattern has shown to be stable over several studies and participants, even though both real and simulated flight, in military and civil settings has been included. The models that have been developed have shown a causal and logical relation connecting mental workload to situation awareness, and awareness, in its turn, to performance when a participant rates himself/herself. An increase in workload and a more demanding task leads to a decrease in situation awareness, which in turn leads to lower performance. Hence, a positive and strong relationship between situation awareness and performance, reported above, was found. The reoccurring pattern also implies that the subjective ratings are truthfully given and that the participants interpret the concepts and rating scales in similar ways, even though they are used at different occasions, with different questionnaires and format, and with different participants.

In the thesis this basic pattern and the same constructs used in Svensson et al. (1999) was the point of origin for the model conceptualization stage. However, both the measurement model and the structural model were modified and extended with performance being measured by "objective" variables. Teamwork was included, and while difficulty was assessed it was not included in the model. The mental reserve capacity and mental workload was combined into one construct. Given that the SBA study, where the data was collected, concerned the effectiveness of a new aircraft radar, the construct of sensor

effectiveness was also included. In all, the following 6 constructs were conceptualized:

- Sensor effectiveness
- Usability of information[8]
- Mental Workload
- Situation Awareness
- Teamwork
- Operative Performance[9]

All of the main concepts have in themselves received extensive attention in the research literature. They have all been identified as multi-dimensional concepts, and for each construct a wide variety of measures have been suggested and evaluated. Each construct can be, and have been, dissected further by other researchers. The constructs can also be represented by other measures than those chosen here. The current thesis focuses what to do after data has been collected and attempts to describe how the constructs relates to each other in something as multi-dimensional as "human activity".

1.7.2 Sensor effectiveness

A pilot in a fighter aircraft needs to actively work with the sensors at his disposal in order to ensure that all relevant information gets to the tactical displays. One primary sensor is the aircraft radar, which has a number of modes and search patterns that the pilot use depending on situation, standard operating procedures (SOPs) and his prediction of where the enemy will appear.

This construct is not a psychological construct, but as the data was collected within a SBA study evaluating the effectiveness of different radar alternatives, the construct was included. The sensor effectiveness, in conjunction with datalink information from team members, also defines the "outer technical envelope" of where the pilot can be aware of, and act against enemies. Also, in order to effectively launch missiles on the enemy, sensor effectiveness needs to be satisfactory, e.g. not lose radar track too easily.

1.7.3 Usability of information

The signals from the aircraft sensors and the information coming through the tactical datalinks are presented on three main displays in the JAS 39 aircraft. This information must be integrated and interpreted correctly by the pilot and this can be quite demanding when there is much information on the screens. The symbols are complex, enemy aircraft and their status is, for example, displayed differently depending on the source of the data. Extra information is also added to target symbols if a target is selected as a priority target. The construct used in Svensson et al. (1999) was labeled information complexity and relates to whether it was possible to read and use the information that was presented on the displays. The present author chooses "the opposite" of information complexity and

[8] In Svensson et al. (1999) this construct was labeled Information Complexity, thus representing an "opposite" concept. Information complexity also consisted of two factors, Information Complexity on the Target Indicator and Information complexity on the Tactical Indicator, but was here collapsed into one construct.

[9] In the model(s) the performance construct have been divided into offensive and defensive performance.

labels it usability of information. The construct includes components of both amount of information (i.e. number of objects on the displays) and complexity (i.e. symbols being built up by many subcomponents).

1.7.4 Mental workload

The construct of workload, with or without mental in front, has received a lot of attention in aviation, ever since it was linked with aircraft performance and safety issues. Workload has often been used as a critical criterion in validation and certification, and workload measurement has been considered a useful tool to aid the evaluation of new cockpit designs or to predict crew performance. Already in Hartman & McKenzie (1979) reference is made to workload as being a concept that has received attention for quite some time. References to the term cognitive workload (e.g. Just, Carpenter & Miyake, 2003; Patton, 2007), and that relate to a very similar construct, have appeared in the scientific literature.

Despite being around for at least 40 years, it is still not "solved" and to date there is consensus that no unified theory of workload exists (Castor et al., 2003).

According to O'Donnell & Eggemeier (1986, p. 42) *"The term workload refers to that portion of the operator's limited capacity actually required to perform a particular task".* Gopher & Donchin (1986, p. 41) defines mental workload as *"...the difference between the capacities of the information processing system that are required for task performance to satisfy performance expectations and the capacity available at any given time".* Kantowitz (1986, 2000) defines mental workload as an intervening variable that modulates allocation of the human's information-processing resources to task demands. Hart & Staveland (1988) describe workload as *"...the perceived relationship between the amount of mental processing capability or resources, and the amount required by the task"*. They also state that mental workload emerges from the interaction between the requirements of a task, the circumstances under which it is performed, and the skills, behaviors, and perceptions of the operator.

Even though many definitions have been proposed, there seems to be general agreement that mental workload is not a unitary, but a multi-dimensional concept that taps both the difficulty of a task and the effort (both physical and mental) brought to bear. Inherent in the notion of mental workload has been the concept that the human operator has a limited capacity to process information. Miller & Hart (1984) identified nine dimensions worth examining in detail when studying total workload: task difficulty, time pressure, own performance, mental effort, physical effort, frustration, stress, fatigue, and activity type. Each of these dimensions affects the information processing capacity of the human operator.

The Human Factors community (e.g. Manning, 2001) often makes a distinction between taskload (the task imposed upon an operator) and their workload (the operator's subjective response). In Castor et al. (2003) the concept was even further dissected, see Figure 6.

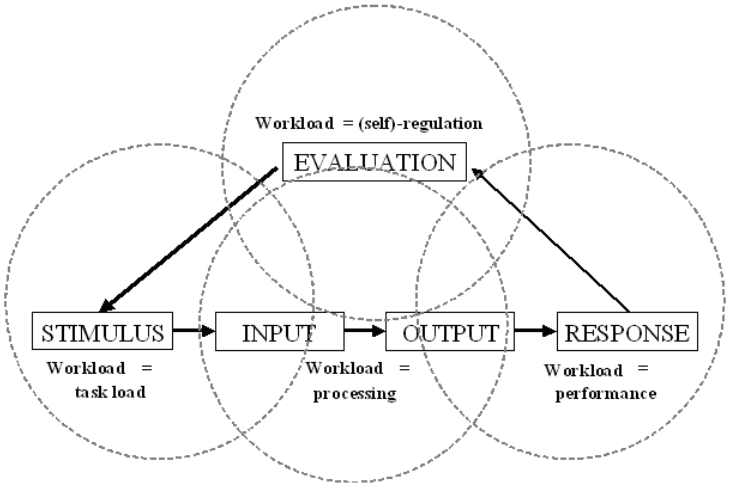


Figure 6. A simplified model of the human operator environment, illustrating the importance of a multi-dimensional approach to workload measurements.

In Figure 6 the importance of a multi-dimensional approach to workload measurements becomes apparent. Four main components of workload (the circles) are identified: task load, processing, performance, and (self)-regulation. These components are difficult to measure and therefore, the measurable "boxes" of these components are suggested: a) STIMULUS, b) INPUT, c) OUTPUT, d) RESPONSE, and e) EVALUATION:

a) STIMULUS. This box represents the external stimuli that can be perceived by an operator. Examples of stimuli are the physical work environment (i.e. including the displays and controls), communication (with crewmembers and air traffic control), mission requirements or the aircraft operation at hand.

b) INPUT. This box refers to the sensory perception of relevant external stimuli. The eyes, ears, and other sensory organs have unique characteristics and limitations that influence the quality and quantity of information flow.

c) OUTPUT. Perceived information can be filtered or processed, affecting pilot situation awareness, decision-making, anticipation, planning, et cetera. These processes may occur without immediate behavioral changes or responses, and is therefore separated from the response box. Mental and physical state changes (e.g. anxiety or fatigue) related to motivation or adopted strategy fall within this box.

d) RESPONSE. Overt behavior, including manual inputs or verbal commands, is the main behavior observed in this box. Measurements of Performance (MoP) adequately indicate its contents.
e) EVALUATION. The information flow through the previous boxes and their effects are evaluated in this box. The outcome of the evaluation may or may not result in a behavioral response. The evaluation process may be long term (i.e. after several stimulus-input-output-response cycles). Operator training and experience are the major factors that affect this box. Measurements of Effectiveness (MoE) relate to the content of this box.

Workload measures can be linked to each of these boxes via the method assessment matrices described in Castor et al. (2003).

1.7.5 Situation awareness

One concept that has received a lot of scientific and operational attention during the last two decades is the concept of situation awareness or SA (Endsley, 1988, 1993a, 1993b, 1995a, 1995b, 2000; Fracker, 1988; McMillan, Bushman & Judge, 1996). Today situation awareness is often used as a summary concept in many domains such as aviation, military command staff work, air traffic control etc. In the thesis by Alfredson (2007) a recent and thorough review of the construct of SA and its measurement is provided.

The definition adopted by the operational community of USAF according to McMillan et al. (1996) is *"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."*

The most frequently cited definition of SA is Endsleys (1995a): *"Situation awareness is 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."*

Situation awareness is a multi-dimensional construct as much as mental workload. Endsley (1995b) divides the concept of SA into the components of perception, comprehension and projection. These components represent three hierarchic levels as described below:

Perception involves monitoring, cue detection and simple recognition; it produces Level 1 SA, the most basic level of SA, which is an awareness of multiple situational elements (objects, events, people, systems, environmental factors) and their current states (locations, conditions, modes, actions).

Comprehension involves pattern recognition, interpretation and evaluation; it produces Level 2 SA, an understanding of the overall meaning of the perceived elements, how they fit together as a whole, what kind of situation it is, what it means in terms of one's mission goals.

Projection involves anticipation and mental simulation; it produces Level 3 SA, an awareness of the likely evolution of the situation, its possible/probable future states and events. This is the highest level of SA.

The multi-dimensionality becomes evident when Endsley (1995b) puts SA in relation to other factors as shown in Figure 7.

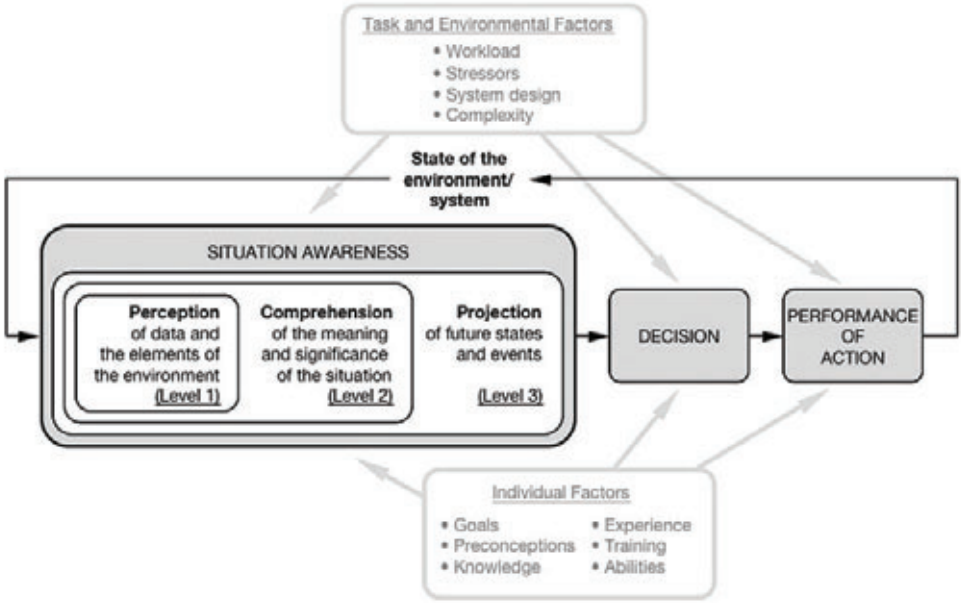


Figure 7. Situation awareness in context, from Endsley (1995b). Figure from Wikipedia article on situation awareness.

As evident from Figure 7, the full knowledge and experience of a person is not part of the SA construct. Thus SA only covers/affects some aspects of a person's decision making and acting in a dynamic situation and the construct is separate from decision making and performance (Endsley, 1995a). As SA is separate from other aspects of the performance of an operator, even a well trained decision-maker can make the wrong decision due to low or unsatisfactory SA. Likewise, an operator with a high and satisfactory SA can make the wrong decision.

The construct of SA has received extensive attention and its definition and concerns of its validity have raised quite a fair amount of debate (e.g. McGuinness, 1996; Artman, 1999). Decompositions, where situation assessment as a process, leads to situation awareness as a state, have also been proposed (e.g. Fracker, 1988). Alfredson (2007) includes action and suggests situation management as a more useful concept.

The concept of shared (or team) situation awareness and suggestions for measurement of it (e.g. Cannon-Bowers, Salas & Converse, 1993; Nofi, 2000) also exists in the scientific literature. Shared situational awareness obviously differs from individual situation awareness because it involves a number of persons that are trying to form a common picture of what is happening. In any given situation each participant are developing their own awareness of the situation, which exists within each persons cognitive processes or mind. Researchers interested in shared situation awareness study how each individual can share that understanding with the other members in a team or group.

To, in excruciating detail, exemplify that SA is intended to be used as a summary concept, Endsley's (1993b) attempt to pin down situation awareness for air-to-air combat is included here. Endsley's SA taxonomy for air-to-air combat is presented in Table 1, and it is intended to reflect "everything" a fighter pilot needs to know in order to have good SA.

Table 1. Endsleys (1993b) taxonomy of SA components for air-to-air combat.

	Ownship	**Flight**	**Friendly**	**Threat**
ID (friendly, flight, threat)				
Aircraft type				
Aircraft flight envelope (vel, acc, capabilities)				
Pilot experience				
Location				
Range				
Azimuth				
Altitude				
Attitude				
Heading				
Aspect				
Acceleration				
Time until intercept other aircraft				
Opening/closing velocity				
G's				
Center of gravity				
Vertical velocity				
Thrust level				

	Ownship	Flight	Friendly	Threat
Time until ground impact				
Current fuel level				
Bingo fuel level				
Fuel flow rate				
Time and distance possible on fuel level				
Sensor search mode				
Sensor search volume				
Sensor limitations (weather, clutter, etc)				
Detections				
Lock-ons				
Sort assignments				
Weapon type selected				
Weapon position selected				
Weapon quantity				
Weapon launch envelope				
Time until other aircraft in weapons range				
Ability to launch on other aircraft				
Current target				
Current missile probability of kill				
Time until weapons employment				
Anticipated firing position				
Targeted by whom				
Emissions (IR, radar, comm., contrails)				
EWS mode				
Com/NAV frequencies				
IFF code				
IFF reply				
Number and type of expendables				
Jamming effects				

	Ownship	Flight	Friendly	Threat
Jammed by whom				
Currently jamming whom				
System status				
Impact of system degrade/malfunctioning				
Current maneuver				
Future maneuver				
Current activity (engaged, free, …)				
Future activities, intentions, objectives				
Energy state				
Tactical posture (offensive/defensive)				
Advantage/disadvantage against others				
Priority threat				
Priority threat imminence				
Mission timing status				
Survivability				
Confidence level of information				
Terrain				
Forward Edge of Battle (for ground forces) location				
Objective location				
Home location				
Waypoint locations				
Landmark locations				
AWACS location				
Tanker location				
Safe areas				
Troops etc				
Ground obstacle location/height				
GCI net detections				
Ground threat type				

	Ownship	Flight	Friendly	Threat
Ground threat ID				
Ground threat active/not active				
Ground threat location				
Ground threat detection volume				
Ground threat weapons volume				
Missile location				
Missile time to impact				
Missile active/ lock-on				
Missile target				
Missile origin				
Kill assessment				
Sun				
Clouds				
Weather				
Prevailing visibility				

To conclude, SA as a summary concept is intended to capture a number of basic questions a pilot or operator has such as: "Where am I?", "What's my status?", "Where are the enemies?", and "What will happen now?".

1.7.6 Teamwork

The construct of teamwork and its effect on performance have also received widespread attention, e.g. Canon-Bowers, et al. (1993) or Salas, Stout & Folwkes (1997). For the tactical tasks of interest in this thesis, the effectiveness of the fourship of pilots to a large extent depends upon how well the individual pilots coordinate their activities. Salas, Dickinson, Converse & Tannenbaum (1992, p. 4) define a team as *"…any distinguishable set of two or more people who interact, dynamically, independently, and adaptively toward a common and valued goal/objective/mission, who have each been assigned specific roles or functions to perform, and who have a limited life-span of membership."*

Glickman et al. (1987) indicate that team performance includes two distinct dimensions of behavior, and separate between taskwork, i.e. the behaviors that are required for the execution of individual task, and teamwork, which are the behaviors needed for cooperative functioning. Brannick, Roach & Salas (1993) state that the understanding of individual competencies is necessary, but not sufficient, in order to understand team behavior. Good teams work together as a unit, not acting only as individuals. For the

missions where the data for this thesis was collected, this is definitely true. A pilot acting alone can only highly hypothetically achieve the team goal.

Similarly to the other constructs used in the thesis, this construct is multi-dimensional and many factors as personal attitudes and abilities build good teamwork. Burke, Salas, Wilson-Donnelly & Priest (2004) describe the "big five" of teamwork as being: team leadership, mutual performance monitoring, backup behavior, adaptability/flexibility, and team/collective orientation. Nofi (2000) describes several factors that are adverse to good shared situation awareness and teamwork: a false group mindset, the adoption of a "press on regardless" philosophy, poor personal communication skills, personnel turbulence, and perception conflicts among the team members.

Cannon-Bowers et al. (1993) state that effective coordination at the team level depends on the emergence of a shared mental model[10], or common understanding among team members regarding expected collective behavior patterns. Well-developed mental models help individuals to process and classify information more efficiently and form more accurate expectations and predictions of task and system events. Cannon-Bowers et al. suggest three kinds of knowledge structures in teams:

The first is a mental model that contains knowledge about the purpose of the teams, and more specifically the task requirements related to this purpose, called a **task model**. This model includes task procedures, strategies, and information on how the task changes in response to changes in the environment. The second model represents knowledge about unit characteristics, including their task knowledge, abilities, skills, preferences, and tendencies, called a **team model**. The third model, and the one that perhaps is most significant when studying collective action, contain information regarding the individual and collective requirements for successful interaction with team members. Cannon-Bowers et al. describe that to be effective, team members must understand their role in the task, which their particular contribution is, how they must interact with the other team-members, who require a particular type of information, and so forth. Related to this, they must also know when to monitor their team members' behavior, and when to step in and help a fellow team member who is overloaded, and when to change behavior in response to the needs of the team. When shared among team members, this model, called the **team interaction model**, is particularly crucial to effective coordinated action.

[10] Canon-Bowers et al. (1993) use the word model in a way that not should be confused with the statistical models presented in this thesis.

1.7.7 Operative performance

As elaborated in Bennett, Lance, & Woehr (2006) the criterion problem continues to be a challenge in the definition of performance. The definition of operative performance is typically application or case specific, and the goals of the mission or task decide appropriate performance measures. According to Locke and Latham (1990, p. 24), goals are "*...desired outcomes in terms of levels of performance to be attained on a task."* The goals of the scenarios of the present study were to shoot down as many as possible of the enemy aircraft (fighter and/or attack aircraft depending on the scenario), preferably without any own losses, and without unnecessary waste of resources, (i.e. returning to base with some missiles remaining is better than returning without any, given the same number of kills).

Care must be taken when selecting an appropriate measure and in the process of defining performance measures for the very complex behavior of skilled operators, there are several obstacles to overcome. There are skills, which are tacit, hidden, and embedded. The operator's handling of the system may have a wide range of effects from immediate to gradual or remote and from trivial to critical. The outcome may be manifest by very simple actions or no action at all. There are at least four aspects to consider: hidden knowledge and embedded performance, lack of practicable theories of performance, lack of studies on measures of validity, and lack of operational performance criteria (Vreuls & Obermayer, 1985; Angelborg-Thanderz, 1990; Svensson et al., 1999).

1.8 Measurement of psychological constructs

Given the complexity of human cognition it has been problematic to develop measures with the same properties as the measures of the natural sciences. The measurement of psychological constructs typically contains a number of trade-off judgments and has received intense attention during the whole history of psychology.

1.8.1 Requirements of measures

In the GARTEUR Handbook of Mental Workload Measurement (Castor et al., 2003) a number of requirements of measures are presented (based on Lysaght et al., 1989). Similar descriptions of requirements of measures are for example provided in ANSI/AIAA (1992) and Rehman (1995).

Validity

In crude terms, validity refers to the extent a variable measures what it is presumed to measure. Content validity refers to the degree a measure assesses appropriate, domain-specific knowledge or behavior. It gives (often multiple) meanings to a variable. At least three different aspects of content validity are important: Factorial or construct validity is based upon factor analysis. From theoretical reasoning and empirical research it is reasonable to conclude, for example that mental workload, pilot performance, as well as SA are multifaceted concepts or constructs. The validity of a manifest measure of one of these constructs or factors is indicated by its correlation with the factor, which is its factor loading. The correlation indicates to what extent the specific measure represents the construct. Both predictive and concurrent validity are expressed by the correlation between a criterion variable and a specific measure (criterion validity). Face validity is related to acceptance of a variable and is of special importance when measuring subjective experience.

