Australasian Transport Research Forum 2011 Proceedings
28 - 30 September 2011, Adelaide, Australia
Publication website: http://www.patrec.org/atrf.aspx

Towards a Code of Best Practice for Evaluating Air Traffic Control Interfaces

Hussein A. Abbass[1,2], William M. Mount[1], Deborah Tuček[1], Jean-Philippe Pinheiro[3]

[1]Cognitive Engineering Laboratory, Australian Defence Force Academy, University of New South Wales, Canberra, ACT 2600

[2]Air Traffic Management Laboratory, Australian Defence Force Academy, University of New South Wales, Canberra, ACT 2600

[3]Human Factors, Thales Air Systems, 3, Av Charles Lindbergh - BP 20351, 94628 Rungis Cedex, France

Email for correspondence: h.abbass@adfa.edu.au

Abstract

The quality of computer interfaces in transportation command and control centres is vital to safe and smooth operations. Air Traffic Control (ATC) is probably the most dynamic area in transportation where a large amount of information is presented to the air traffic controller within a short timeframe. Future Air Traffic Interfaces (ATI) are on the horizon with more information and added levels of sophistication. Safety is becoming a default constraint in current systems and evaluating the usability of these interfaces has been seen traditionally as crucial for ensuring high operational safety standards. To this end, a strong business case for evaluating the usability of interfaces necessarily requires a full-scale justification of the usability study and its associated cost. The benefits of performing such an evaluation also need to be communicated to decision makers in terms of economic values and gains. It is at this point that the field of operational analysis intersects with human factor research.

This paper outlines a methodology for conducting usability studies for ATI. The methodology has been designed to connect higher-level organisational objectives with low-level usability metrics. The methodology will be presented towards establishing a code of best practice for the design and conduct of usability studies in this domain. While the results can be generalised to other transportation command and control interfaces, this paper focuses on ATC because this code of best practice is tailored towards ATC functions.

1. Introduction

The evaluation of air traffic control interfaces usually consists of a battery of heuristic, standards-based and usability assessments throughout the development phase. As a result, there are many standards developed for usability of hardware and software systems. Standards are written to be general and mostly focus on a single dimension of a complex problem. Standards such as ISO 9241-11:1998 and technical reports such as ISO/TR 16982:2002 and ISO/TR 9241-100:2010 cover both usability design and evaluation perspectives. The basic assumption is that criteria used for the evaluation stage can inform the design stage. While this is may be true in many systems, it is not the case for all.

To use an example specific to the context of this paper, safety-critical command and control systems (SC-C2S) are very large, built by multi-national companies and mostly exist in whole or part before the target market companies can begin tendering for them. While the technology companies can rely on user-centric design during the evolution of these systems and can customise them for different users, evaluating usability of these systems would normally be done by end-users that either have a different mind-set from the users or testers who were involved in the design, and/or in environments that were not fully anticipated during the design stage. Here, evaluating usability is no trivial matter.

Usability is defined by ISO as the "*extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context*

of use" (ISO 9241-11:1998), and such an evaluation is required by decision-makers to ensure the system will provide additional user and organisational benefits while maintaining or exceeding safety standards. Usability standards, however, tend to focus on the 'use' of a product rather than the 'impact' on the user. It may sound obvious that the impact of a product on a user will be reflected in how they use the product, and vice versa. However, in cases such as safety-critical command and control systems, one needs to understand the interrelationship of use and impact with a more in-depth analysis of the dynamic occurring between the two.

Another challenge that we are faced with in relying on standards – even those written specifically for interactive software systems and multi-media – is that critical aspects of SC-C2S, such as situational awareness, are not considered. Most usability studies in the software area are driven by marketing research and, as a result, focus on web and multi-media applications in less safety-critical systems. For example, the usability of a web page for personal banking is characterised by situations which are highly predictable, less dynamic, and occur within a relatively controlled environment. The focus of these studies would be on information layout, the type of content to be included, timing and display of that content as well as aesthetic or readability features such as colour, font type and size. All of these aspects are equally important in SC-C2S, but a usability study for an SC-C2S needs to go beyond the simple interface. It needs to take into account the user's cognitive processes, decision-making processes and the highly dynamic mental picture that is being constantly formed and updated within the user's mind.