Reliability

According to the reliability theory, reliability can be defined as the proportion of the total variance of a measure that is true variance. An obtained measure or score is assumed to be the sum of a true measure and an error component. Test-retest reliability (stability) refers to the capability of a measure to provide the same results when the exact conditions are replicated on two or more separate occasions. Internal consistency refers to the extent different measures are similar with respect to factorial content.

Sensitivity

The sensitivity of a measure is closely related to its reliability (relationship between true and total variance). It indicates a measure's capability to distinguish between different conditions of interest imposed on an operator or pilot, e.g. different levels of mental workload. The sensitivity of a mental workload measure would increase with the measurement technique's capacity to measure mental workload variations during a flight. Sensitivity is a very important criterion and critical in the selection of empirical measures.

Diagnosticity
Diagnosticity refers to the extent a measure expresses not only overall assessments, but also gives information about specific components of that assessment. According to Lysaght et al. (1989) the essence of diagnosticity is to be able to identify, for example, the specific mechanism (sensory, perceptual, cognitive, and psychomotor), the process involved during the performance of a particular task or which part of an interface an operator has problems with.

Data properties
Properties of data on different scale levels (nominal, ordinal, interval, ratio) and the resulting appropriateness of analysis by different parametric or non-parametric statistical methods are valid issues in every psychometrically oriented study. Behavioral researchers often use parametric methods on data from measurement of psychological phenomena, collected through Likert-type scales, despite the fact that they cannot be guaranteed to yield data on interval scales.

Sjöberg (2006) present some arguments in favor of analyzing Likert-type scales of data as if they were interval measurements and why puristic statistical criticism is right in principle, but wrong in practice. Three of Sjöbergs arguments are:

- It is being done and it yields meaningful results and a coherent body of knowledge. That would probably not happen if the data were vastly different from an interval scale.
- Comparisons between parametric and non-parametric methods tend to give the same results, but non-parametric methods usually have lower power.
- It is reasonable to assume that deviations from a true interval scale are modest. Only drastic and very large deviations would lead to erroneous conclusions.

1.8.2 Methods of human performance measurement

The measurement approaches used in human performance measurement can broadly be categorized into three major categories:

- Subjective ratings (during or after task/mission execution, by subjects or observers)
- Performance measures (primary task measures or secondary task measures)
- Psychophysiological measures

Subjective ratings
When subjective ratings are used, the subjects participating in a study quantify their experience or opinion with regard to a number of questions that tap the essential aspects the researcher wants to assess. Subjective ratings are most often given after the completion of a mission or a task, but can also be given during mission execution in order to provide quasidynamic data during the process (Svensson, Rencrantz, Lindoff, Berggren, & Norlander, 2006). The subjective ratings can be provided by both subjects and observers and then represent different perspectives on the same event. Subjective measures have turned out to be extremely useful in many contexts. It is sometimes claimed that the reliability and validity of subjective ratings of, for example, mental

workload and performance are insufficient (e.g. Muckler & Seven, 1992), and that it can be difficult to fully determine what has been measured. Doubts about their validity, see for example Stanton & Stammers (2002), although sometimes justified, should not be exaggerated. Even if the precision of any single rating is modest, data may still be sufficiently rich in information to be useful. The present author wants to highlight the experienced operator or subject as an intelligent filter against the complexity of the world who has the capability to integrate his or her experience into a balanced measure. Note that this capability is dependent upon the nature of the construct that is being assessed, as not all mental processes are introspectively available. For the constructs used in the thesis the present author adopts a similar standpoint as the one expressed by Johanssen, Moray, Pew, Rasmussen, Sanders & Wickens (1977) concerning workload: *"Despite all the well-known difficulties of the use of rating scales we feel that these must be regarded as central to any investigation. If the person feels loaded and effortful, he is loaded and effortful whatever the behavioral and performance measures may show*." The discussion easily gets stuck in whether a measure is objective or subjective, while it is more important to get a grasp of its reliability.

One major advantage of subjective ratings is the cost and ease of administration as they are often administered in paper-and-pencil form. Well designed rating questionnaires also usually have high face validity within an operational community and can be analyzed relatively quickly and straightforwardly.

Performance measures

One problem with primary task measures, as opposed to the more general class of secondary task measures, e.g. counting backward from thousand in steps of seven, is that a measure must be developed on an individual basis for each application. Traditionally, primary task measures in Human Factors have been variants of speed and accuracy measures. Speed measures would be based on the reaction time, for example, to perceive an event and initiate an action. In Angelborg-Thanderz (1990) deviations from prescribed standard procedures or routines were used as one of several accuracy measures. In Svensson et al., (1997) precise navigation at specified altitudes and flight speeds were the accuracy measures used.

Techniques for performance assessment can be categorized in a number of ways. One aspect of categorization concerns the level of objectivity, i.e. whether the assessments derive from technical measurement or from humans' subjective ratings, observations and statements. Even if this classification de facto is empty or insignificant, it is still in frequent use. The fundamental factors of all measures (independent of level of objectivity) concern their validity and reliability, i.e. whether they measure the right aspects, and whether the measures are precise.

Psychophysiological measures

Psychophysiological measures, as changes in, for example, heart rate or blink frequency, are reflecting psychological state changes and have been used by many researchers (e.g. Wilson, 1993; Wilson, Fullenkamp & Davis, 1994); Veltman & Gaillard, 1998) to describe changes in the functional state of an operator.

Several reviews of mental workload measures (Lysaght et al., 1989; ANSI/AIAA, 1992; Castor et al., 2003) describe a variety of different psychophysiological measures that can be used to assess variations in mental workload. It is in workload assessment that the psychophysiological measures primarily have been used, but they have also been used to assess situation awareness. These various measures can be classified into three major categories, or related to different physiological subsystems: 1) eye related measures, 2) brain related measures, and 3) heart related measures. Other measures, such as skin conductivity and muscle activity, have also been used.

In a recent thesis (Nählinder, 2009) psychophysiological measurement was used in order to do fine-grained comparisons between simulated and real flight. Especially eye-related measures have been used in SA research to provide information on where an operator is directing his or her visual attention to aspects critical for SA. A comparison of different eyetracking techniques is found in Alfredson, Nählinder & Castor (2004). The thesis by Alfredson (2007) summarizes several studies where eyetracking is used to investigate SA of pilots in military and civil flight simulators.

1.8.3 Choosing appropriate measures

In Harris et al. (1992) review of workload measurement techniques the following criteria are used: Operator availability, System/Hardware availability, Subject Matter Expert (SME) availability, Comparable system, Time constraints (preparation, data collection, scoring, analysis), Ease of use (preparation, data collection, scoring, analysis), Available task data (task descriptions, general task times, detailed task times), Workload dimensions, Operator contact, Real-time applications, Environment, Training times, Operator interference, Operator intrusiveness, Desired outputs, Diagnosticity, Anonymous results, and, Sensitivity. Although Harris et al. focus on workload measurement techniques, the same criteria are applicable to most choices of human performance measurement techniques.

In Alfredson, et al. (2003, 2004) a meta-method for selection of appropriate Human Factors evaluation methods is presented. The selection method is based on the answer to four general questions concerning the status of a specific project, that non Human Factors experts can answer. The four questions are: a) “Are you early or late in the development process, i.e. is there a prototype of the system available?”, b) “What resources are available for the Human Factors study?”, c) “What is the capability/experience of the Human Factors staff?”, and d) “Are end-users available?”. This meta-method divides the selection space, and narrows down the choice of appropriate evaluation methods between the 26 methods described.

The choice of method chosen for this study follow the logic of the meta-method, i.e. there were prototypes of the systems and a simulator setting in which to perform the studies, the amount of resources were not that extensive (which, for example, removed psychophysiological methods as an alternative), the availability/capability of the Human Factors staff was rather large but the present author was alone, and there was an adequate number of end-users available. Thus, the FOI Pilot Performance Scale (FOI PPS) was found to be a suitable choice of evaluation method to assess the pilots’ cognitive status,

complemented by data extracted from the simulator log. The FOI PPS and the data from the simulator are further described in chapter 2.

1.8.4 Sources of variance

Variation is the foundation for statistical analyses. Without variation of different phenomena no conclusions about their relationships can be drawn. If two or more phenomena vary (i.e. are variables), and they vary in systematic and concordant ways, there are covariances or intercorrelations between the variables. The "first generation" statistics concerns comparisons and tests of changes and differences between variables. The "second generation" statistics, such as multivariate statistical techniques including factor analysis and structural equation modeling, are based on the covariation of variables.

In classical experimentation the total variance of a database consists of *inter-individual variance*, *situational variance*, and *error variance*. Differences between cases or individuals are a main source of variance. Typically, groups with a large number of cases are measured after different forms of treatment. For instance, the performance of groups of subjects can be measured and compared after different kinds of training, or their performance in system with different designs can be compared.

In repeated measurement designs, *intra-individual variance* is added to the inter-individual and situational variances. It is reasonable to conclude that intra-individual variance, to a large extent, reflects experimental or external influences. This is due to the reduced proportion of inter-individual variance in repeated measures designs. Repeated measurement designs make possible descriptions of changes over time or over consecutive phases of a task. Accordingly, from the consecutive changes correlations as a function of the dynamics of the situation can be estimated.

In spite of the fact that the variance is the basis for statistical analyses and conclusions, the sources of variance are, to the author's knowledge, seldom discussed. In the studies performed by the Swedish Defence Research Agency (FOI) on the performance of skilled operators, in real as well as in simulated settings, a central conclusion is that the intra-individual variance as a function of situational dynamics is of specific importance. In operational studies we are usually more interested in the actions of the operators as a function of the situational changes rather than the differences between the participating subjects. The total variance of the databases presented here includes situational variance and inter-individual variance as well as intra-individual variance.

1.8.5 Structure theory

The radex theory (Guttman, 1954) describes how interesting phenomena sometimes can be described in simplex and circumplex structures. Simplex structures describe sequences where each entity or construct is affected by one preceding and affects one succeeding entity. In circumplex structures the entities or constructs are ordered in a circle. A simplex model is a type of covariance structure, which for example, can occur in longitudinal studies. The correlations between performance measures of pupils over the

school years can, for example, be represented by a simplex structure – the performance during a preceding year form base for the performance during the next.

A typical feature of a simplex correlation structure is that the correlations decrease as one move away from the main diagonal – meaning that the correlations decrease as a function of the size of lags between the entities or constructs. A distinction can be made between perfect and quasi-simplex structures. In a perfect simplex the measurement errors in the test scores are negligible. A quasi-simplex structure, on the other hand, allows for sizeable errors of measurement (Jöreskog & Sörbom, 1984). Figure 8 illustrates a simplex structure in which entity A affects entity B and so on.

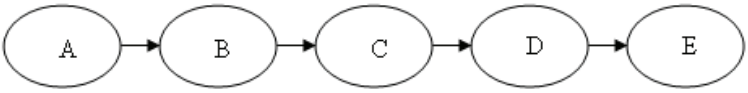


Figure 8. A simplex sequence in which A affects B, and B affects C, and so on.

As compared to simplex structures, circumplex configurations are seldom found to represent psychological processes, even though, for example, the perception of similarities between spectral colors can be represented by a circumplex structure (Ekman, 1954). In this structure the colors of e.g. red and green, and blue and yellow, respectively are located as contrary opposites on the circle. The structure of mood or emotions can also be expressed with circumplex representations. Svensson (1978) found that adjectives of moods or feelings are distributed on the surface of spheres, where similar feelings are close to each other (e.g. active–alert) and opposites (e.g. happy–sad) placed contrary to each other. He also found that the factor analytical model was suboptimal for an unambiguous reduction of adjectives evenly distributed or ordered in strings on the surface of the spheres. Figure 9 illustrates a circumplex structure in which the entities have a circular order.

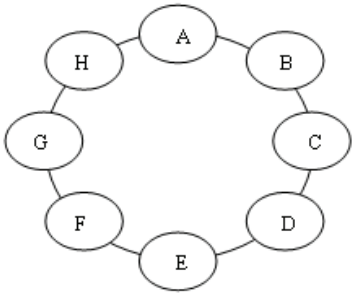


Figure 9. A circumplex structure in which A affects B, and B affects C and so on in a circular order.

1.9 Hypothesis

The formulation of the hypothesis below was driven by the following research question:

Can the complex reality and interactions between and within technical systems and individual operators be expressed in something as simplified as a simplex structure? In other words, is it possible to develop a simple statistical model where aspects/characteristics of a technical system, individual's cognition, the interaction between individuals and final operative performance can be described and quantified?

The hypothesis is best expressed graphically as in Figure 10:

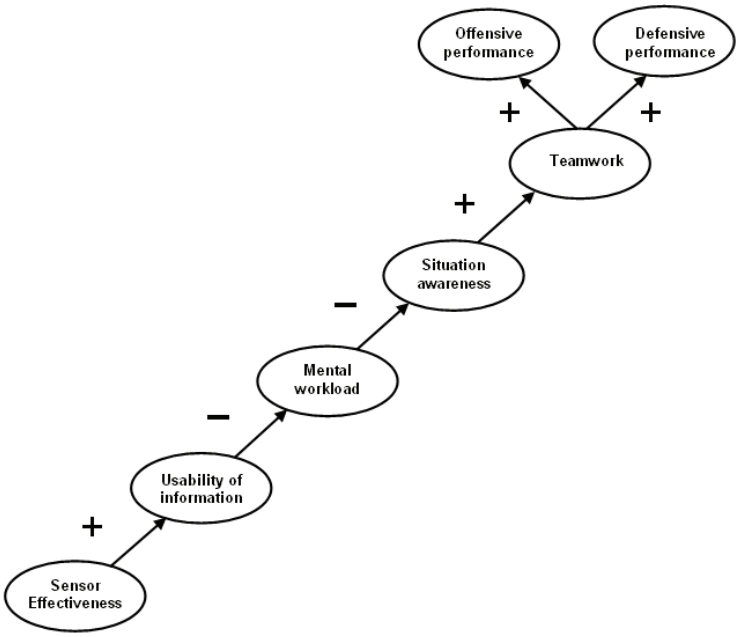


Figure 10. Conceptual model and main hypothesis of the thesis.

The figure should be interpreted as follows:
If sensor effectiveness is high → usability of information is high → mental workload is low → situation awareness is high → teamwork is high → performance is high.

The rationale behind the graphical representation of the hypothesis was the author's conceptualization that lower level processes (to the left in the figure) form base of higher and higher level of processes that eventually builds up to final performance.

However this main, or general, hypothesis can also be expressed as a number of sub-hypotheses.

Table 2. Sub-hypotheses.

H1: SENSOR EFFECTIVENESS is positively correlated to usability of INFORMATION.
H2: Usability of INFORMATION is negatively correlated to MENTAL WORKLOAD.
H3: MENTAL WORKLOAD is negatively correlated to SITUATION AWARENESS.
H4: SITUATION AWARENESS is positively correlated to TEAMWORK.
H5: TEAMWORK is positively correlated to OFFENSIVE PERFORMANCE.
H6: TEAMWORK is positively correlated to DEFENSIVE PERFORMANCE.
H7: All chosen constructs are useful and appropriate (valid and reliable) in a SEM model describing the collected data. H7 is a consequence of H1-H6.
H8: It is possible to describe the constructs in a simplex or quasi-simplex structure, while retaining reasonable model fit.

2 Method

2.1 Participants

A total of 37 fighter pilots from the Swedish Air Force participated in the radar study during nine weeks of simulation over the two years duration of the SBA study.

During each of the nine weeks the simulated aircraft were manned by eight fighter pilots, with four being the blue side versus four on the red side (except during one of the weeks, when two simultaneous 2 versus 2 engagements were flown).

The mean age of the pilots was 31 years and their experience as pilots, expressed in terms of flight hours, was from 600 to 2300 with a mean of 1340, and with a standard deviation of 533 hours. The amount of flight hours on military system aircraft, i.e. the JA37 or AJS37 Viggen system, and the JAS 39A[11] Gripen system was between 370 to 1750 hours with a mean of 888 hours, and a standard deviation of 457 hours. All the pilots were FFSU[12] qualified, i.e. were operational pilots and not students, i.e. they were not GFSU[13] or GTU[14] pilots.

Thus, the pilots participating in the study form a representative and experienced sample including some very senior pilots. The Air Force provided the pilots based on the timing of the study weeks in relation to other duties of the pilots, so there was no selection process, except that no novice pilots were accepted.

Table 3 shows the participation for each pilot, for each week. The numbers indicate how many engagements in which each pilot participated on the blue side during the week. As evident from the table the availability of pilots was such that they participated to various extents. Accordingly, some pilots have more influence over the database than others. Similarly the number of engagements flown over the weeks varied. This is one example of effects that are to be expected when the real operational pilots are used as subjects in a multi-year study.

[11] As of 2009 the JAS 39C version is the current single seat aircraft version.
[12] FFSU = Fortsatt flygslagsutbildning.
[13] GFSU = Grundläggande flygslagsutbildning.
[14] GTU = Grundläggande taktisk flygutbildning.

Table 3. Pilot participation in the simulations.

Week number / Pilot number	w 450	w 409	w 350	w 325	w 323	w 321	w 248	w 237	w 235	Total amount of participation on blue side
1	19	16	9	18			6		20	88
2		16				18	18	12	20	84
3				18	18		18	17		71
4		16	12	18	18					64
5		16	15	15	18					64
6			18	15				16	11	60
7			18				20	17		55
8		16					19	13		48
9					18	18				36
10								16	20	36
11		16							20	36
12	18					18				36
13	18					18				36
14			18			18				36
15				15					20	35
16	17			18						35
17							7	12	15	34
18							20	13		33
19				15	18					33
20			18		12					30
21							16		7	23
22									20	20
23			18							18
24			18							18
25						18				18
26						18				18
27						18				18
28							18			18
29	18									18
30	18									18
31	18									18
32	18									18
33					16					16
34					14					14
35					12					12
36							10			10
37									7	7
Number of engagements	36	24	36	33	36	36	38	29	40	308
Number of cases from week[15]	144	96	144	132	144	144	152	116	160	1232

[15] The number of cases from each week is the number of engagements times four.

2.2 Experimental design

The SBA study evolved over a two-year period and represents an experimentation campaign where the SBA questions and systems under study matured over time.

The first three weeks focused on tactics development and during the remaining six weeks radar characteristics were manipulated as primary independent variables. Missile performance, scenario and existence of fighter controllers also varied during the weeks according to Table 4.

Table 4. Description of the weeks.

Week No	Independent variable (& number of alternatives)	Blue vs Red	Fighter controller	Scenario	Radar	Missile
235	Tactics (3) & Radar (2)	4 v 4	No	Defensive	SECRET	SECRET
237	Tactics (3) & Radar (2)	4 v 4	Yes	Offensive	SECRET	SECRET
248	Tactics (3) & Radar (2)	2 v 2 + 2 v 2	Yes	Defensive & Offensive	SECRET	SECRET
321	Radar performance (3) & Missile performance (2)	4 v 4	Yes	Defensive & Offensive	SECRET	SECRET
323	Radar performance (3) & Missile performance	4 v 4	Yes	Defensive & Offensive	SECRET	SECRET
325	Radar performance (3) & Missile performance	4 v 4	Yes	Defensive & Offensive	SECRET	SECRET
350	Radar performance (3)	4 v 4	Yes	Defensive	SECRET	SECRET
409	Radar performance (2)	4 v 4	No	Defensive	SECRET	SECRET
450	Radar performance (3)	4 v 4	Yes	Defensive	SECRET	SECRET

As evident when combining the information in Table 3 and Table 4, i.e. participating pilots and study alternatives, the total database does not represent a classical experimental design. Variance in the data can be both due to the repeated measurement during different conditions and individual differences. Accordingly, the total variance is a composition of inter-individual, intra-individual, and situational variance.

For the current study concerning SEM modeling of the cognitive status of the pilots the whole dataset have been compiled into one database. This database contains data from a rather varied set of conditions, which strengthens claims concerning the generalizability of findings and proposed model(s).

2.3 Apparatus/instruments

2.3.1 The simulator facility

The simulator facility of the Swedish Air Force Combat Simulation Centre[16], FLSC, consists of eight pilot stations where the JAS 39 Gripen and other aircraft can be simulated. During the missions executed during the actual study four pilots were on the blue team and four on the red team. The facility also contains four FSL positions (fighter controller/fighter allocator) that were manned during seven of the nine weeks. The simulator facility has been developed in order to be used as a platform for training and materiel acquisition studies, where the training or study goals are on a technical to tactical level. The facility has been operational since 1998 and is used extensively by the Swedish Air Force.

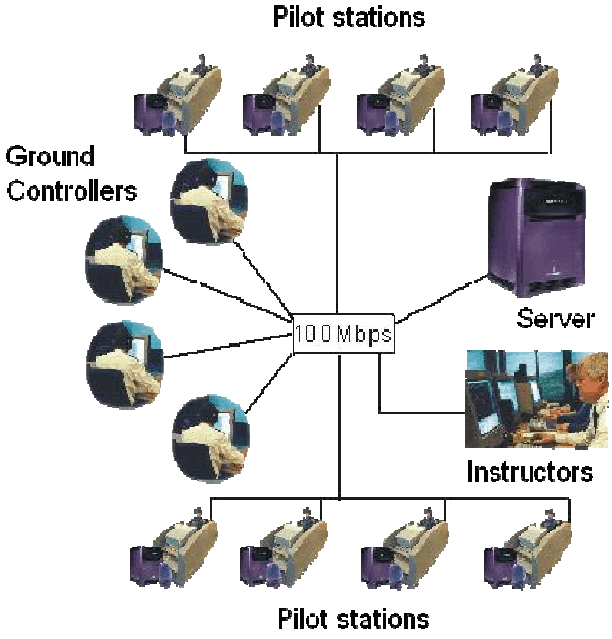


Figure 11. A schematic description of the FLSC simulator facility.

[16] FLSC, Flygvapnets luftstridssimuleringscenter

Figure 12. One of the pilot stations of FLSC with the layout from the time of the study[17]. Other pilot stations and the exercise manager's "God's Eye presentation" can be seen in the background. During the study covers were drawn between the pilot stations.

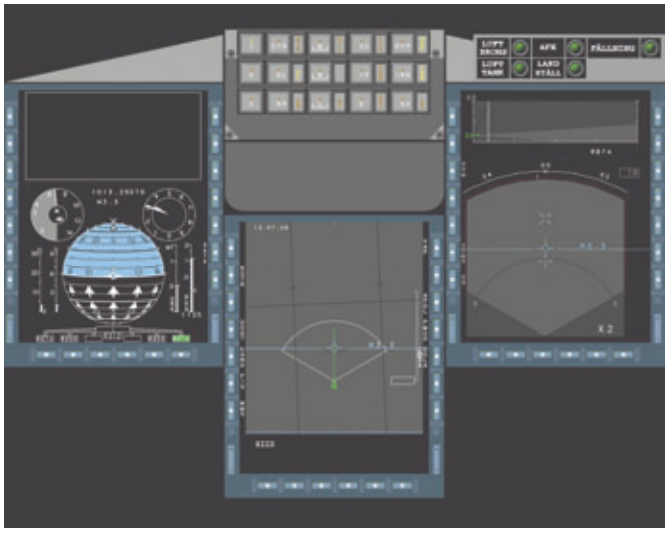


Figure 13. A head-down display from one of the pilot stations. The Tactical Indicator is the center display and the Target Indicator is on the right side.

[17] In 2006 the pilot stations were updated with dome displays, which provide much larger fields of view.

2.3.2 Data collection instruments

During the SBA radar study a large number of different measures were collected and analyzed. The data presented in the thesis are of two main categories, subjective ratings from the pilots on the FOI PPS (Pilot Performance Scale) and "objective" performance measures extracted from the simulator logs. A translated version of the FOI PPS questionnaire that was used is provided in Appendix 2.

The FOI PPS builds on previous efforts to develop practicable scales and subjective measures for operational use. In these studies central aspects have been identified by expert pilots and have been tested again and again in different studies. These rating scales initially started out with a large set of items that have been reduced in the search for a practicable instrument usable in studies outside of a lab setting. Some rewordings and updating of the questions were done, and questions concerning the system effectiveness and the teamwork constructs were added for this study.

- **Sensor effectiveness** was operationalized by three subjective questions concerning the performance of the radar and one technical performance measure that measured the amount of time during which the enemy was within a certain threat distance and the own aircraft had radar contact with the enemy aircraft.

- **Usability of information** was operationalized by four subjective questions concerning the usability of information on the tactical head-down displays. The questions refer both to the information on the Tactical Indicator, and the Target Indicator.

- **Mental workload** was operationalized by four subjective ratings. One rating on the well known Bedford Scale for mental workload assessments (Roscoe, 1987; Roscoe & Ellis, 1990), one question whether the level of mental workload was an obstacle to performance, and two questions relating to the pilots ability to meet the temporal demands of the situation.

- **Situation awareness** was operationalized by four subjective ratings. The questions chosen was one general question of overall SA, and three other questions were intended to measure some aspect of Endsley's (1995) three levels of SA: a) perception, i.e. were you surprised by enemy positions, b) understanding, i.e. were you surprised by enemy behavior and c) projection, i.e. could you predict the course of events.

- **Teamwork** was operationalized by four questions concerning the quality of the teamwork within the pilots of the fourship formation, including the ability to keep the original plan.

- **Performance** was split into two constructs, offensive performance and defensive performance. Each construct was operationalized by two performance measures extracted from the simulator logs.

The constructs and the manifest variables chosen for their operationalization are presented in Tables 5 and 6 below. Recommended practice (e.g. Bollen, 1989) in SEM is that each construct or factor should be measured by at least three manifest variables.

Table 5. Constructs and manifest variables.

Construct	**Full name of construct**	**Manifest variables**
SENSOR	Sensor effectiveness	1 performance measure, 3 subjective ratings
INFO	Usability of information, i.e. how useful was the information presented on the tactical head-down displays, i.e. the information on the Tactical Indicator and the Target Indicator.	4 subjective ratings
MWL	Mental workload	4 subjective ratings
SA	Situation awareness	4 subjective ratings
TEAM	Teamwork quality	4 subjective ratings
OFFPERF	Offensive performance, i.e. the results against the enemy	2 performance measures
DEFPERF	Defensive performance, i.e. survival	2 performance measures

Table 6. Description of manifest variables. Translated and sorted for thesis.

Manifest variable	**Construct**	**Description of questionnaire items (shortened) and objective variables. Full description available in Appendix 2.**
RADARCONTACT % of TIME → RADARCO7	SENSOR	The percent of time during which an enemy aircraft was within a certain threat distance and that the fighter had radar contact with it. The data was extracted from the simulator logs and was recoded into a seven step scale before modeling began.
SENSOR1	SENSOR	Could you use the radar's full potential?
SENSOR2	SENSOR	Could you keep radar contact with targets?
SENSOR3	SENSOR	Problems using the radar?
INFO1	INFO	Was it possible to survey the information on the target indicator?
INFO2	INFO	Could you use the information on the target indicator?

INFO3	INFO	Was it possible to survey the information on the tactical indicator?
INFO4	INFO	Could you use the information on the tactical indicator?
MWL1	MWL	Difficult to manage all actions? (reversed)
MWL2	MWL	Did you have “mental lead time” with respect to your tasks?
MWL3BED	MWL	What was your mental workload? Rated on the Bedford Rating Scale (Roscoe, 1987; Roscoe & Ellis, 1990).
MWL4	MWL	Was mental workload an obstacle to optimal performance?
SA1	SA	Situation awareness?
SA2	SA	Surprised by enemy positions?
SA3	SA	Surprised by enemy behavior?
SA4	SA	Could you predict the course of events?
TEAM1	TEAM	To what extent could you follow the initial plan of the team?
TEAM2	TEAM	How well did the teamwork in the unit work out?
TEAM3	TEAM	To what extent could you predict the behavior of the other pilots in your unit?
TEAM4	TEAM	How long did it take for the unit to “straighten out” situations when something became confused?

HITS → HITS7	OFFPERF	The number of hits on enemy aircraft by the pilot, both on enemy fighter aircraft (manned by the red side) and attack aircraft (computer generated, only during the defensive scenario). HITS data was extracted from the simulator and recoded into a seven step scale before modeling began. The HITS variable was recoded according to the following rule: 3 or 4 hits = 7, 2 hits = 5, 1 hit = 3 and 0 hits = 1 for the HITS7 variable.
TEAMHITS minus own → THITSMO7	OFFPERF	The number of hits by the whole fourship formation on enemy aircraft (enemy fighter and attack aircraft minus the value of the HITS variable. The data was extracted from the simulator logs and recoded into a seven step scale before modeling began. TEAMHITS minus own was recoded according to the following rule: 4 or more hits = 7, 3 hits = 5, 2 hits = 4, 1 hit = 3 and 0 hits = 1 for the THITSMO7 variable.
SURVIVAL → SURV7	DEFPERF	The pilot survived the mission. The data was extracted from the simulator logs. SURVIVAL is a dichotomous variable of 0 or 1 and was recoded into 2 = killed and 6 = survived for the SURV7 variable.
TEAMSURVIVAL minus own → TSURVMO7	DEFPERF	The number of surviving pilots in the team, apart from the rating pilot. The data was extracted from the simulator logs and recoded into a seven step scale before modeling began. TEAMSURVIVAL minus own was recoded into: 3 survivals = 7, 2 survivals = 5, 1 survival = 3 and 0 survival = 1 for the TSURVMO7 variable.

All subjective ratings, except the mental workload rating on the Bedford Rating Scale, were made on seven point Likert scales with anchors at the ends. In the Bedford scale the ratings are presented according to a decision tree structure to aid the selection of appropriate descriptors. See Appendix 2.

2.4 Scenarios

Two tactical scenarios were used in the study to broaden the space of sorties for which results from the radar study were valid. In one scenario the blue side was a fourship formation with a basically defensive mission goal. In the other scenario the blue side also flew a fourship formation, but with an offensive mission goal. In both scenarios the blue and the red side were evenly matched in terms of fighter aircraft, although system performance (aircraft, missile, radar) varied across study weeks.

Both scenarios were beyond visual range (BVR) scenarios, where it was appropriate and possible to engage the enemy before visual contact was made, i.e. the radar is the primary sensor that can provide the pilot with the information necessary for him to develop his situation awareness. The radar warning receiver is the only other onboard sensor that provided relevant information. Information from the other aircraft in the formation and the ground controllers is sent through the tactical data links, which greatly enhance the pilots' possibility to keep track of their own team-members and enemy aircraft.

In the defensive scenario the blue fourship formation was tasked to defend the airspace against an approaching red formation of computer generated attack aircraft, escorted by a manned fourship formation of fighter aircraft. The blue fighter had to destroy the attack aircraft or destroy their escorts and thereby forcing the attack to be aborted. The target of the red attack formation was a blue naval vessel and the attack had to be countered before the red side could reach their weapons release point. There were no other aircraft in the area, and thus a very "clean" scenario.

In the offensive scenario, roles were basically reversed and the blue fighters were tasked to escort a blue attack column, i.e. bombers, whose target was a number of red surface to air missile sites (i.e. an escort of a SEAD/DEAD[18] mission). In this scenario a number of neutral aircraft, both civilian airliners, helicopters and the fighters of a neighboring country provided a much more cluttered airspace.

[18] SEAD = Suppress Enemy Air Defenses. DEAD = Destroy Enemy Air Defenses.

2.5 Procedure

At the start of each week in the SBA radar study the pilots flying during the week were briefed for approximately two hours on the goal of the radar study, the performance of the different radar alternatives, the tactics to be used and experimental protocol. The pilots then had approximately two hours of training in the simulator with the new systems before the engagements of the study began.

In order to minimize the effects of tactics choice on the outcome, the engagements were flown in "triplets", where the same basic tactic and risk level was maintained for three consecutive engagements, one with each of radar alternatives of the current week. Formations were allowed to be mirrored or changed to a minor extent, but the same basic behavior was to be maintained within the triplet.

When the engagement was over, or when a pilot was shot down, each pilot on the blue side answered a number of questions on the FOI PPS questionnaire. The questionnaire typically was completed within five minutes. After questionnaire completion all pilots, both from the blue and red team, watched a replay of the engagement. After the replay there was time for a short tactics discussion, change of sides if a triplet was completed and then a new engagement started. Each engagement cycle typically was 40 minutes.

3 Results

3.1 Data collection

As presented above, the SBA radar study consisted of nine weeks of simulation. In total 308 engagements were flown were different radar system alternatives were evaluated against each other. From each engagement data from all 4 pilots on the blue side were collected, resulting in 1232 surveys being administered. Consequently the database consists of 1232 cases of data strings with 24 variables. For 9 cases there were some missing data, a result from when a pilot missed to score on a few of the items in the questionnaire. For these 9 cases, where there were missing data for typically 6 variables, rounded means have been imputed in the database. In the data there are 1232 distinct response patterns.

The initial model development, presented in 3.4 and 3.5, was performed with a dataset consisting of data from weeks 235, 237, 248, 321, 323, and 325; summing up to 848 cases in the database. Weeks 350, 409 and 450, i.e. 384 cases, were kept aside as a cross-validation sample.

The pilots and fighter controllers were motivated and disciplined during the study, both with regard to behavior in the simulator and in adherence to the experimental protocol. A few mission instances were lost due to simulator or logging problems, but were repeated with the same basic setup to preserve the number of missions flown with each radar alternative.

3.2 Normality of data

To examine the normality of the data histograms all variables have been inspected and in Figure 14 a typical example is provided.

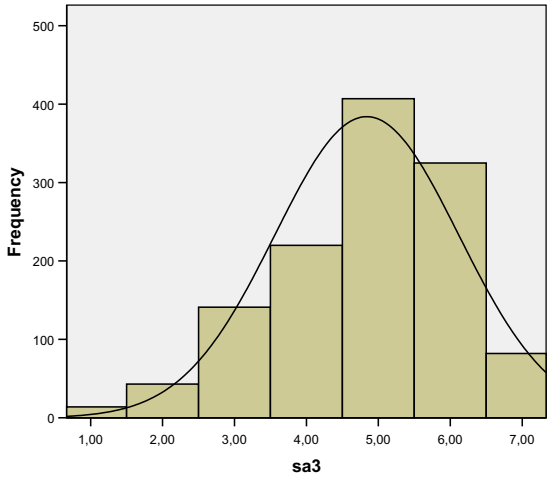


Figure 14. Histogram describing the distribution of the SA3 manifest variable.

In Table 7 the means and standard deviations along with skewness and kurtosis are reported for the manifest variables. The variables are ordered in groups representing the hypothesized constructs. Data for the five variables from the simulator log are restricted.

Table 7. Description of the data.

Construct	Question	Mean	Median	Std. Dev.	Min	Max	Skew	Kurtosis
Sensor Effectiveness	RADARCO7	SECRET			1	7		
	SENSOR1	4,60	5	1,40	1	7	-0,69	0,02
	SENSOR2	4,42	4	1,48	1	7	-0,39	-0,49
	SENSOR3	4,38	4	1,43	1	7	-0,43	-0,50
Information	INFO1	4,33	4	1,36	1	7	-0,55	-0,27
	INFO2	4,24	4	1,37	1	7	-0,50	-0,30
	INFO3	4,85	5	1,27	1	7	-0,43	-0,39
	INFO4	4,79	5	1,30	1	7	-0,30	-0,52
Mental Workload	MWL1	3,31	3	1,16	1	7	0,45	0,04
	MWL2	3,36	3	1,25	1	7	0,63	-0,09
	MWL3BED	3,13	3	1,02	1	10	1,08	3,38
	MWL4	2,86	3	1,07	1	7	0,71	0,48
Situation Awareness	SA1	4,65	5	1,20	1	7	-0,50	0,15
	SA2	4,77	5	1,35	1	7	-0,57	-0,07
	SA3	4,84	5	1,28	1	7	-0,56	0,02
	SA4	4,38	4	1,17	1	7	-0,35	-0,15
Teamwork	TEAM1	4,66	5	1,38	1	7	-0,62	-0,07
	TEAM2	4,60	5	1,27	1	7	-0,52	-0,20
	TEAM3	4,59	5	1,19	1	7	-0,55	0,05
	TEAM4	4,47	4	1,25	1	7	-0,42	0,21
Offensive performance	HITS7	SECRET			1	7		
	THITSMO7	SECRET			1	7		
Defensive performance	SURV7	SECRET			2	6		
	TSURVMO7	SECRET			1	7		

Studies by Jaccard & Wan (1996) indicate that maximum likelihood solutions are robust to skewness with only small effects on the estimation of parameters and standard errors.

Kurtosis distributions that depart from normality assumption represent another issue and can lead to increased risks of making the wrong conclusions. As a rule of thumb, data may be assumed to be normal if skewness and kurtosis is within the range of +/- 1.0 (Schumacker & Lomax, 2004). SURV7, HITS7 and MWL3BED are the only three variables that do not meet these criteria. As an illustration, the histogram for MWL3BED is presented in Figure 15. As can be seen from inspection of the figure, the pilots' responses are approximately normally distributed, even if the variable does not fulfill the skewness- and curtosis criteria.

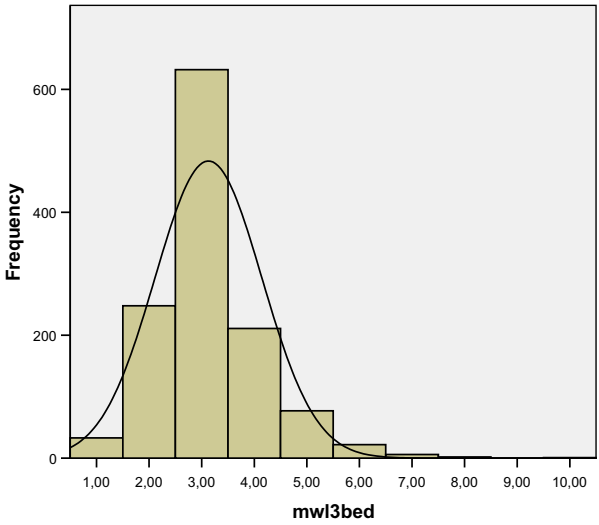


Figure 15. The distribution of the MWL3BED manifest variable.

The potential problems with non-normality are that we might not be able to trust our results because they rest on the normality assumption. However, Boomsma & Hoogland (2001, p. 14) indicate that normal maximum likelihood theory works well under "reasonable" non-normality, which was the case for the data used here. Furthermore, non-normal distributions tend to reduce optimal covariances between variables, which will counteract e.g. simple structures of factor analyses.

Different researchers have suggested recommendations concerning the sample size in relation to the number of manifest variables. The ratio of 1:15 is one of the more stringent recommendations, although this type recommendations should be seen in the light of the "clarity of structure" and strength of relations within the correlation or covariance matrix. For the current sample with n = 1232 the recommendations are easily met.

Structural Equation Modeling (SEM) also assumes linear relationships between manifest and latent variables, and between latent variables. In order to detect any violations of this assumption a large number of bivariate plots were analyzed, but no non-linear effects were detected. Note that in other corresponding studies, but with other mission profiles,

non-linear relations, e.g. underload effects between mental workload and performance, have been found (Svensson et al., 1999).

Multicollinearity, i.e. when there are very high correlations between two or more variables, can lead to problems when conducting multivariate analyses; standard errors of parameter estimates, and coefficient estimates can be affected. A correlation of more than 0.85 between variables represents high multicollinearity (Garson, 2008). The correlation matrix between all manifest variables was inspected and only the correlation between INFO4 and INFO5, i.e. the two questions assessing the usability of information of the tactical indicator reach a r = of 0.86.

3.3 Factor analysis and development of measurement model

Mulaik & Millsap (2000) suggeste a stringent four-step approach to modeling:

1. Common, i.e. exploratory factor, analysis to estimate the number of latent variables or factors.
2. Confirmatory factor analysis to confirm the measurement model. As a further refinement, factor loadings can be constrained to 0 for any measured variable's crossloadings on other latent variables, so every measured variable loads only on its latent (i.e. fulfilling Thurstone's (1932) simple structure criterion). Schumacker & Lomax (2004) note that this could be a tough constraint, leading to model rejection.
3. Test the structural model.
4. Test nested models to get the most parsimonious one. Alternatively, test other researchers' findings or theory by constraining parameters as they suggest should be the case.

In 3.3, the first two steps are presented, while step 3 and 4 are presented in 3.5.

A subset of the data was kept aside for cross-validation, and thus the analyses presented in 3.3 and 3.4 are based on a sample size of 848.

3.3.1 Exploratory factor analysis

A common exploratory factor analysis was performed by the statistical analysis software SPSS 14 (SPSS, 2008) and the output is presented below. Oblique rotation, as opposed to orthogonal, was chosen as relations between the factors were assumed. Oblique rotation maximizes the loadings when factors are correlated. Maximum Likelihood was chosen as extraction method. In this procedure the extracted parameters are improved iteratively according to a fit function.

Table 8. The pattern matrix from the exploratory factor analysis.