We pose the question of how organisations can reliably evaluate interfaces for safety critical command and control systems for their specific needs and environments. While the proposed methodology is generic enough to be adapted for various SC-C2S, the domain knowledge and expertise underpinning this paper lies within the field of air traffic management (ATM); hence, the code will be presented from this perspective and we encourage others to deploy it within differing contexts.

The remainder of this paper will present the motivation underpinning the development of this code of best practice (COBP) to evaluate interfaces of SC-C2S, discuss the important challenges faced within this domain, and present the code using examples to help illustrate its implementation for the evaluation of an ATI.

2. The Need for a Code of Best Practice

The establishment of a COBP to evaluate SC-C2S interfaces was motivated by the following issues:

1. There is growing demand on acquiring new or upgrading existing safety-critical systems in defence and government organisations. Major transformations are occurring worldwide to provide better networking and upgraded command and control (C2) systems. Defence is leading the way with concepts of operation (CONOPs) such as Network Centric Warfare and Network Centric Operations (DOD-NCW, 2007), which are impacting on almost all C2 systems in Defence. In ATM, network-based operations have been a major area of discussion in both the Single European Sky ATM Research Programme (SESAR) (SESAR, 2011) and the Next Generation Air Transportation System (NEXTGEN) (FAA, 2007) in Europe and the USA, respectively. In Australia, the move to integrate the Civil-Military ATM systems necessitates a closer look at how to evaluate the interfaces of these systems to ensure their safe future operation. The lack of a COBP puts these organisations at risk of spending unnecessary resources. Moreover, there is a serious concern that some of these studies will be done in an ad-hoc manner, under time and resource constraints.

2. The area of ATM has witnessed a large number of Human Factors studies. Organisations such as Eurocontrol, FAA, NASA, and ICAO produced many reports establishing principles for Human Factors studies in ATM. Similar to the area of usability, many of these studies were either generic for both hardware and software, or focused on a specific test. One common thread in Human Factors studies is safety as the primary motivation for the conduct of such a study. It is natural to motivate such studies using a safety lens. However, to establish a business case for a human factors study from an organisational perspective, the objectives of the study need to be linked to the objectives of the organisation. Here, we propose a methodology to establish this link in a systematic manner.
3. Practitioners face real moral, legal and technical challenges when asked to sign off on the acceptance or purchase of an interface for a safety-critical system. Mostly, those who are accountable for accepting these interfaces are faced with the moral dilemma of whether or not they have performed sufficient due diligence in the testing of these interfaces. The concept of due diligence unfortunately relies heavily on an individual's knowledge and abilities at a particular time. Consequently, practitioners will try to use this knowledge and level of understanding to predict, anticipate and extrapolate many 'possible' scenarios or negative impacts in an attempt to secure against them. A COBP offers both practitioners and decision-makers greater peace of mind when it comes to shouldering the responsibility of accepting a new system into their organisation.

In developing a COBP, it is important that one realises the following:

1. A COBP represents a philosophy. While this philosophy can be challenged, if its premises are accepted, it provides a unified framework for policy makers to judge on the validity of a usability study of safety-critical systems.
2. A COBP is a compilation of lessons learned within a domain. As lessons evolve, the code itself needs to continue to evolve. A COBP should be seen as a best practice for our current level of knowledge. It defeats its own purposes if it is seen as a compilation of untouchable and unquestionable facts.
3. A COBP does not dictate a process to follow, but rather offers a set of guidelines and principles to ensure the quality of a process. Through a COBP, principles for designing experiments to evaluate interfaces in safety-critical systems are compiled using a scientifically rigorous approach.
4. A COBP promotes different type of experiments that can be conducted to evaluate an interface in safety-critical systems, its strengths and weaknesses, and use and misuse. The word 'practice' also emphasises what is 'doable'. It combines scientific rigour with practical feasibility.