	Factor						
	SENSOR	INFO	INFO	MWL	SA	TEAM	PERF
radarco7		.316	-.189			.160	
sensor1	**.601**		.131	.134			.444
sensor2	**.586**	.139					.346
sensor3	**.682**	.143					
info1		**.808**					
info2		**.917**					
info3			**.881**				
info4			**.944**				
mwl1	-.704			**.256**			.203
mwl2	-.476			**.159**		-.159	.225
mwl3bed				**.711**			
mwl4	-.145			**.631**			
sa1			.251		**.385**	.212	
sa2					**.855**		
sa3					**.783**		
sa4					**.488**	.132	.171
team1	.179				.132	**.321**	.338
team2						**.891**	
team3						**.834**	
team4					.295	**.238**	.234
surv7							**.218**
tsurvmo7		-.181				.144	
hits7							**.288**
thitsmo7							**.134**

Extraction Method: Maximum Likelihood.
Rotation Method: Oblimin with Kaiser Normalization.
Values smaller than +/- .13 are suppressed.

In Table 8 the pattern matrix from the exploratory factor analysis is presented. The bold values in the table represent the hypothetically driven clusters hypothesized during the questionnaire design. The factor names in the heading are the labels specified by the author.

According to Kaiser's criterion (retaining factors with eigenvalues > 1.0), a 7 factors solution was suggested. This solution explains 65% of the total variance in the data. Figure 16 presents the plot of eigenvalues as a function of number of factors.

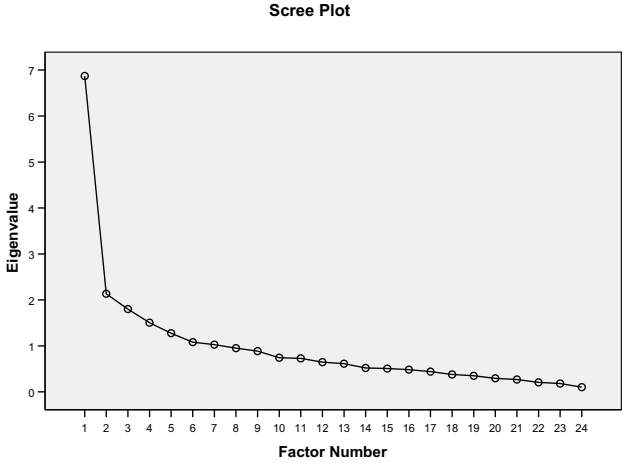


Figure 16. Plot of eigenvalues as a function of number of factors (Scree plot).

As indicated above the exploratory factor analysis finds seven factors with an eigenvalue over 1. Thus the analysis partly supports the constellation of hypothesized constructs and their manifest variables. Some comments concerning the conclusions drawn from the analysis are:

- The three subjective questionnaire items asking about sensor effectiveness are loaded in the same factor, while the technical sensor variable loads in the same factor as where the two questions concerning the information on the target display are loaded.

- The factor analysis places the four items that relate to usability of information in two different factors, which is in concurrence with earlier findings in Svensson (1999), i.e. the information on the two displays represents something different. Despite this, they will be combined into one construct in the current model development.

- For the factor labeled mental workload by the author, the loadings of the four questions are divided, and the two questions relating to temporal demands load more in the sensor factor than the two that are more directly related to mental workload assessments. Supported by Miller and Hart's (1984) view that time pressure is an important aspect of mental workload assessments the four questions were combined into one factor.

- The factor labeled situation awareness emerges rather distinctly in the factor analysis.

- The four items concerning teamwork appear together in the factor analysis even though two of the factor loadings are rather weak.

- For both the performance constructs that were hypothesized, the factor analysis fails to support them. The two variables related to offensive performance load on the same factor, but not strongly.

To conclude, the exploratory factor analysis partly supports the constructs chosen. However, it is important to remember that the explorative analysis gives **one** possible solution among many. An important aspect of the explorative analysis is that a tentative number of reliable factors can be estimated, as well as the amount of variance explained by the factors.

3.3.2 Confirmatory factor analysis

In accordance with Mulaik & Millsap's (2000) recommendations a confirmatory factor analysis (CFA) was also conducted with LISREL 8.80. In the CFA, the measurement model is tested, i.e. the relations between the manifest variables and the latent variables or constructs are specified, and all latent variables or constructs are allowed to covary. In Table 9 the factor loadings of the manifest variables on their latent variables are presented.

Table 9. Factor loadings in CFA.

	Factor						
	SENSOR	INFO	MWL	SA	TEAM	OFFPERF	DEFPERF
radarco7	**.27**						
sensor1	**.85**						
sensor2	**.82**						
sensor3	**.69**						
info1		**.54**					
info2		**.60**					
info3		**.32**					
info4		**.40**					
mwl1			**.76**				
mwl2			**.68**				
mwl3bed			**.24**				
mwl4			**.37**				
sa1				**.83**			
sa2				**.56**			
sa3				**.40**			
sa4				**.66**			
team1					**.80**		
team2					**.69**		
team3					**.57**		
team4					**.69**		
surv7						**.27**	
tsurvmo7						**.24**	
hits7							**.36**
thitsmo7							**.38**

Table 10 presents the factor intercorrelations (standardized solution) found in the confirmative analysis. All loadings and factor intercorrelations are significant ($p < 0.05$).

Table 10. Output matrix from CFA.

	SENSOR	INFO	MWL	SA	TEAM	DEFPERF	OFFPERF
SENSOR	1.00						
INFO	0.84	1.00					
MWL	-0.60	-0.62	1.00				
SA	0.51	0.65	-0.53	1.00			
TEAM	0.66	0.59	-0.47	0.76	1.00		
OFFPERF	0.33	0.13	-0.11	0.33	0.62	1.00	
DEFPERF	0.57	0.09	-0.40	0.53	0.74	0.93	1.00

```
Root Mean Square Error of Approximation (RMSEA) = 0.070
Comparative Fit Index (CFI) = 0.94
Standardized RMR = 0.067
Goodness of Fit Index (GFI) = 0.90
```

Indicated by analysis of the fit-indices of the CFA, the proposed measurement model is adequate as input in subsequent structural equation modeling analyses.

3.4 Structural model development I. Submodels

A number of submodels were developed as they can be considered the "building blocks" of the hypothesized model. The models have been modified with respect to the suggestions, or modification indices, provided by LISREL in order to improve model fit. The inclusion of error covariances must be justified, but all the allowed error covariances can be explained when the wordings of the manifest variables are analyzed. For example, in the submodel below with sensor effectiveness and usability of information, two error covariances (the curved arrows) are allowed. The INFO1 and the INFO2 variables both refer to the information on the target indicator, and it is reasonable to assume covariance in error as they relate to the same indicator. Similar cases can be made for the other error covariances allowed in the models. If these error covariances are removed, model fit decreases to some extent, but the basic structure of the model still holds.

3.4.1 Sensor effectiveness & usability of information

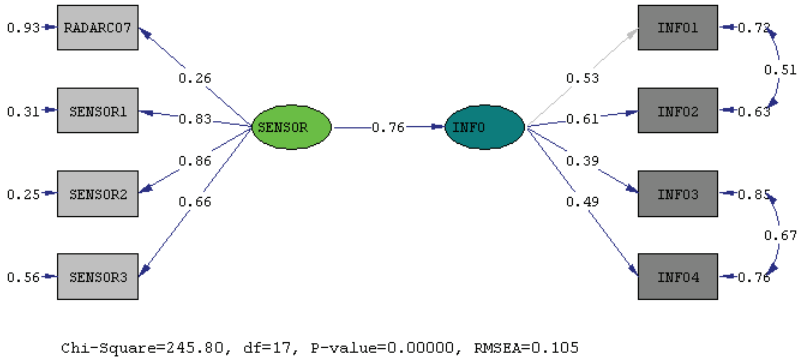


Figure 17. A screenshot from LISREL with the submodel SENSOR and INFO.

In Figure 17 a screenshot from LISREL is provided. In order to increase legibility and graphical control, the models from Figure 18 and onward will be presented in another graphical format. In these figures, the error terms, i.e. the numbers to the left of the RADARCO7-SENSOR3 variables, will be omitted.

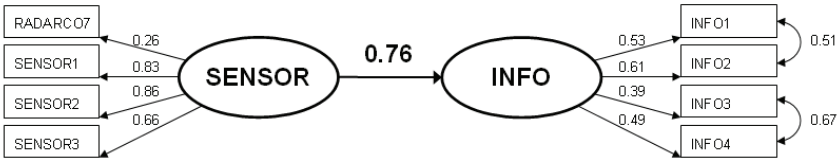


Figure 18. Submodel with SENSOR and INFO.

Model information and fit indices:

```
Chi-Square = 245.80
df = 17
P-value = 0.00000
RMSEA = 0.105
Comparative Fit Index (CFI) = 0.96
Standardized RMR = 0.063
Goodness of Fit Index (GFI) = 0.95
```

Here the error variance of the two questions concerning information on the target indicator was allowed to covary as they relate to the same indicator and an error covariance can be justified. The same is valid for the two questions concerning the tactical indicator. The conclusion from this submodel is that there is a strong effect, i.e. beta-weight, between SENSOR and INFO.

3.4.2 Usability of information & mental workload

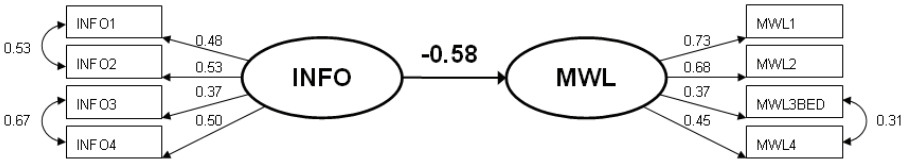


Figure 19. Submodel with INFO and MWL.

Model information and fit indices:

```
Chi-Square = 176.88
df = 16
P-value = 0.00000
RMSEA = 0.090
Comparative Fit Index (CFI) = 0.96
Standardized RMR = 0.047
Goodness of Fit Index (GFI) = 0.97
```

Here error covariance between two questions concerning mental workload was allowed. The conclusion from fit indices of this submodel is that there is a strong effect between INFO and MWL.

3.4.3 Mental workload & situation awareness

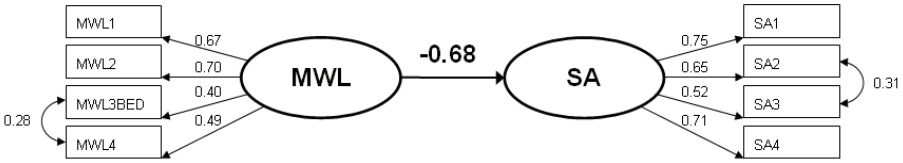


Figure 20. Submodel with MWL and SA.

Model information and fit indices:

```
Chi-Square = 137.40
df = 17
P-value = 0.00000
RMSEA = 0.076
Comparative Fit Index (CFI) = 0.97
Standardized RMR = 0.044
Goodness of Fit Index (GFI) = 0.97
```

In this model the error covariance between the SA2 and the SA3 variable was allowed along with the previous error covariance of two mental workload variables. Both SA questions ask whether the pilot was surprised, by the enemy position (SA2) and the enemy's actions (SA3), during the mission. The conclusion from fit indices of this submodel is that there is a strong negative effect between MWL and SA.

3.4.4 Situation awareness and teamwork

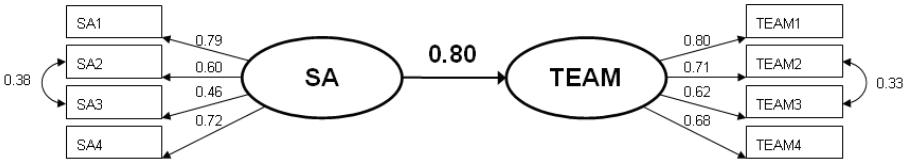


Figure 21. Submodel with SA and TEAM.

Model information and fit indices:

```
Chi-Square = 172.42
df = 17
P-value = 0.00000
RMSEA = 0.086
Comparative Fit Index (CFI) = 0.98
Standardized RMR = 0.046
Goodness of Fit Index (GFI) = 0.97
```

In this model the error covariance between the two SA items from the previous model is retained. Two of the questions asking about the teamwork were also allowed to have an

error covariance. The conclusion from fit indices of this submodel is that there is a strong effect between SA and TEAM.

3.4.5 Teamwork and performance

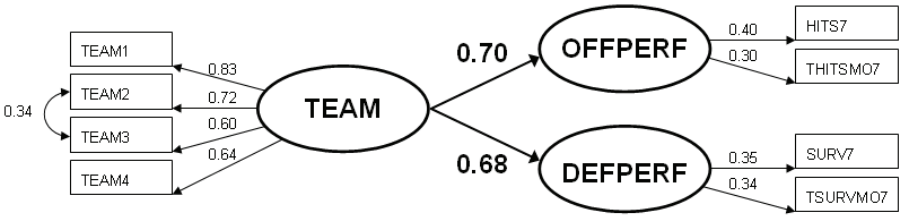


Figure 22. Submodel with TEAM and OFFPERF & DEFPERF.

Model information and fit indices:

```
Chi-Square = 115.23
df = 17
P-value = 0.00000
RMSEA = 0.069
Comparative Fit Index (CFI) = 0.97
Standardized RMR = 0.044
Goodness of Fit Index (GFI) = 0.98
```

In this model the effects between TEAM and DEFPERF and OFFPERF is presented. The conclusion from fit indices of this submodel is that there are rather strong effects between teamwork and the two performance-constructs.

3.4.6 Summary of submodel development

As can be seen from Figures 17–22, significant effects were found in all sub-models, and the model fit-indices were satisfactory. In Table 11, the results of the submodel analyses, that are related to the subhypotheses H1 – H6, are compiled. From the table it can be concluded that all the subhypotheses were corroborated.

Table 11. Summary of submodels analysis.

Hypotheses	Supported
H1: SENSOR EFFECTIVENESS is positively correlated to usability of INFORMATION	Yes
H2: Usability of INFORMATION is negatively correlated to MENTAL WORKLOAD	Yes
H3: MENTAL WORKLOAD is negatively correlated with SITUATION AWARENESS	Yes
H4: SITUATION AWARENESS is positively correlated with TEAMWORK	Yes
H5: TEAMWORK is positively correlated with OFFENSIVE PERFORMANCE	Yes
H6: TEAMWORK is positively correlated with DEFENSIVE PERFORMANCE	Yes

3.5 Structural model development II. Final model

3.5.1 Conceptual model

To reiterate, the hypothesized relationships between the selected constructs, i.e. the conceptual structural model, is shown in
Figure 23. The model represents a simplex structure, i.e. the constructs are ordered in a sequence from sensor effectiveness to operative performance.

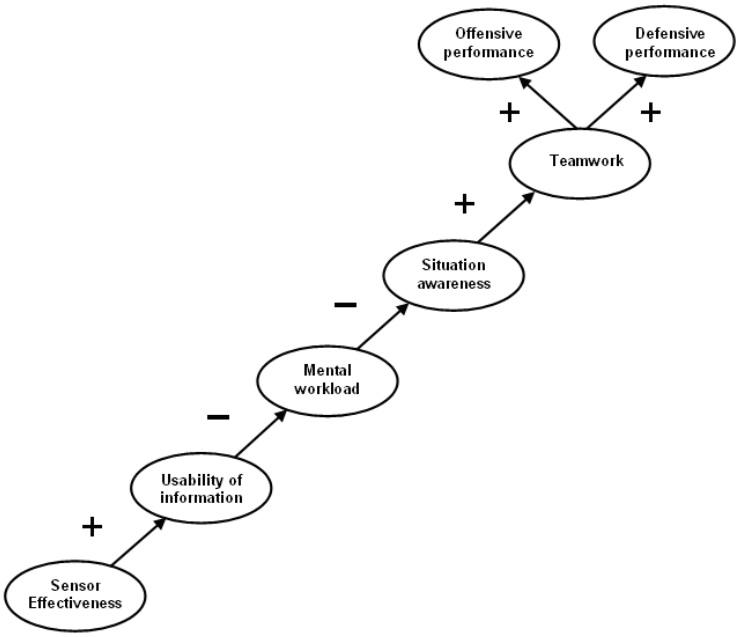


Figure 23. Same as Figure 10. Conceptual model and main hypothesis of the thesis.

3.5.2 Model with data, n = 848

The larger conceptual model, composed of the submodels, was specified and tested on a database with 848 cases.

The full path diagram, with estimates for the measurement model and the structural model, along with associated fit indices, is provided in Figure 24. In the figure a screenshot from the LISREL software is shown, slightly adjusted to enhance legibility.

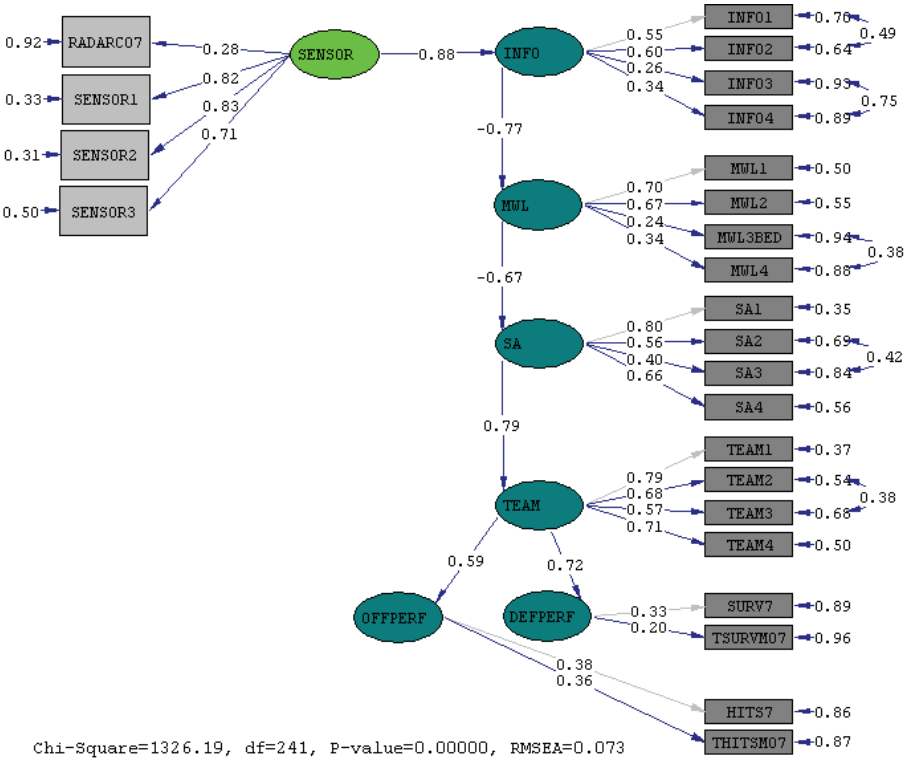


Figure 24. LISREL screenshot of full structural model with n = 848. All effects are significant ($p < 0.05$).

In Figure 25 a graphically enhanced version of the same model is presented.

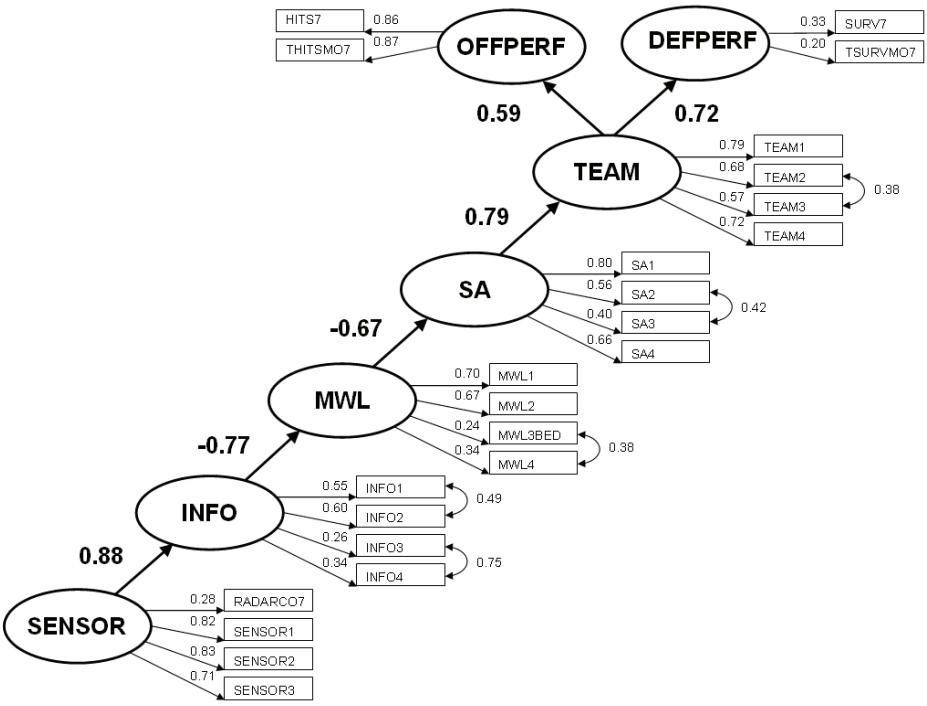


Figure 25. Full model with parameter estimates, n = 848. All effects are significant ($p < 0.05$).

Model information and fit indices:

```
Chi-Square = 1326.19
df = 241
P-value = 0.00000
RMSEA = 0.073
Model CAIC = 1783.02
Comparative Fit Index (CFI) = 0.93
Standardized RMR = 0.080
Goodness of Fit Index (GFI) = 0.88
```

The model is very restrictive, which explains the high chi-square value. The chi-square statistic is also dependent on the number of cases (Diamantopoulus, 2000, p. 94) and the large sample of this study is a circumstance explaining the high chi statistic and the significance.

Table 12 presents the correlations between the constructs. As can be seen, the correlations decrease as one move away from the main diagonal, meaning that the correlations decrease as a function of the number of lags between the factors – a typical feature of a simplex correlation structure (Jöreskog & Sörbom, 1984).

Table 12. Output matrix from full model with data, n = 848.

	SENSOR	INFO	MWL	SA	TEAM	DEFPERF	OFFPERF
SENSOR	1.00						
INFO	0.88	1.00					
MWL	-0.68	-0.77	1.00				
SA	0.45	0.51	-0.67	1.00			
TEAM	0.36	0.41	-0.53	0.79	1.00		
DEFPERF	0.26	0.29	-0.38	0.57	0.72	1.00	
OFFPERF	0.21	0.24	-0.31	0.46	0.59	0.42	1.00

After analysis of fit indices and factor intercorrelations, the conclusion of the author is that the model can, in terms of a simplex, explain the empirical relationships between the manifest variables measured. Accordingly, the relations between 24 empirical measures can be explained in terms of a 7-factors simplex structure.

In the cross-validation step, the fit of the proposed model was tested in a validation sample with 384 cases, with data from the weeks 350, 409 and 450. The result of the cross-validation is presented in Figure 26.

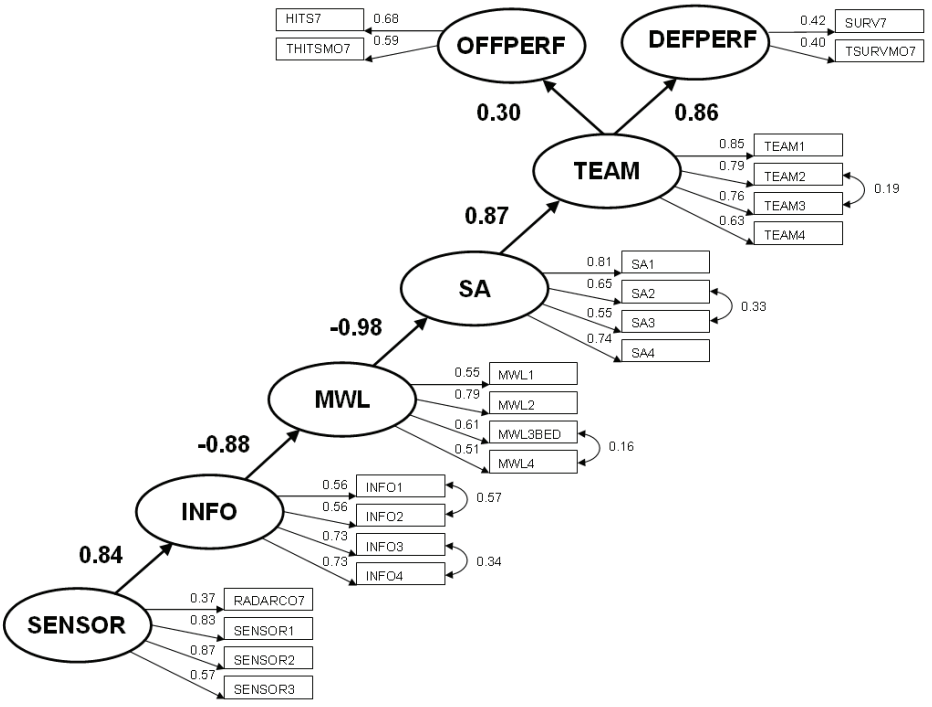


Figure 26. Cross-validation of full structural model with n = 384. All effects are significant ($p < 0.05$).

Model information and fit indices:

```
Chi-Square = 809.15
df = 241
P-value = 0.00000
RMSEA = 0.078
Model CAIC = 1219.23
Comparative Fit Index (CFI) = 0.96
Standardized RMR = 0.067
Goodness of Fit Index (GFI) = 0.85
```

Table 13 presents the correlations between the constructs or factors. The matrix is very similar to the one presented in Table 12, and the correlation between the two matrices is 0.96, ($p < 0.05$).

Table 13. Output matrix from cross-validation with full model, n = 384.

	SENSOR	INFO	MWL	SA	TEAM	DEFPERF	OFFPERF
SENSOR	1.00						
INFO	0.84	1.00					
MWL	-0.74	-0.88	1.00				
SA	0.73	0.86	-0.98	1.00			
TEAM	0.63	0.75	-0.85	0.87	1.00		
DEFPERF	0.54	0.65	-0.73	0.75	0.86	1.00	
OFFPERF	0.19	0.22	-0.25	0.26	0.30	0.26	1.00

Based on the fit indices and similarities in the factor intercorrelations, the conclusion was that the cross-validation supported the main model. The differences in beta-weights between the constructs of the models are summarized in Table 14.

Table 14. Differences between beta-weights of the main and the cross-validation model.

Relation	n = 848	n = 384	Difference
SENSOR to INFO	0.88	0.84	0.04
INFO to MWL	-0.77	-0.88	0.11
MWL to SA	-0.67	-0.98	0.31
SA to TEAM	0.79	0.87	0.08
TEAM to OFFPERF	0.59	0.30	0.29
TEAM to DEFPERF	0.72	0.86	0.14

The major differences between the beta-weights for the main model and the cross-validation model are that the weight between TEAM and OFFPERF is substantially lower, and the weight between MWL and SA is substantially higher in the cross-validation model. However, and most important, the number of constructs, and the structure of the main model are strongly supported by the cross-validation model.

3.5.3 Combining the datasets

In order to explore further the similarities and differences between the individual weeks the correlations (product moment) between the correlation matrix of all variables for all weeks and the correlation matrices of each week were computed. As can be seen from Table 15, there is a range in correlation from 0.69 to 0.89. Thus, the shared variance of the correlation structures range from 47 to 79 percent. A conclusion is that there is a basal structural similarity over the weeks, but, for some weeks, situational conditions changed the correlational structure. This is to be expected as the sensor alternatives studied during different weeks provided different preconditions for the pilots. This analysis can also be regarded as evidence concerning the test-retest reliability between the weeks. Note that the figures contain some auto-correlation as each separate week was compared with the full database.

Table 15. The correlations between correlation structures of each week and the correlation structure based on the total sample.

	w450	w409	w350	w325	w323	w321	w248	w237	w235
Correlation (*r*) with full database	0,89	0,88	0,88	0,88	0,87	0,84	0,76	0,74	0,69

Plots of two examples of weeks, representing the highest and lowest similarity, are presented in Figure 27 and Figure 28, respectively.

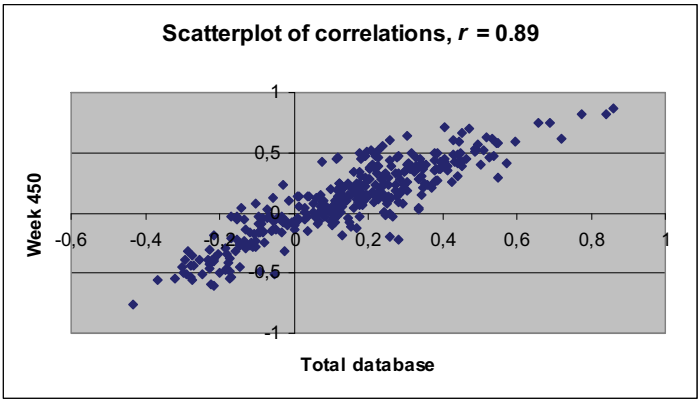


Figure 27. Scatterplot of correlation of correlation structures for week 450 vs the full database.

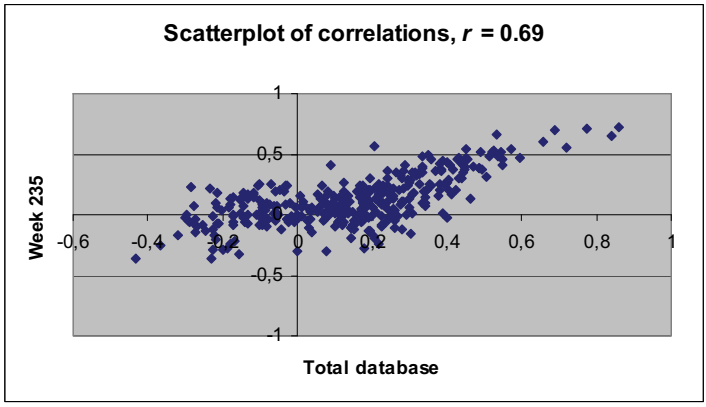


Figure 28. Scatterplot of correlation of correlation structures for week 235 vs the full database.

The relative consistency of the correlation structures found over the weeks supports the integration of both datasets to one database with 1232 cases. All analyses below use the full database.

3.5.4 Final model, n = 1232

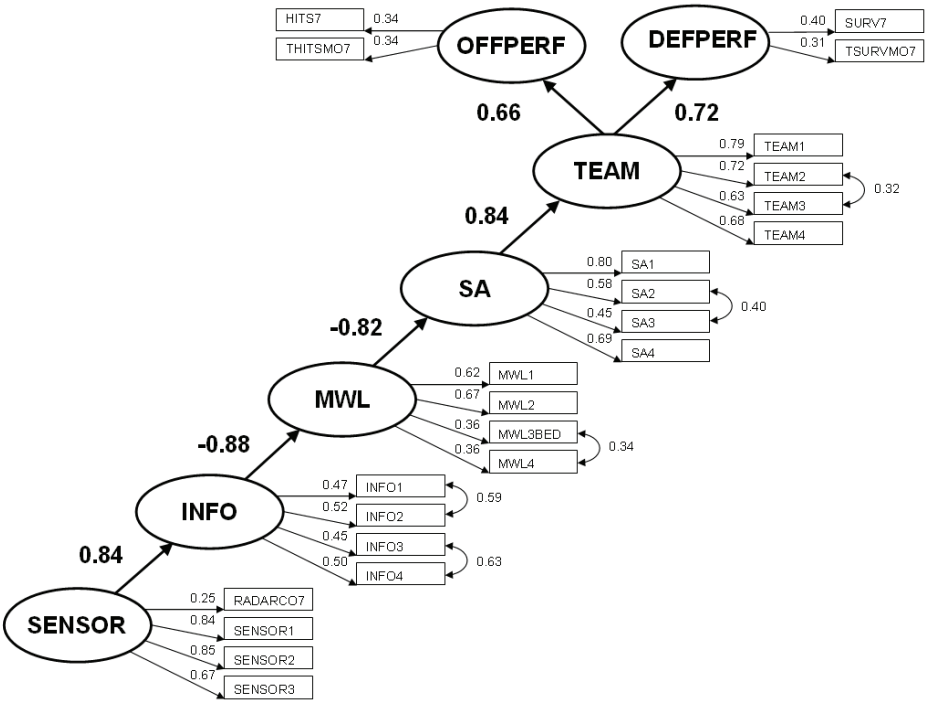


Figure 29. Full structural model with n = 1232. All effects are significant ($p < 0.05$).

Model information and fit indices:

```
Chi-Square = 1756.42
df = 241
P-value = 0.00000
RMSEA = 0.071
Model CAIC = 2235.29
Comparative Fit Index (CFI) = 0.94
Standardized RMR = 0.066
Goodness of Fit Index (GFI) = 0.89
```

Table 16. Correlations between the constructs in the full and final model.

	SENSOR	INFO	MWL	SA	TEAM	DEFPERF	OFFPERF
SENSOR	1.00						
INFO	0.84	1.00					
MWL	-0.74	-0.88	1.00				
SA	0.60	0.72	-0.82	1.00			
TEAM	0.50	0.60	-0.68	0.83	1.00		
DEFPERF	0.36	0.43	-0.49	0.60	0.72	1.00	
OFFPERF	0.33	0.40	-0.45	0.55	0.66	0.48	1.00

As shown earlier, the correlations decrease as one move away from the main diagonal, meaning that the correlations decrease as a function of the number of lags between the factors, and which is a typical feature of a simplex correlation structure (Jöreskog & Sörbom, 1984). This is illustrated in Figure 30, which shows the effects (direct and indirect) of the SENSOR construct on the other constructs of the simplex.

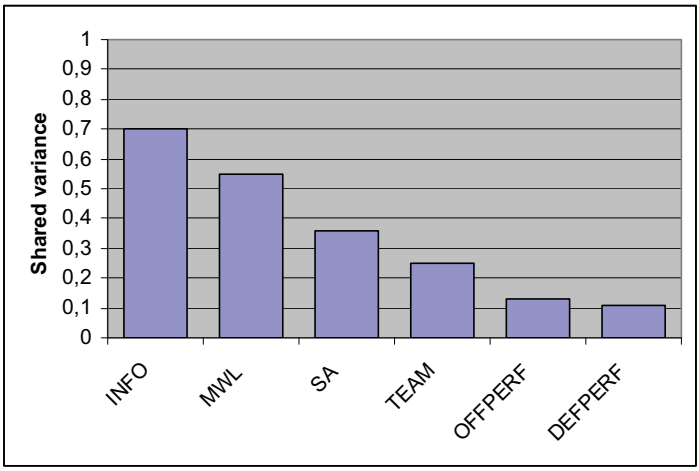


Figure 30. Direct (SENSOR to INFO) and indirect effects of SENSOR on the other constructs of the model in terms of shared variance.

Comparison between fit indices and factor intercorrelations from the analysis of the full and final model (n = 1232) with the initial model (n = 848), shows that the general fit is as good for the full and final model. Accordingly, the correlations between 24 empirical measures can be reduced to and explained in terms of a 7-factors simplex structure. The theoretical and practical implications will be discussed later.

3.5.5 Validation of the simplex structure

The simplex structure of the initial hypothesis, investigated in section 3.5.5, rests on strong assumptions. In order to analyze further whether these assumptions bias the conclusions from the final model analyses, the output factor intercorrelation matrix from the CFA was compared with the corresponding output matrix of the full and final model analysis. When the model is specified and the constructs forced into a simplex structure, it might result in an output matrix more reflecting the command file rather than the true factor intercorrelations of the CFA. If these two matrices were radically different it would be an indication that the proposed simplex structure is forced, and faulty. However, the correlation between the two matrices, illustrated in Figure 31, is 0.98.

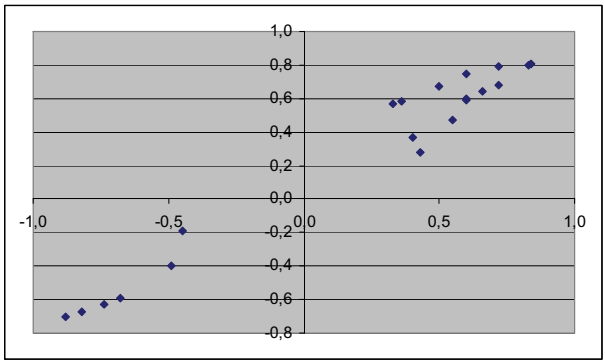


Figure 31. Scatterplot of correlation between output matrix from CFA and output matrix from the full structural model, ***r*** = 0,98.

A Multi Dimensional Scaling (MDS) analysis (Wilkinsson, 1990) was run with the covariance matrix from the full structural model as input. The result is illustrated in Figure 32. For a correct transformation of covariances to similarities, the negative covariances of MWL were reversed to positive covariances. As can be seen from the figure, the MDS analysis completely replicates the conceptual model or main hypothesis presented in Figure 23. Dimension 1 of the figure represents the simplex structure.

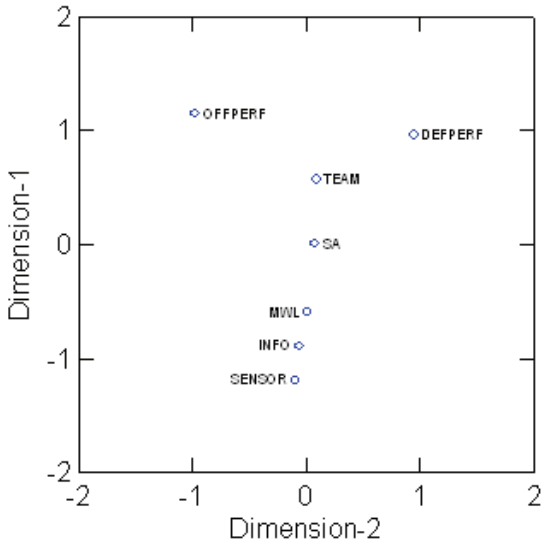


Figure 32. MDS analysis of output matrix from full model. Distances interpreted as similarities. (Guttman/Lingoes coefficient of alienation = 0.02).

3.5.6 Nested model. Model with "perfect measurement"

In order to further examine the database a model where the factor covariance matrix, instead of the raw data was used as input, and the latent variables used earlier were used as manifest variables[19].

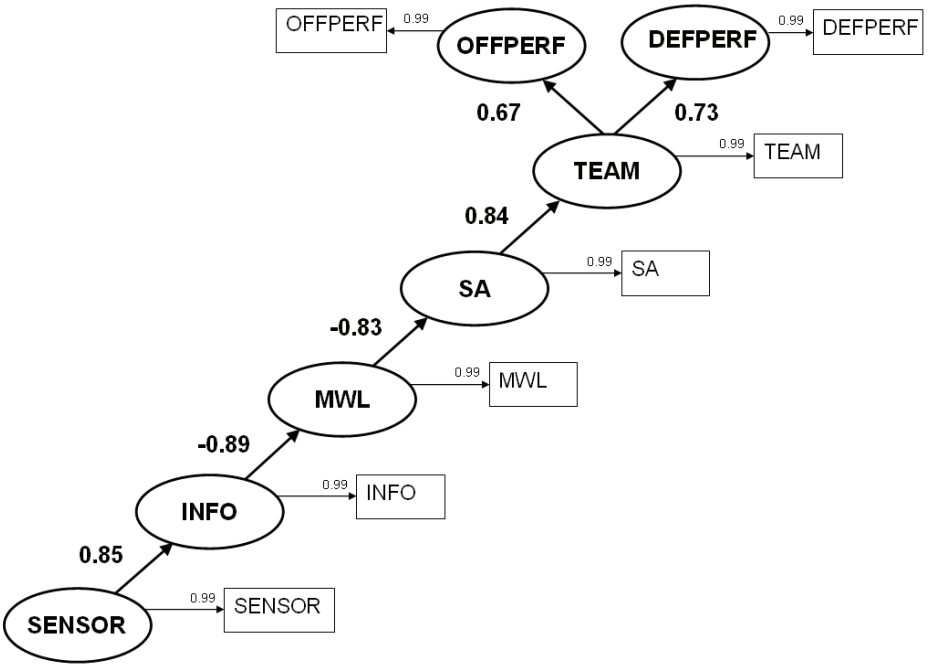


Figure 33. Model with "perfect measurement".

Model information and fit indices:

```
Chi-Square = 3.83
df = 15
P-value = 0.99825
RMSEA = 0.000
Model CAIC = 109.34
Comparative Fit Index (CFI) = 1.00
Standardized RMR = 0.0079
Goodness of Fit Index (GFI) = 1.00
```

Since the new constructs were almost perfectly "measured", the error variances estimated were close to zero (they were fixed to 0.01). With respect to the measurement errors, the model represents a perfect simplex structure. As expected, the fit indices become close to

[19] Note that this is not a recommended approach, as LISRELs ability to manage measurement error is "disabled", but it was conducted in order to further explore the properties of the dataset.

perfect. The weights in the model are quite the same as those found for the full and final model in Figure 29.

3.5.7 Nested model. Model with reduced number of manifest variables

Here a nested model is presented where only two choice manifest variables per construct have been kept – in all 14 measures. Thus, the same structural model is investigated, but the measurement model is reduced, in order to see how fit indices change. A reduction of the number of markers can deflate construct validity and reliability of the constructs, but if a model with a reduced number of manifest variables can be used, the practicality and easy of use of the survey increase as the pilots have to answer fewer questions.