We may argue that the following three generic steps would exist in any study. As such, we will structure the code around each of these steps.

1. Understanding the context

   This involves understanding the interface design principles, the organisation's context, the operator's duties and functions, and the operator's cognitive functions.

2. Designing the experiment

   This step looks at different factors that need to be considered during experimental design, including the design of the tasks that will be used for testing and the data collection plan. Many variables in a usability study are not easily measured. In many cases, we use indicators rather than direct measures. The concept of usability itself is not measurable, although some indicators can be used for it. This step connects the context established in step 1 with the experiments.

3. Analysing the results

   Usability studies can produce a large amount of data. Without a proper design for the analysis, significant resources can be devoted to collecting data that will not be used for the analysis. We refer the reader to the Code of Best Practice in Experimentation (Alberts, 2002) and the Technical Cooperation Committee Program (TTCP) publication on Experimentation (Bowley et. al., 2006) to cover a portion of the second step and this third step. The rest of this paper will focus on the first two steps.

3. Understanding the Context

3.1. Interface design principles

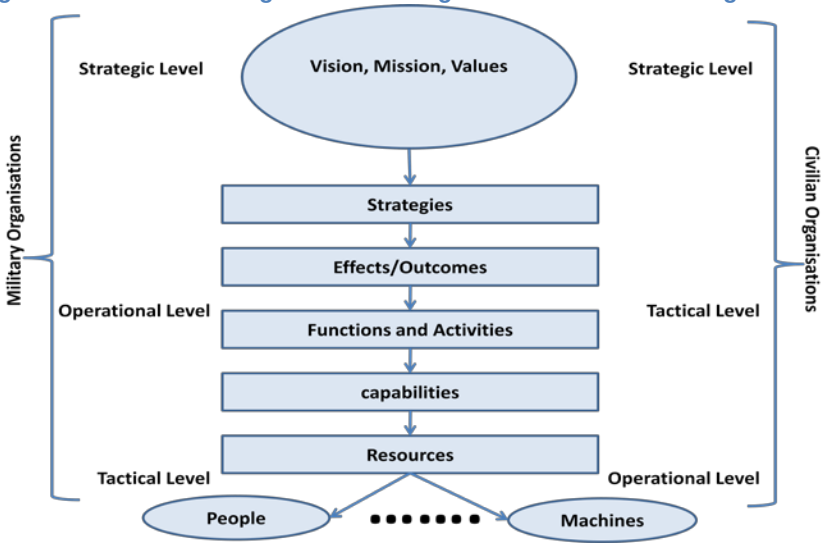
When evaluating an interface, it is always worthwhile understanding the design principles for the software at hand. Many recent developments in the software industry rely on user-centric design. Therefore, we can expect to find that representatives of the user community have guided the design of the software, and consequently the design of the interface during product development. Understanding the principles of the design can assist the usability analyst in developing appropriate metrics. It also helps during the experimental design phase.

It is important to identify those design principles that do not have a direct impact on the interface. Modularity of design and code reusability, for example, do not necessarily translate directly into tools or functionalities visible from the interface. The impact of such design principles is normally outside the scope of a usability study. While these design principles can bring direct benefits to the organisation in terms of increased reliability, reduced development cycle time and lower software maintenance costs, the design impacts require an economic evaluation rather than a usability study. Nevertheless, it is important to also become aware of these design principles since the usability study itself will form part of the overall evaluation report; a report that would include the economic benefits of the software.

3.2. Organisation's context

A business case for a study on usability or human factors needs to be situated within the wider context of an organisation. A study is commissioned by an organisation for a purpose, and unless the purpose is properly linked to organisational objectives, management is left with no logical justification to accept the study. The following schematic diagram links organisational functions together to clarify the role and importance of each function and the interdependencies among them.

Figure 1 A schematic diagram connecting different levels of an organisation.



It is assumed that a healthy organisation will have an overall vision, mission and a defined set of values. The vision captures the ambition of the organisation, the mission sets out its core purpose, achievable goals and objectives, and the values represent the principles defining an organisation's culture. In most cases, the vision, mission and values are formulated simultaneously. Strategic management transforms these into strategies.