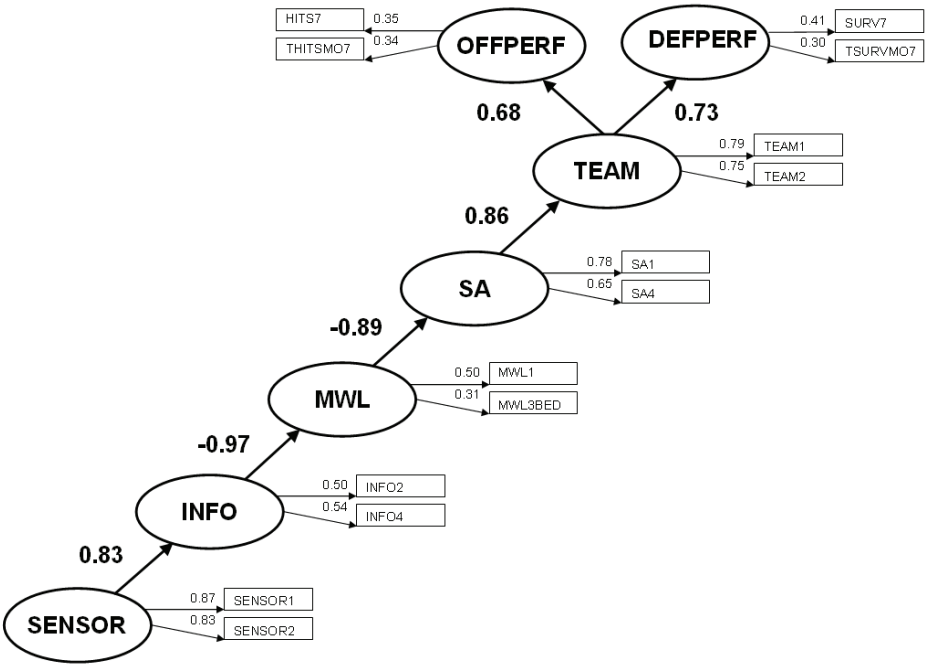


Figure 34. Model with reduced number of manifest variables.

Model information and fit indices:

```
Chi-Square = 462.47
df = 71
P-value = 0.00000
RMSEA = 0.067
Model CAIC = 738.43
Comparative Fit Index (CFI) = 0.95
Standardized RMR = 0.049
Goodness of Fit Index (GFI) = 0.95
```

3.5.8 Nested model. Minimalistic model of SENSOR directly to OFFPERF & DEFPERF

A minimalistic model, where the whole "model of the operator" was omitted, was also specified. Thus, this model is almost a black box model (although the SENSOR construct still contains three subjective ratings) with little diagnosticity with regards to the functional state of the operators.

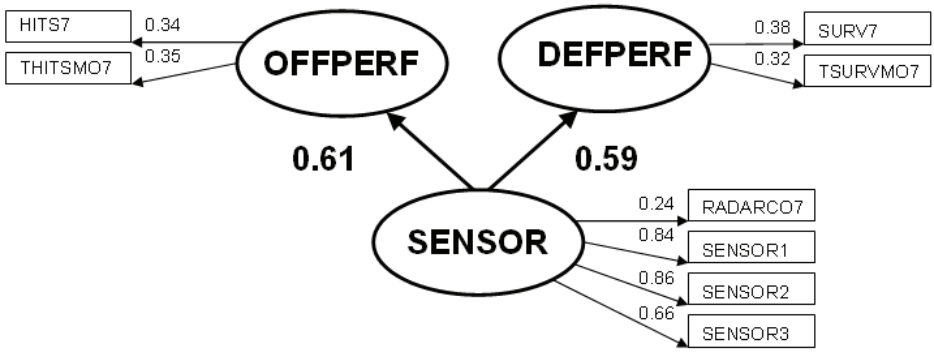


Figure 35. Minimalistic model.

Model information and fit indices:

```
Chi-Square = 160.38
df = 18
P-value = 0.00000
RMSEA = 0.080
Model CAIC = 306.48
Comparative Fit Index (CFI) = 0.94
Standardized RMR = 0.055
Goodness of Fit Index (GFI) = 0.97
```

3.5.9 Nested models. Models with one or more constructs left out

A number of nested models were developed where one or more of the endogenous constructs were left out. Descriptions of these models, beta weights and fit indices, are presented in Table 17.

The models have been named according to which construct(s) that have been omitted. For example, the model labeled No INFO, MWL or SA in Table 17 below is a model where only the teamwork construct is used as the mediating factor between sensor effectiveness and performance. As apparent a number of nested models can be described, which exhibit similar properties in terms of fit. The two best fitting models are the models where the number of manifest variables have been reduced and the model where only the TEAM construct mediates between SENSOR and OFFPERF/DEFPERF.

Table 17. Comparison of a number of nested models.

Beta weight in to construct in model	Full model	Red. Measm. model	No INFO	No MWL	No SA	No TEAM	No INFO or MWL	No INFO, MWL or SA	No INFO, MWL or TEAM
→ INFO	0.84	0.83		0.80	0.87	0.84			
→ MWL	-0.88	-0.97	-0.73		-0.89	-0.84			
→ SA	-0.82	-0.89	-0.78	0.83		-0.76	0.67		0.61
→ TEAM	0.84	0.85	0.83	0.85	-0.75		0.86	0.69	
→ OFFPERF	0.65	0.68	0.67	0.68	0.67	0.52	0.68	0.71	0.59
→ DEFPERF	0.72	0.73	0.72	0.72	0.71	0.64	0.73	0.73	0.68
GFI	0.89	0.95	0.92	0.92	0.91	0.90	0.94	0.97	0.96
CAIC	2235	738	1531	1198	1658	1768	891	482	546
CFI	0.94	0.95	0.95	0.95	0.94	0.92	0.97	0.97	0.95
S RMR	0.066	0.049	0.060	0.061	0.063	0.069	0.051	0.042	0.051
RMSEA	0.071	0.067	0.070	0.069	0.076	0.078	0.063	0.058	0.066
ChiSquare	1756	463	1142	1100	1291	1370	583	255	319

3.5.10 Alternative model. Conceptual model from Nählinder et al., 2004

As the model, describing the reoccurring patterns of difficulty→ mental workload→ situation awareness→ performance, (Svensson, 1999; Nählinder et al., 2004) was used in the conceptualization of the models of this thesis the fit of the data to this model was also tested.

Difficulty was operationalized and measured by four subjective questions:

- How difficult do you think the engagement was?
- How challenging was this engagement?
- How dangerous do you think the enemy was?
- How large risks did you take during this engagement?

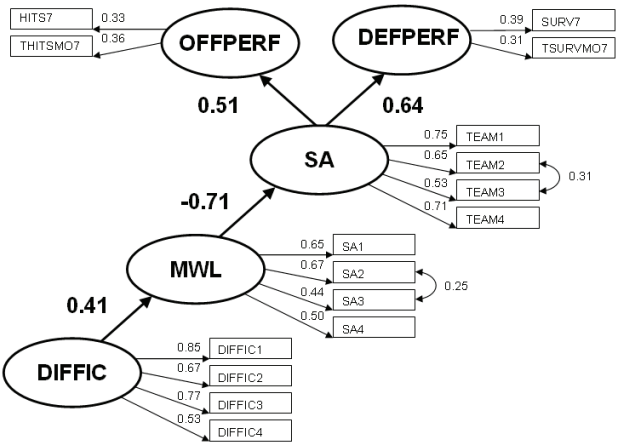


Figure 36. Alternative model based on Nählinder, et al., 2004.

Model information and fit indices:

```
Chi-Square = 771.98
df = 98
P-value = 0.00000
RMSEA = 0.075
Comparative Fit Index (CFI) = 0.92
Standardized RMR = 0.069
Goodness of Fit Index (GFI) = 0.93
```

This model describes how the (perception of) difficulty of the engagements relates to mental workload, situation awareness and operative performance. Accordingly, the model found in several other studies was confirmed.

3.5.11 Sources of variance - intra- , inter-individual, and situational variance

As noted earlier, the covariances or the shared variance estimated, are based on inter-individual, intra-individual, and situational variance. In order to compare differences between pilots with respect to their intra-individual contribution, the correlations between covariance structures for the 24 variables for each pilot, were compared to the corresponding structure based on the whole database. The individual structures represent the intra-individual variances, i.e. the variance dependant on the repeated measures, and the situational variances. Accordingly, the correlations represent estimates of each pilot's (intra-individual) variance contribution to the total variance. The individual matrices, for the 24 manifest variables, for half of the pilots (from the one with most engagements to the number 19), were compared to the total matrix. The mean correlation (product-moment) found was 0.75, with a range from 0.89 to 0.57. The squared correlation (r^2) represents the extent to which the intra-individual variances explain the variances of the total covariance matrix, and thus, on average, 56 percent of the total variance is explained.

A conclusion of the author is that the intra-individual variance component contributes significantly to the covariance structures used in the SEM-analyses. The conclusion is of importance for several reasons.

From an empirical point of view, we are interested in variance sources reflecting situational changes or changes in experimental conditions, and we are less interested in the variance between subjects. Accordingly, in this operational study, our interests concern the actions of the pilots as a function of the situational changes and not the differences between Subject Matter Experts (SME's). The estimates of proportions of explained variance indicate that systematic changes in the operational situations affect the correlational structure on which the models are based.

Also, from a statistical point of view, it is important to identify and be in control of the sources of variance behind a covariance measure. The military pilots of the study are considered a homogenous group of qualified specialists on military aircraft system handling, i.e. SME's. Accordingly, from a statistical point of view, they do not represent independent measures. The fact that they are repeatedly used as measurement instruments may also in the same way, violate statistical independency of measures criteria. From a puristic statistical point of view these circumstances constitute a problem, because they can affect the covariances, and the statistical analyses. Unfortunately, the impact of these circumstances is hard to estimate, and accordingly, in most cases the effects are not known.

From a practical point of view, it is often difficult to meet statistical assumptions in the selection of cases and conditions in operational situations. For example, the number of Swedish SME's as military pilots or cardiac surgeons is generally too low for multivariate statistical analyses. On the other hand, by combining inter- intra-, and situational variance sources in repeated measurement studies, databases of practical and theoretical importance can be developed. Studies by Angelborg-Thanderz (1990),

Svensson et al. (1999), and Magnusson (2002) are examples representing the practicability of the procedure.

Questions referring to the sources of variance accounting for the covariances used in multivariate analyses are seldom asked for, and quite often statistical independency criteria are violated in applied studies. Multivariate techniques are, as here, often used for data reduction and dimensionality estimates of different kinds of questionnaires. The present state of the art on, for example, studies of personality dimensions, are entirely based on series of questions reflecting different personality aspects, and the subjects have to answer the questions, one after the other, and at the same time (e.g., Svensson et al., 2008). In this type of databases, halo-effects or cognitive biases are to be expected. As a contrast, clinical medical research (e.g. Alehagen, Svensson & Dahlström, 2007), where specific clinical examinations by cardiologists, measures of blood pressure, echocardiographic measures, and measures of peptides represent almost independent measures as a base for multivariate analyses.

A theoretically possible design could have been to use different observers for rating each of the constructs or their markers. In that case, the independency of measures criteria had been fulfilled. However, the study of that design had not been possible to perform. Accordingly, there has to be a balance between statistical stringency and practicable designs in operational settings with SME's. It is the present author's opinion that the utility advantages of using multivariate analyses as statistical tools in operational settings outweigh the disadvantages in terms of risks for false conclusions.

4 Discussion

The popularity of scientific constructs waxes and wanes. Mental workload received extensive attention during the 1980s, situation awareness was discussed intensely in the 1990s, while teamwork currently receives more attention. The final model that is proposed here is an unusually large structural equation model that spans over and integrates these theoretical constructs, anchored in system-oriented variables at both the beginning and the end of the model. As compared to most former studies, this study encompasses the lion's share of the operational constructs in use, and the model proposed spans over and relates them in a logical way.

Thus, the model stretches from effectiveness in a technical system, to the cognition of the individual, to the team and ends in the operative outcome or performance – a model of the operators and their functional state mediating between technical parameters. The relations between constructs found in the data are stronger, and thus contain a larger amount of shared variance, than what usually has been found, which allows for stronger theoretical statements. Compared to earlier similar modeling efforts, the teamwork construct have been added and the operationalization of performance is done through more "undebatable" measures. The model fit is reasonable given the size of the dataset and the model. Thus, all the subhypotheses, as well as the general simplex hypothesis, formulated in section 1.9 are supported, as presented in Table 18.

Table 18. Conclusions concerning subhypotheses.

Hypotheses	Supported
H1: SENSOR EFFECTIVENESS is positively correlated to usability of INFORMATION.	Yes
H2: Usability of INFORMATION is negatively correlated to MENTAL WORKLOAD.	Yes
H3: MENTAL WORKLOAD is negatively correlated to SITUATION AWARENESS.	Yes
H4: SITUATION AWARENESS is positively correlated to TEAMWORK.	Yes
H5: TEAMWORK is positively correlated to OFFENSIVE PERFORMANCE.	Yes
H6: TEAMWORK is positively correlated to DEFENSIVE PERFORMANCE.	Yes
H7: All chosen constructs are useful and appropriate (valid and reliable) in a SEM model describing the collected data. H7 is a consequence of H1-H6.	Yes
H8: It is possible to describe the constructs in a simplex or quasi-simplex structure, while retaining reasonable model fit.	Yes

The fact that the empirical data is expressed as representing a simplex structure implies a rather strong theoretical statement, and is a scientific contribution for the researchers of the field to debate. That, for example, the SA construct precedes the TEAM construct in the model, while the communication and teamwork within the team surely must be a part of the SA building process, could for example, be an interesting issue to debate. Similarly, the confirmatory factor analysis shows a 0.53 correlation between SA and DEFPERF, but no direct effect have been included in the model, so the effect is mediated by TEAM. For a pilot that flies first in a formation, and thus first gets radar contact and launches a missile, his individual situation awareness probably is more important than the teamwork, at least during the initial stage of the engagement. As stated before, any model represents a trade-off between simplicity and explanatory power. The proposed model captures the lion's share of the variance of the empirical data, while retaining reasonable model fit. The data necessary to specify alternative and competing structural models are presented in the thesis, providing a researcher with an alternative hypothesis with the possibility to formulate a competing model.

From a methodological perspective, the thesis describes a quasi-experimental approach that is rarely seen in the Human Factors field and is by that a unique contribution. Given the attention these constructs have received and the existence of suitable statistical model development methods (e.g. structural equation modeling) that also have received extensive attention, it is still very unusual within Human Factors studies to see models quantifying the relations between the constructs. Modeling of this kind can perhaps provide the type of quantitative justification sought for by the Human Factors field. The measurements used also demonstrate how rather straightforward operationalization, that is able to distinguish between multi-dimensional constructs, can be used in applied settings.

From an epistemological perspective, the modeling effort describes a way to increase knowledge generation in a system or concept development process. The knowledge produced in the radar study where data was collected, is an example of how knowledge typically is developed in military acquisition studies – a new technology is developed, introduced and tested in simulators. Knowledge is generated by researchers and developers together with a user group in a joint production of knowledge. The radar study is an example of a rather broad sociotechnical study where the introduction of new technological capabilities change the established work practice of the pilots, with a resulting need to develop new tactical behavior. The questions of the radar study matured and changed during the course of the study, and the more knowledge generated in order to track the changes in the pilots, the better the understanding of their performance.

The proposed model represents a rather strong statement in terms of a simplex structure, and several points of criticism and questions can be raised against the model and the general research approach. Below, a number of issues are presented that have to be considered in order to be able to accept the models, their consequences, and the general message of the thesis. When reading through these possible points of criticism, some of Herbert Simon's thoughtful words from *Models of my life* (Simon, 1996) should be kept in mind:

"The true line is not between 'hard' natural science and 'soft' social sciences, but between precise science limited to highly abstract and simple phenomena in the laboratory and inexact science and technology dealing with complex problems in the real world."

Multi-dimensional constructs
Issue: All the constructs that are used in the thesis are complex, multifaceted and have been discussed with regards to their scientific justification. No final and precise definitions exist.

Answer: Final definitions of these types of constructs probably never will exist. Phenomena or concepts like intelligence, emotion, trust, common ground, et cetera, that are studied within the social sciences are inherently complex and multi-dimensional. New variables or markers of concepts are developed and refined, which also change the characteristics of the concepts.

Issue: The constructs used in the modeling are described at a too shallow level and they share meaning with each other.

Answer: The constructs certainly can be, and have been elsewhere, further dissected and discussed. The thesis attempts to present an integrative approach and a perspective where the findings of other researchers are used to be able to describe “the big picture”.

Sources of variance
Issue: The observations or cases in the database are treated as independent, while some structural variance in the data must be assumed to be a result of the fact that is was 37 pilots who generated the 1232 cases. The pilots also participated to varying degrees and this is not controlled for.

Answer: It is probably true that each pilot rate according to an individual response pattern and hence parts of the correlations in the database can be a result of this. This assumes that individual response patterns are alike, otherwise the bias of each pilot contributes with unique variance. A number of uncontrolled sources of variance exist, and it is uncertain whether response patterns, along with differences between engagements, also depend on pilot’s experience, trim on system, or the fact that they are members of a very select group of persons, which in some ways, due to training or selection, are unusually homogenous.

Other similar issues also exist. For the teamwork construct the four pilots are relating to the same teamwork process, although from their own perspective. This means that the ratings for teamwork cannot be said to be entirely independent, even within one engagement. The THITSMO7 (team hits minus own) and TSURVMO7 (team survival minus own) also depend on the performance of the same team and cannot be said to be entirely independent.

Issues like those above are prevalent in many Human Factors studies outside the laboratory. It would be very helpful to have more statistical methods available to help in the analysis of sources of variance, but this goes beyond the scope of the thesis.

Causality in the model
Issue: LISREL or SEM in itself provides no clues with regard to causality in a model and the causality in a model must always be justified.

Answer: For the final model the causality is based on both patterns found earlier and the fact that the manifest performance variables are final outcomes of the process. The conceptualization goes from technical systems to larger and larger systems and from individual cognition to team interaction and then interaction with the enemy in terms of hits and survival.

It is correct that a model that starts in performance that leads to teamwork, which leads to situation awareness, et cetera, i.e. a model with "reversed causality" is equally valid from a statistical standpoint. However, the empirical and logical reasons, and temporal ordering of events underlying such a model would be hard to defend.

Specification error
Issue: Do the measurement model really capture all relevant aspects, or could there be other factors that contribute with variance?

Answer: A rather large proportion of unexplained variance exists in the data, although not more than in many other studies. The unexplained variance is largest for the "objective" variables extracted from the simulator logs. Two major factors that affect the outcome of these variables are skill and tactics choice and these two factors were not assessed. The skill level of the pilots is dependent on a number of factors such as experience in terms of flight hours and current flight trim, both in the real aircraft and the simulator. The differences in understanding and skill in execution of the new tactics, that some of the studied radar alternatives allowed, also generated variance.

Demographic data, e.g. the number of flight hours, could be used to form a general experience factor that probably could be included in the model. Also a post hoc ratio between HITS and SURVIVAL could be computed as used as a skill variable. However, as the purpose of the SBA study was to study the effects of sensor effectiveness on performance, the skill construct was not included in the modeling effort.

Another major confounding factor was how the tactics and risk level chosen by one side, for that current engagement was "compatible" with the other side's choice of tactic and risk level. If both sides chose a very aggressive tactic, the engagement quickly becomes very intense and a higher number of kills on both sides are to be expected. Explanations on a very fine-grained level build up much of the variance of the performance variables. For example, the pilot had to break his radar contact just a few seconds too early in order to avoid an incoming missile or his evasive maneuver was only almost good enough.

These types of minidecisions govern whether a missile hits or misses and whether a pilot survives the mission or not.

Although for example Alberts & Hayes (2002) state that it is unrealistic to expect nearly perfect predictability given the complexity of human behavior at these individual and organizational levels, it would be desirable to develop ways to assess the "compatibility" of tactics. This would probably lead to that more variance could be explained.

Learning effects
Issue: Subjects in any study can exhibit significant learning effects when repeatedly faced with a similar task. During a week in the radar study the pilots flew in similar conditions 15–20 times on the blue side and as many engagements on the red side. The pilots also reviewed the course of events after each engagement, and learning effects must be assumed.

Answer: Apart from the development of the structural equation model(s) presented above, a number of other statistical analyses have been performed. In order to check whether any learning effects can be seen over the course of each separate week, a number of correlation analyses between engagement number in a specific week and all other variables were performed. Here, the question was whether any structural differences in the objective or subjective variables were seen over a week, due to the presumption that the pilots became better or learned the scenario. Only a few noteworthy correlations appeared in the data for a few of the variables. The patterns are inconsistent over the weeks, which led to the conclusion that any consistent learning effects (such as learning when and where to expect the enemy, and any specifics of the scenario) is countered by the fact the knowledge on the enemy side also increased over the week, as the pilots switched sides after three or six engagements.

Operationalization of constructs
Issue: Each construct could have been operationalized more extensively or differently. More globally comparable measures of, for example, mental workload exist.

Answer: The operationalization of constructs can almost always be expanded and be made more resource intensive, and perhaps, even more valid and reliable through additional ways of measurement. However, the thesis focuses on what is practically useful and to describe a methodological approach that researchers can use in order to increase the scientific and operational value of human performance measurement.

For example, NASA-TLX (Hart & Staveland, 1988) is probably the subjective rating technique for mental workload measurement that has been in most widespread use. The multi-dimensional nature of the NASA-TLX was designed in order to provide greater diagnosticity than, at the time, more traditional global measures of workload. NASA-TLX assesses 6 major dimensions: 1) mental demand, 2) physical demand, 3) temporal demand, 4) satisfaction in performance, 5) effort and, 6) frustration level. However, in Svensson et al. (1997) problems with the NASA-TLX workload construct was presented. In an MDS analysis the performance aspect was clearly an outlier, while the other markers cluster close to each other, i.e. they express the same thing. The construct

validity of the NASA-TLX is thus in question and this is the main reason why this measure was not used in the study. Also, as the purpose of the current study was not to further dissect workload, the assessment provided by the Bedford scale was deemed sufficient.

When contemplating the specific measures used in the measurement model it could be that the measures can be considered as instances of the current status of questionnaire items and their applicability within a specific study. The constructs are more "eternal" and can be found again with new or updated measures as manifest variables.

Subjective ratings
Issue: Subjective ratings can be biased by a number of factors.

Answer: The subjective ratings may be, for example, by the overall outcome of the engagement, i.e. halo error, rather than being independent assessments of the questions. During the course of the model development the underlying database have been visually inspected and even engagements where the pilot did not survive usually contain ratings which seem unaffected by the fact that the pilot was shot down. At least severe halo effects, e.g. only extremely low ratings, are non-existent.

Even "objective" measures often contain problems. In this study the chosen performance measures represent the primary information that pilots and other stakeholders are interested in, i.e. number of hits and survival. For the hits variable there is a potential ceiling effect due to the fact that there only was four enemy fighters to shot down in the defensive scenario. Thus, the most realistic engagement scenario, i.e. fourship versus a fourship contains a ceiling effect that is hard to avoid.

Properties of data
Issue: The subjective questionnaire items produce data with ordinal scale level properties. Why have they been analyzed as having interval scale level properties? Also, some of the variables do not fulfill suggested criteria for a fully normal distribution.

Answer: With concern to the subjective questions, see Sjöberg's (1996) comments in chapter 1.8. As the number of categories in the subjective questions was seven and the data shows normal distribution this is not a severe issue. The non-normality is an issue, especially for the technically logged measures. However, correlational statistics are rather robust with regard to "reasonable" non-normality.

Model fit
Issue: The models show higher than recommended values for the RMSEA fit index. The chi-square is high and the models become significant.

Answer: The RMSEA of the final model lie above the 0.05 threshold, while according to many SEM manuals it should be under the threshold to indicate a good fit. However, the model is still under the 0.08 recommendation found in some sources. Also both the chi-square and the RMSEA are sensitive to sample size. Several other fit statistics, such as the GFI and the CFI, indicate good fit.

Model complexity
Issue: Why is a simplex structure used when there are several statistically significant and strong factor intercorrelations in the dataset?

Answer: The structure of a model must always be chosen with regard to the purpose of the model. A rather large number of statistically significant factor intercorrelations exist in the data, but the simplex structure was chosen a) because it can handle the strong and direct factor intercorrelations in terms of indirect effects, b) in order to present a model that could be used in support of a procurement decision. The search in this modeling effort was not to necessarily find the simplest model, but to find a simple model where all the chosen constructs could be integrated. Through this, the model retains diagnosticity, even while it is easy to grasp conceptually.

One of many models
Issue: A number of statistically almost equivalent models can usually be expressed, based on the same data.

Answer: The measurement model provides a number of "building blocks" that can be arranged in several meaningful ways, depending upon model purpose. If more relations are allowed in the model, model fit typically increase, but the model become harder to grasp. It should also be noted that the statistical machinery effectively stops many models. During the modeling effort several alterative models have been tested that have been rejected due to bad model fit.

Usefulness in a SBA process
Issue: Does models of this kind really contribute with any information usable in a procurement decision? Do they contribute with something more than "common sense"?

Answer: In all the models presented above the whole database have been used or week-number have been used to separate the datasets, given that the purpose was to describe how the constructs relate to each other. For practical use in a SBA study, where the purpose is to make conclusions regarding sensor alternatives, the questions arises how parameter estimates differs between sensor alternatives rather than weeks, as each week contain two or three sensor alternatives. The pursuit of these types of questions quickly makes conclusions and models secret. When the proposed model is tried on subsets of the database with the same sensor alternative, parameter estimates similar to the ones presented appear. SEM software also allows for multigroup comparisons and latent mean structure analysis, but it will not be presented here.

Comparison with other models
Issue: What's new with this model?

Answer: The author has seen few other statistical models, SEM or otherwise, which include all these constructs. Other conceptual models where all or some of the constructs are included exist, see for example Figure 7, but they are conceptual models without any attempt to quantify relations. In comparison with previous models by Svensson et al. (1999), the teamwork construct has been included. A simplex structure was chosen in

order to make the model even easier to grasp, and perhaps more usable in a procurement decision. Also, both the start and the end of the model are anchored in technical system-oriented, rather undebatable variables.

Relation to cognitive science
Issue: What's the relation between the constructs used here and "real" cognitive science?

Answer: The constructs used, for example situation awareness and mental workload, lies at the heart of the intersection between basic cognitive psychology and applied Human Factors. Although the essence of, for example, the mental workload construct can be linked to working memory limitations (Baddeley, 1986; 1992), the position of the author is that this issue relates to which level of analysis that is appropriate. Compare with Marr's (1982) well known description of three distinct levels in analysis of information-processing systems. Marr distinguishes between: a) a computational level, at which a system's goal is described, b) an algorithmic level, at which a system's method is described, and c) an implementational level, at which a system's means are described. Where basic cognitive psychology often seeks explanations on the algorithmic, e.g. cognitive processes, or implementational level, e.g. neural structures, Human Factors researcher often seek explanations on the computational level, e.g. mental representations.

One of Marr's points was that all levels of description should be taken equally seriously, to eventually arrive at a comprehensive theory, consisting of three complementary descriptions which together explain "how the goal is reached with a method that is allowed by the means of the system".

Generalization of model
Issue: What is the generalizability of the proposed model?

Answer: In terms of generalizability to a larger population, the sample of pilots that participated represents a fair percentage of the total Swedish fighter pilot population and generalizability is very high. The generalizability to the larger population of fighter pilots of the world should also be high, and a similar model structure probably appears.

In terms of generalizability to other air combat situations, the database includes data from engagements with varied preconditions for the pilots (different sensor alternatives, missiles, existence of fighter controllers, and scenarios). This is a strength of the database, and support claims that the model covers the variation in the situation that "reality" represents, at least for BVR scenarios.

The question whether the model is generalizable to real flight and real engagements is of course very interesting. Hopefully, we will never be able to collect this type of dataset. It is also not the purpose of a model in a SBA study. The procurement and design decisions have to be made years before a real incident, and the model, based on data from simulation really needs to be a predictive tool and provide basis for a "best guess". This is the envisioned world problem in a nutshell.

The generalizability to other work domains is a more open question. The basic methodological approach is generic, but the relevant constructs and relations between constructs might be different. This is an empirical issue, but similar Human Factors problems, job environments, and job task characteristics exist in any vehicle or platform driving situation, in control rooms, et cetera, where Human Factors researchers and practitioners typically work. If more data-based statistical models were developed and compared, it might well be the case that many experimental studies could provide additional scientific value and further contribute to the much needed theory infusion mentioned by Salas (2008).

5 Conclusions

The thesis presents several statistical models describing how the functional state of the operators or pilots mediate the effect between sensor effectiveness in one end of the model and the "objective" outcome variables of hits and survival in the other end, i.e. a "model of the operator" placed in the context of technical parameters. Human performance data describing approximately 700 hours of experienced fighter pilots' work in a complex and realistic context are thus captured and summarized in the models. The constructs' effects are strong, and they explain each other to an unusually high degree.

Several different types of conclusions can be drawn from the modeling effort.

5.1 Empirical conclusions

It has been said that nothing is more practical than a good model or theory. Through the modeling effort of the thesis, it was possible to reduce 24 manifest variables to seven constructs with their six effects. Thus, a substantial data reduction has been performed, and a larger set of data is summarized in a simple model that is easy to overview. Accordingly, a main characteristic of a practicable model is achieved. The model gives a simplified explanation of a complex situation and we can see how, and to what extent sensor data are transformed into information, how this information is handled by the operators, transformed to awareness of the situation, which relate to teamwork and the final outcome of the process.

The degrees of freedom and practical possibilities for a Human Factors researcher to collect data were within expectations when working within a real simulation based acquisition study. Full experimental control over the participating pilots/SME's can not be expected. It is hard to separate between the different sources of variance in the data, and methods to assess the impact of this needs to be developed.

5.2 Methodological conclusions

Structural equation modeling is a powerful statistical tool and is possible to use the method in this type of studies. The resulting statistical models provide an efficient and convenient way of describing the latent structure underlying a set of manifest variables. Expressed either graphically or mathematically via a set of equations, the models explain how the manifest and latent constructs are related to one another. An important feature of the methodology is that it makes it possible to draw conclusions from quasi- or non-experimental settings, where classical experimental control and manipulation of independent variables are circumscribed.

The multi-dimensional constructs are distinguishable from each other through rather straightforward operationalization, i.e. a number of subjective ratings. This is probably an

effect of the fact that the ratings have been developed and tested iteratively in close contact with the pilots and relates to concepts they talk about themselves.

More of the "objective" measures could be valuable, but the most relevant new measure would probably be a way to quantify how different tactics and the amount of risk taking chosen by the two sides result in intensity of the engagement.

5.3 Practical conclusions

The practical application of the models in the SBA study was not elaborated in the thesis due to the sensitive nature of conclusions. But it is possible to develop similar models to the ones presented above that describe differences between the sensor alternatives.

For team training (e.g. training in the FLSC simulator facility, or other team training situations in a number of domains), models of the kind that have been described here are also important. Feedback is a central aspect in all training and in the development of adaptive expertise, i.e. skills or expertise useful even in non-routine situations (Holyoak, 1991). Freeman, Salter, & Hoch (2004) describe how feedback during debriefs is particularly important in team training, since the teamwork itself does not necessarily produce the immediate feedback from which the team members can learn. Models including teamwork are thus especially important. Ericsson (e.g. Ericsson & Charness; 1994, Ericsson, 2002; 2006) has clarified the need for deliberate practice in the development of expertise for a number of domains of human activity. Feedback and deliberate practice go hand in hand. Direct, almost instantaneous, feedback would be necessary in order for models of this type to support deliberate practice (cf. Angelborg-Thanderz, 1990). Provided that a model could be specified based on earlier experiences and models, e.g. the ones of this thesis, this would be possible if data was entered in a digitized form. Visualization of Human Factors such as those chosen in the current thesis could lead to improved training value.

5.4 Theoretical conclusions

All the constructs used in the thesis showed their worth and justification. Theoretical discussions whether these constructs represent the same thing, are not supported by the data. If they are seen as being "constructs on a string", as expressed by the simplex structure, where each constructs to a rather large extent explains the preceding and succeeding constructs, further understanding of the whole system can be achieved.

The statement implicit in the simplex model represent a rather strong theoretical claim, to be challenged!

6 References

Aagard-Nielsen, K., & Svensson, L. (Eds.). (2006). *Action and Interactive Research – Beyond practice and theory*. Maastricht: Shaker Publishing.

Ahl, V. & Allen, T.F.H. (1996). *Hierarchy Theory: A Vision, Vocabulary and Epistemology*. New York, NY: University of Columbia Press.

Alberts, D., & Hayes, R. (2002). Code of best practice: Experimentation. Washington: CCRP Publication Series.

Alehagen, U., Svensson, E., & Dahlström, U. (2007). Natriuretic peptide biomarkers as information indicators in elderly patients with possible heart failure followed over 6 years: A head to head comparison of 4 cardiac natriuretic peptides. *Journal of Cardiac Failure,* 13 (6).

Alfredson, J., Nählinder, S., & Castor, M. (2004). *Measuring eye movements in applied psychological research – five different techniques – five different approaches* (*FOI-R--1406--SE)*. Linköping: Swedish Defence Research Agency.

Alfredson, J., Oskarsson, P-A., Castor, M., & Svensson, J. (2003). Development of a meta instrument for evaluation of man-system interaction in systems engineering. *Proceedings of NES 2003, 35th Annual conference of the Nordic Ergonomics Society*, Reykjavik: Nordic Ergonomics Society.

Alfredson, J., Oskarsson, P-A., Castor, M., & Svensson, J. (2004). *Metodvalsverktyg - Ett hjälpmedel vid planering av MSI-utvärdering* (FOI-R--1295--SE) [Instrument for choice of methods - A means of assistance in planning of a man-system interaction evaluation]. Linköping: Swedish Defence Research Agency.

Alfredson, J. (2007). Differences in Situational Awareness and How to Manage Them in Development of Complex Systems (Linköping Studies in Science and Technology Dissertation No. 1132). Linköping: University of Linköping.

Anderson, J.R. (2002). Spanning seven orders of magnitude: a challenge for cognitive modelling. *Cognitive Science*, 26 (1), 85-112.

Angelborg-Thanderz, M. (1982). Assessing pilot performance and mental workload in training simulators. *Proceedings of Flight simulation, aviation medicine, and avionics system groups two day international conference*. London: The Royal Aeronautical Society.

Angelborg-Thanderz, M. (1989). Assessing pilot performance in training simulators. A structural analysis. *Proceedings of 1989 Spring Convention on flight simulation.* London: The Royal Aeronautical Society.

Angelborg-Thanderz, M. (1990). *Military flight training at a reasonable price and risk* (Doctoral dissertation no. 311). Stockholm: School of Economics. In Swedish.

Annett, J., & Stanton, N. (Eds.) (2000). *Task analysis*. London: Taylor & Francis.

ANSI/AIAA (1992). *Guide to Human Performance Measurement*. Washington, D.C.: AIAA.

Artman, H. (1999). *Fördelade kunskapsprocesser i ledningscentraler vid nödsituationer – koordination och situationsmedvetenhet* (Linköping Studies in Arts and Science Dissertation no 186) [Distributed Cognition in Control Rooms for Emergency Situations]. Linköping: University of Linköping.

Baddeley, A. (1986). *Working Memory*. Oxford: Clarendon Press.

Baddeley, A. (1992). Working memory. *Science*, 255, 556-559.

Bannon, L.J. (2001). Toward a Social and Societal Ergonomics: A Perspective from Computer-Supported Cooperative Work. In M. McNeese, E. Salas & M. Endsley (Eds.), *New Trends in Cooperative Activities: Understanding System Dynamics in Complex Environments*. Santa Monica, CA: Human Factors and Ergonomics Society.

Bennett, W.Jr., Lance, C.E., & Woehr, D.J. (Eds.) (2006). *Performance Measurement – Current perspectives and future challenges*. Mahwah, NJ: Lawrence Erlbaum.

Boomsma, A., & Hoogland, J.J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit & D. Sörbom (Eds.), *Structural equation models: Present and future. A Festschrift in honor of Karl Jöreskog*. Lincolnwood, IL: Scientific Software International.

Bollen, K. (1989). *Structural equations with latent variables.* New York: Wiley.

Brannick, M., Roach, R., & Salas, E. (1993). Understanding team performance: A multimethod study. *Human Performance, 6*, 287-308.

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta Psychologica*, 81, 211-241.

Burke, C.S., Salas, E., Wilson-Donnelly, K., & Priest, H. (2004). How to turn a team of experts into an expert medical team: guidance from the aviation and military communities. *Quality & Safety in Health Care*, 13 (1), 96-104.

Byrne, B.M. (1998). Structural equation modeling with LISREL, PRELIS, and SIMPLIS. Hillsdale, NJ: Lawrence Erlbaum.

Byrne, B.M. (2001). Structural Equation Modeling with AMOS. Mahwah, NJ: Lawrence Erlbaum.

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

Carlshamre, P. (2001). *A Usability Perspective on Requirements Engineering - From Methodology to Product Development* (Linköping Studies in Science and Technology Dissertation No. 726). University of Linköping.

Carroll, J.M. (1997). Human-Computer Interaction: Psychology as a Science of Design. *Annual Review of Psychology*, 48, 61-63.

Castor, M., Hanson, E., Svensson, E., Nählinder, S., LeBlaye, P., MacLeod, I., Wright, N., et al. (2003). *GARTEUR Handbook of Mental Workload Measurement*. Group of Aeronautical Research and Technology in Europe Technical Paper 145.

Coulter, K., Jones, R., Kenny, P., Koss, F., Laird, J., & Nielsen, P. (1999). Automated Intelligent Pilots for Combat Flight Simulation. *Proceedings of the Tenth Annual Conference on Innovative Applications of Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Dekker, S., & Hollnagel, E. (2004). Human factors and folk models. *Cognition, Technology and Work, 6*, 79–86.

Diamantopoulos, A., & Siguaw, J. (2000). *Introducing LISREL*. London: Sage Publications.

Doane, S.M., & Sohn, Y.W. (2000). ADAPT: A Predictive Cognitive Model of User Visual Attention and Action Planning. *User Modeling and User-Adapted Interaction* 10: 1-45. Kluwer Academic Publishers.

DMSO (1999). The Simulation Based Acquisition Vision. *Proceedings of the Army's Simulation & Modeling for Acquisition, Requirements and Training (SMART) conference*. Defence Modeling and Simulation Office.

Durso, F.T., & Sethumadhavan, A. (2008). Situation Awareness: Understanding Dynamic Environments. *Human Factors*, 50 (3)3, 442–448.

Ekman, G. (1954). Dimensions of color vision. *Journal of Psychology*, 38, 467-474.

Endsley, M.R. (1993a). Situation awareness and workload: Flip sides of the same coin? *Proceedings of the 7th International Symposium on Aviation Psychology*, Columbus, OH: Ohio State University.

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

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.

Endsley, M.R. (2000). Theoretical underpinnings of situation awareness: A critical review. In M.R. Endsley & D.J. Garland (Eds.), *Situation Awareness Analysis and Measurement*. Mahwah, NJ: Lawrence Erlbaum.

Engeström, Y., Miettinen, R., & Punamäki, R.L. (Eds.) (1999). *Perspectives on Activity Theory*. Cambridge: Cambridge University Press.

Ericsson, K.A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, *49* (8), 725-747.

Ericsson, K.A. (2002). Attaining excellence through deliberate practice: Insights from the study of expert performance. In M. Ferrari (Ed.), *The pursuit of excellence in education*. 21-55. Hillsdale, N.J.: Erlbaum.

Ericsson, K.A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. Feltovich, & R. R. Hoffman, R. R. (Eds.). *Cambridge handbook of expertise and expert performance*. 685-706. Cambridge: Cambridge University Press.

Fracker, M.L. (1988). A theory of situation assessment: Implication for measuring situation awareness. *Proceedings of the Human Factors Society 32nd Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.

Freeman, J., Salter, W.J., & Hoch, S. (2004). The users and functions of debriefing in distributed, simulation-based team training. *Proceedings of the Human Factors Society 48th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.

Glickman, A.S., Zimmer, S., Montero, R.C., Guerette, P.J., Campbell, W.J., Morgan, B.B. & Salas, E. (1987). *The evolution of teamwork skills: An empirical assessment with implications for training* (Tech. Rep. 87-016) Orlando, FL: Naval Training Systems Center.

Gunzelmann, G., Gluck, K.A., Price, S.C., Van Dongen, H.P.A., & Dinges, D.F. (2007). Decreased arousal as a result of sleep deprivation: The unraveling of cognitive control. In W.D. Gray (Ed.), *Integrated Models of Cognitive Systems*. New York, NY: Oxford University Press.

Guttman, L.A. (1954). A new approach to factor analysis: the radix. In P. F. Lazarsfield (Ed.) *Mathematical thinking in the social sciences*. New York, NY: Columbia University Press.

Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C. (1995). *Multivariate Data Analysis with readings*. Englewood Cliffs, NJ: Prentice-Hall.

Hart, S.G., & Staveland, L. E. (1988). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam: Elsevier.

Harris, R.M., Hill, S.G., Lysaght, R.J., & Christ, R.E. (1992). *Handbook for Operating the OWLKNEST Technology* (ARI Research Note 92-49). U.S. Army Research Institute.

Harré, R. (2002). *Cognitive Science: A Philosophical Introduction*. London: Sage.

Hollnagel, E., & Woods, D.D. (2005). *Joint cognitive systems: Foundations of cognitive systems engineering*. Boca Raton, FL: CRC Press / Taylor & Francis.

Holyoak, K.J. (1991). Symbolic Connectionism: Toward third-generation theories of expertise. In K. A. Ericsson & J. Smith (Eds.), *Toward a General Theory of Expertise: Prospects and Limits*. 301-335. Cambridge, UK: Cambridge University Press.

Hoyle, R.H. (Ed.) 1995. *Structural Equation Modeling*. Thousand Oaks, CA: Sage.

Hutchins, E. (1995). *Cognition in the Wild.* Cambridge, MA: MIT Press.

Jaccard, J., & Wan, C.K. (1996). *LISREL approaches to interaction effects in multiple regression.* Newbury Park, CA: Sage.

Johannsen, G., Moray, N., Pew, R., Rasmussen, J., Sanders, A., & Wickens, C. (1977). Final report of experimental psychology group. In N. Moray (Ed.) Mental workload: Its theory and measurement. New York, NY: Plenum Press.

Johansson, C. (1999). Modellering av luftstridsavdömningar [Air Combat Resolution Modeling] (LiTH-ISY-EX-2049). Linköping: Linköping Institute of Technology.

Just, M.A., Carpenter, P.A., & Miyake A. (2003). Neuroindices of cognitive workload: neuroimaging, pupillometric and event-related potential studies of brain work. *Theoretical Issues in Ergonomics Science*, 4 (1-2), 56-88. London: Taylor and Francis.

Jöreskog, K.G. (1973). A general method for estimating a linear structural equation system. In A.S. Goldberger & O.D. Duncan (Eds.), *Structural equation models in the social sciences*. New York: Seminar Press/Harcourt Brace.

Jöreskog, K.G., & Sörbom, D. (1984). *Analysis of linear structural relationships by maximum likelyhood, instrumental variables, and least squares methods*. Uppsala: University of Uppsala.

Jöreskog, K.G., & Sörbom, D. (1993). *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Chicago, IL: Scientific Software International.

Kantowitz, B.H. (2000). Attention and Mental Workload. *Proceedings of IEA 2000/HFES 2000 Congress.* Santa Monica, CA: Human Factors and Ergonomics Society.

Klein, G. (1989). Recognition-Primed Decisions. In W.B. Rouse (Ed.) *Advances in man-machine system research*. 5, 47-92. Greenwich, CT: JAI Press.

Klein, G., Orasanu, J., Calderwood, R., & Zsambok, C. (Eds.) (1993). *Decision making in action: Models and methods*. Norwood, NJ: Ablex Publishing.

Klein, G., Moon, B., & Hoffman, R.R. (2006). Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems,* 21, 70–73.

Lachman, R., Lachman, J.L., & Butterfield, E.C. (1979). *Cognitive psychology and information processing*. Hillsdale, NJ: Lawrence Erlbaum.

Laird, J.E., Coulter, K.J., Jones, R.M., Kenny, P.G., Koss, F., & Nielsen, P.E. (1998). Integrating intelligent computer generated forces in distributed simulations: TacAir-Soar in STOW-97. *Proceedings of the 1998 Spring Simulation Interoperability Workshop*. Orlando, FL: SISO.

Locke, E.A., & Latham, G.P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice Hall.

Lotens, W., Allender, L., Armstrong, J., Belyavin, A., Cain, B., Castor, M., et al. (in press) *Human Behavior Representation in Constructive Simulation*. Paris: NATO Research and Technology Organisation.

Lützhöft, M. (2004). *"The technology is great when it works" - Maritime Technology and Human Integration on the Ship's Bridge* (Linköping Studies in Science and Technology Dissertation No. 907). Linköping: University of Linköping.

Lysaght, R.J, Hill S.G., Dick A.O., Plamondon B.D., Linton P.M., Wierwille W.W, et al (1989). *Operator workload: comprehensive review and evaluation of operator workload methodologies* (ARI Technical Report No. 851). Fort Bliss, TX: U.S. Army Research Institute.

McMillan, G.R., Bushman, J., & C.L.A., Judge (1996). Evaluating Pilot Situational Awareness in an Operational Environment. In *Situational Awareness: Limitations and Enhancement in the Aviation Environment* (AGARD-CP-575). Paris: NATO Research and Technology Organisation.

Magnusson, S. (2002). On the similarities and differences in psychophysiological reactions between simulated and real air-to-ground missions. *International Journal of Aviation Psychology*, 12 (1). 49-61.

Marr, D. (1982). *Vision: A Computational Approach*. San Francisco: Freeman.

Mathieu, J.E., Marks, M.A. & Zaccaro, S.J. (2001). Multi-team systems. In N. Anderson, D. Ones, H. Sinangil, & C. Viswesvaran, C. (Eds.), *International Handbook of Work and Organizational Psychology*. 289-313. London: Sage

McGuinness, B. (1996). Situational awareness measurement in cockpit evaluation trials. In *Situation Awareness: Limitations and Enhancement in the Aviation Environment* (AGARD-CP-575). Paris: NATO Research and Technology Organisation.

Moray, N. (Ed.) (1979). *Mental workload: Its Theory and Measurement*. New York: Plenum.

MVSOFT (2005). EQS homepage. http://www.mvsoft.com. Visited 2008-01-10.

Newell, A. (1973). You can't play 20 questions with nature and win. In W.G. Chase (Ed.), *Visual Information Processing*. New York, NY: Academic Press.

Newell, A. (1990). *Unified theories of cognition. Cambridge, MA:* Harvard University Press.

Nofi, A.A. (2000). *Defining and Measuring Shared Situational Awareness* (CRM DOOO2895A1). Alexandria, VA: Center for Naval Analyses.

Nählinder, S., Berggren, P., & Svensson, E. (2004). Reoccurring LISREL patterns describing mental workload, situation awareness, and performance. *Proceedings of Human Factors and Ergonomics Society 48th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.

Nählinder, S. (2009). Flight Simulator Training: Assessing the Potential (Linköping Studies in Science and Technology Dissertation No. 1250). Linköping: University of Linköping.

O'Donnell, R.D., & Eggemeier, F.T. (1986). Workload assessment methodology. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.) *Handbook of perception and human performance*. New York, NY: John Wiley.

Patten, C.J.D. (2007). *Cognitive Workload and the Driver: Understanding the Effects of Cognitive Workload on Driving from a Human Information Processing Perspective*. Stockholm: University of Stockholm.

Pew, R.W., & Mavor, A.S. (Eds.) (1998). *Modeling Human and Organizational Behavior Application to Military Simulations.* Washington, D.C.: National Academy Press.

Rehmann, A.J. (1995). *Handbook of Human Performance Measures and Crew Requirements for Flightdeck Research* (DOT/FAA/CT-TN95/49). Atlantic City, NJ: Federal Aviation Administration.

Roscoe, A. H. (1987). In-flight assessment of workload using pilot ratings and heart rate. In A. Roscoe (Ed.) *Practical assessment of pilot workload* (AGARD Agardograph No 282). Paris: NATO Research and Technology Organisation.

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

Salas, E., Dickinson, T.L., Converse, S.A., & Tannenbaum, S.I. (1992). Toward an understanding of team performance and training. In R. W. Swezey & E. Salas (Eds.), *Teams: Their training and performance.* Norwood, NJ: Ablex.

Salas, E. (2008). At the Turn of the 21st Century: Reflections on Our Science. *Human Factors*, 50 (3), 351–353.

Schumacker, R.E., & Lomax, R.G. (2004). *A beginner's guide to structural equation modeling, 2nd edition*. Mahwah, NJ: Lawrence Erlbaum.

Stout, J., Salas, E., & Fowlkes, J.E. (1997). Enhancing teamwork in complex environments through team training. *Journal of Group Psychotherapy, Psychodrama and Sociometry*, 49(4), 163-187.

Sanders, P. (1999). Simulation Based Acquisition – The Revolution is coming. *Army RD&A*, 8-10.

Simon, H. (1996). *Models of my life.* Cambridge, MA: MIT Press.

Sjöberg, L. (2006). Worry about scale levels is alive and well, but should it be? Department of Psychology, University of Trondheim.

SPSS (2008). AMOS homepage. http://www.spss.com/amos. Visited 2008-05-10.

SSI (2008). LISREL homepage. http://www.ssicentral.com/lisrel/index.html. Visited 2008-05-10.

Stanton, N.A., & Stammers, R.B. (2002). Creative (dis)agreement in ergonomics. *Ergonomics*, 45(14), 963-965.

Svensson, E. (1978). *Mood: its structure and measurement* (Göteborg Psychological Reports Dissertation No 6). Göteborg: University of Göteborg.

Svensson, E., Angelborg-Thanderz M., Sjöberg, L., & Gillberg, M. (1988). Military flight experience and sympatho-adrenal activity. *Aviation, Space, and Environmental Medicine*, 59, 411-416.

Svensson, E., Angelborg-Thanderz, M., & Sjöberg, L. (1993a). Mission challenge, mental workload and performance in military aviation. *Aviation, Space, and Environmental Medicine*, 64, 985-991.

Svensson, E., Angelborg-Thanderz, M., & Sjöberg, L. (1993b). Information overflow? Mental workload and performance in combat aircraft. *Proceedings of Workload Assessment and Aviation Safety*. London: The Royal Aeronautical Society.

Svensson, E., & Angelborg-Thanderz, M. (1995). Mental workload and performance in combat aircraft: systems evaluation. In R. Fuller, N. Johnston, & N. McDonald (Eds.), *Human Factors in Aviation Operations*. Aldershot: Avebury aviation, Ashgate Publishing.

Svensson, E., Angelborg-Thanderz, M., Sjöberg, L., & Olsson, S. (1997). Information complexity - mental workload and performance in combat aircraft. *Ergonomics*, 40 (3), 362-380.

Svensson, E. (1997). Pilot mental workload and situational awareness – psychological models of the pilot. In R. Flin, M. Salas, M. Strub, & L. Martin (Eds.), *Decision making under stress: Emerging themes and applications*. 261-267. Aldershot: Ashgate.

Svensson, E., Angelborg-Thanderz, M., & van Awermaete, J. (1997). *Dynamic measures of pilot mental workload, pilot performance, and situational awareness* (VINTHEC-WP3-TR01). Amsterdam: NLR.

Svensson, E., Angelborg-Thanderz, M., & Wilson, G. (1999). *Models of pilot performance for systems and mission evaluation - psychological and psychophysiological aspects* (AFRL-HE-WP-TR-1999-0215*)*. Dayton: U.S. Airforce Research Lab.

Svensson, E., & Wilson, G. (2002). Psychological and psychophysiological models of pilot performance for systems development and mission evaluation. *International Journal of Aviation Psychology*. 12 (1), 95-110.

Svensson, E., Rencrantz, C., Lindoff, J., Berggren, P., & Norlander, A. (2006). Dynamic measures for performance assessment in complex environments. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.

Svensson, E., Lindoff J., & Sutton, J. (2008). Predictive modelling of personality traits – Implications for selection of operational personell (NATO RTO MP-HFM-142-06). Paris: NATO Research and Technology Organisation.

Svensson, E., Angelborg-Thanderz, M., Borgvall, J., & Castor, M. (in press). Skill Decay, Re-Acquisition Training, and Transfer Studies in the Swedish Air Force: A Retrospective Review. In W. Bennett & W. Arthur (Eds.), Tentative book name: *Skill decay and retention.* Taylor Francis or Ashgate.

Swedish Armed Forces Headquarters (2005). Doktrin för luftoperationer [Air operations doctrine]. Stockholm: Swedish Armed Forces.

Tabachnick, B.G., & Fidell, L.S. (1996). *Using multivariate statistics* (3rd ed.). New York, NY: Harper Collins.

Thompson, B. (2000). Ten commandments of structural equation modeling. In L. Grimm & P. Yarnell (Eds.), *Reading and understanding more multivariate statistics*. Washington, D.C.: American Psychological Association.

Thurstone, L.L. (1932). *The vectors of mind.* Chigago, IL: University of Chigago Press.

Veltman, J.A., Gaillard, A.W.K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, 41 (5), 656-669.

Vicente, K. J. (1999). *Cognitive Work Analysis: Towards Safe, Productive, and Healthy Computer-based Work*. Mahwah, NJ: Erlbaum.

Vidulich, M.A. (2003). Mental Workload and Situation Awareness: Essential Concepts for aviation psychology practice. In P.S Tsang & M. A Vidulich (Eds.), *Principles and Practice of Aviation Psychology*. Mahwah, NJ: Lawrence Erlbaum.

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

Weick, K.E. (1995). *Sensemaking in organizations*. Thousand Oaks, CA: Sage.

Westlander, G. (1999). *People at Work – Investigating the social-psychological contexts*. Lund: Studentlitteratur.

Wickens, C.D. (1984). Engineering psychology and human performance. New York, NY: Harper Collins.

Wickens, C.D. (2008). Multiple Resources and Mental Workload. *Human Factors*, 50 (3), 449-455.

Wierwille, W.W., & Casali, J. (1983). Validated rating scale for global mental workload measurement applications. *Proceedings of the Human Factors Society 27th Annual Meeting.* Santa Monica, CA: Human Factors and Ergonomics Society.

Wilkinson, L. (1990). *SYSTAT: The System for Statistics*. Evanston, IL: SYSTAT.

Wilson, G.F. (1993). Air-to-ground training missions - a psychophysiological workload analysis. *Ergonomics,* 36 (9), 1071-1087.

Wilson, G.F., Fullenkamp, P., & Davis, I. (1994). Evoked-potential, cardiac, blink, and respiration measures of pilot workload in air-to-ground missions. *Aviation Space Environmental Medicine*, 65 (2), 100-105.

Wilson, G.F., Balkin, T., Beamont, M., Burov, A., Edgar, G., Fraser, W., et al. (2004). *Operator Functional State Assessment* (RTO-TR-HFM-104). Paris: NATO Research and Technology Organisation.

Wikipedia (2008). http://en.wikipedia.org/wiki/AIM-120_AMRAAM. Visited 3rd January 2008.

Wikipedia (2008). http://en.wikipedia.org/wiki/BVR. Visited 3rd January 2008.

Woods, D., Christoffersen, K., & Tinapple, D. (2000). Complementarity and synchronization as strategies for practice-centered research and design. Plenary address, 44th Annual Meeting of the Human Factors and Ergonomics Society and International Ergonomic Association, 1 August 2000.

Woods, D., & Decker, S. (2000). Anticipating the effects of technological change: a new era of dynamics for human factors. *Theoretical Issues in Ergonomics Science*, 1(3), 272-282.

Wright, S. (1918). On the Nature of Size Factors. *Genetics,* 3, 367-74.

Zsambok, C.E. (1997). Naturalistic decision making: Where are we now? In C. Zsambok & G. Klein (Eds.), *Naturalistic decision making*. Mahwah, NJ: Lawrence Erlbaum.

7 Appendices

7.1 Appendix 1. SIMPLIS command files

```
Raw Data from file 'C:\Documents and
Settings\Administrator\Desktop\thesis\thesis-n1232.psf'

Sample Size = 1232
Latent Variables SENSOR INFO MWL SA TEAM OFFPERF DEFPERF
Relationships

! Definition of the structural model
INFO = SENSOR
MWL = INFO
SA = MWL
TEAM = SA
OFFPERF = TEAM
DEFPERF = TEAM

! Definition of the measurement model
RADARCO7 = SENSOR
SENSOR1 = SENSOR
SENSOR2 = SENSOR
SENSOR3 = SENSOR
INFO1 = INFO
INFO2 = INFO
INFO3 = INFO
INFO4 = INFO
MWL1 = MWL
MWL2 = MWL
MWL3BED = MWL
MWL4 = MWL
SA1 = SA
SA2 = SA
SA3 = SA
SA4 = SA
TEAM1 = TEAM
TEAM2 = TEAM
TEAM3 = TEAM
TEAM4 = TEAM
HITS7 = OFFPERF
THITSMO7 = OFFPERF
SURV7 = DEFPERF
TSURVMO7 = DEFPERF

! Results of model respecification
Set the Error Covariance of INFO1 and INFO2 Free
Set the Error Covariance of INFO3 and INFO4 Free
Set the Error Covariance of MWL3BED and MWL4 Free
Set the Error Covariance of SA2 and SA3 Free
Set the Error Covariance of TEAM2 and TEAM3 Free

Path Diagram
```

7.2 Appendix 2. Data collection instrument

Translated from Swedish and sorted for thesis. Manifest variable names added in parenthesis.

ENGAGEMENT NUMBER…….. DATE…….. TIME……

PILOT SIGNATURE …….
FLSC PILOT STATION NUMBER……
NUMBER IN FORMATION…… (1 = FOURSHIP LEAD)

The questions below cover different aspects that are important when assessing workload, performance, system functions and teamwork. From your answers we can analyze requirements of missions and systems along with the pilots' possibility to meet the requirements. Difficulty and workload naturally varies during the engagements so please answer the question with regard to the **highest workload** you encountered during the engagement. Mark the number that seems most fitting in order to reflect your experience from this engagement.

1. To what extent was it possible to use the full capacity of the radar? (SENSOR1)

   Not at all 1 2 3 4 5 6 7 Completely

2. To what extent was it possible to keep satisfactory radar contact with "your" targets? (SENSOR2)

   Not at all 1 2 3 4 5 6 7 Completely

3. Did you experience any problems using the aircraft radar? (SENSOR3)

   Not at all 1 2 3 4 5 6 7 Very large

4. To what extent was it possible to survey the information on the Target Indicator? (INFO1)

   Not at all 1 2 3 4 5 6 7 Completely

5. To what extent was it possible to use the information on the Target Indicator effectively? (INFO2)

   Not at all 1 2 3 4 5 6 7 Completely

6. To what extent was it possible to survey the information on the Tactical Indicator? (INFO3)

   Not at all 1 2 3 4 5 6 7 Completely

7. To what extent was it possible to use the information on the Tactical Indicator effectively? (INFO4)

   Not at all 1 2 3 4 5 6 7 Completely

8. To what extent did you have situation awareness, with regard to the whole situation? (SA1)

   Very small 1 2 3 4 5 6 7 Completely

9. To what extent were you surprised by the positions of enemy aircraft? (SA2)

   Never 1 2 3 4 5 6 7 Very often

10. To what extent were you surprised by the behavior of enemy aircraft? (SA3)

    Never 1 2 3 4 5 6 7 Very often

11. To what extent did you perceive that you could predict the course of events? (SA4)

    Not at all 1 2 3 4 5 6 7 All the time

12. To what extent could you follow the original plan? (TEAM1)

    Not at all 1 2 3 4 5 6 7 All the time

13. How was the teamwork of the unit? (TEAM2)

    Not good at all 1 2 3 4 5 6 7 Best possible

14. To what extent did you perceive that you could predict the behavior of the other pilots in the unit? (TEAM3)

    Not at all 1 2 3 4 5 6 7 Best possible

15. How long time did it take for the unit to “straighten out” situations where something had become confused? (TEAM4)

Much too long 1 2 3 4 5 6 7 Very quickly

16. How hard do you think the engagement was? (DIFFIC1)

Very easy 1 2 3 4 5 6 7 Very hard

17. How challenging was the engagement? (DIFFIC2)

Not at all 1 2 3 4 5 6 7 Very challenging

18. How dangerous did you perceive the threat to be? (DIFFIC3)

Not at all 1 2 3 4 5 6 7 Very dangerous

19. How large risks to you take during the engagement? (DIFFIC4)

Very small 1 2 3 4 5 6 7 Very large

20. Was it hard to manage all tasks? (MWL1)

Not at all 1 2 3 4 5 6 7 Very hard

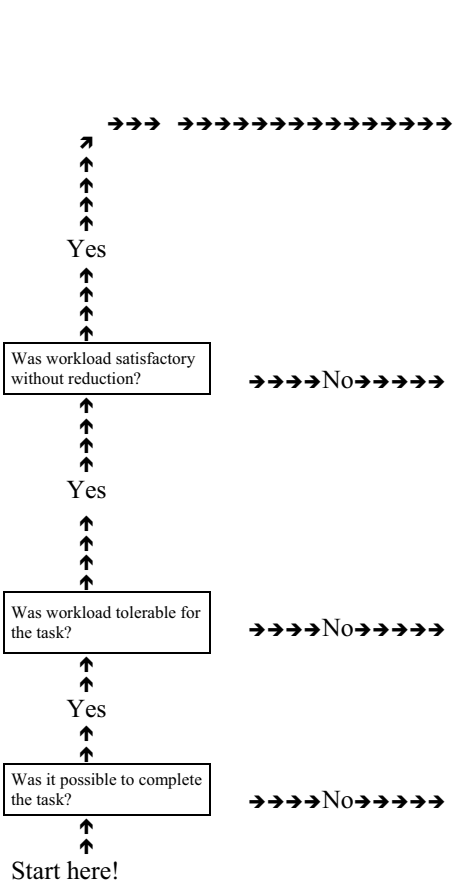
21. Was it possible to have “mental lead time” with respect to your tasks? (MWL2)

Not at all 1 2 3 4 5 6 7 Definitely

22. To what extent was mental workload an obstacle to optimal performance? (MWL4)?

Not at all 1 2 3 4 5 6 7 Completely

23. How large was your mental workload?(MWL3Bed)



Workload description	Rating
Workload insignificant.	1
Workload low.	2
Enough spare capacity for all desirable additional tasks	3
Insufficient spare capacity for easy attention to additional tasks.	4
Reduced spare capacity, additional tasks cannot be given the desired amount of attention.	5
Little spare capacity, level of effort allows little attention to additional tasks.	6
Very little spare capacity, but maintenance of effort in the primary task not in question.	7
Very high workload with almost no spare capacity. Difficulty in maintaining level of effort.	8
Extremely high workload. No spare capacity. Serious doubts as to ability to maintain level of effort.	9
Task abandoned. Pilot unable to apply sufficient effort.	10