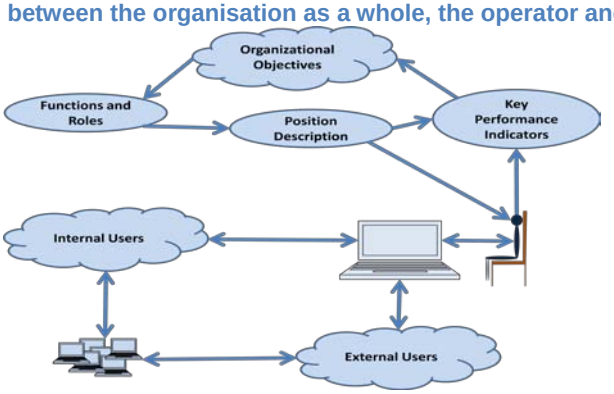
The executive level in an organisation transforms strategies into organisational capabilities. This is done through determining the organisation's desired outcomes, what functions and activities need to be done to generate or achieve these outcomes, what capabilities need to be established to perform these functions, and what resources need to exist to form these capabilities. There are different models that can be used here, but this paper follows a capability-based approach (CDG-CDM 2006) that is common within the Australian Defence context and can be easily used to describe many organisations. For example, the capabilities can be represented by a divisional structure within an organisation, not necessarily as major defence acquisitions. The resources will include people and machines. A usability study is situated at the interface of these two resources.

3.3. Operator's duties and functions

The purpose of a usability study needs to be linked with the organisation; that is, the study can be justified through its link or relationship to a capability, function, outcome, strategy, value, mission, or vision. A usability study for a new ATM interface, for example, can be justified in the form of improving the efficiency of performing certain functions – such as integrating and coordinating air elements – and to generate an operational effect or outcome – such as maintaining a certain level of safety in the airspace.

The relationship between the operator and the user-interface needs to be understood in more depth; thus the following diagram depicts this relationship within an organisational context.

Figure 2 The relationship between the organisation as a whole, the operator and the rest of the system.



Here, organisational objectives are prescribed as functions that need to be executed and that require people to perform certain roles. These functions and roles are normally summarised in the position description document for a job. An example of such a document will be given later on in this paper. The position description defines what is expected from a user. As such, users/operators are chosen according to the position description document, while these documents are also used to establish the key performance indicators of the user. The end-user interfaces with other end-users within the organisation via the machine (e.g. computer interface or software). This computer-based system enables them to interface with external users or stakeholders as well.

The previous two diagrams are essential to establish a methodology to communicate the benefits of a usability study to both the organisation as a whole and to the user him/herself.

The discussion above represents the thought process required to establish an organisational need for a usability study. A human factors researcher, usability analyst, or a cognitive

scientist would normally join the study after this need has been established and justified within the wider context of the organisation. The COBP emphasises the need to do this analysis so that the domain expert becomes aware of the wider organisational context. Studies have failed simply because this communication and understanding did not occur.

To use a hypothetical scenario to illustrate this point, imagine that a human factors specialist in one study only focuses on the study without understanding its purpose within the wider organisational context. While this hypothetical study was established to balance workload among operators, the human factors specialist focuses on finding the minimum model to measure workload instead. The focus, from a human factors perspective, would be scientifically valid. However, the minimum model to measure workload does not necessarily help in distributing workload. For example, task switching is an important factor if a re-distribution of workload is performed but the complexity and nature of task switching during redistribution of load is different from that occurring during normal tasks. This gap between the real and intended objective of a study can cause many studies to fail in the real world.

3.4. Understanding the Operator's Job

The duty specifications document of the operator's job – in their position description – is a valuable source of information. This document can be easily overlooked in a study despite its importance to link the operator's expected duties with the wider organisational objectives on the one hand, and the operator's key performance indicators on the other.

The Australian Standard Classification of Occupations (ASCO) provides a clear list of duties expected from an air traffic controller. The job of a controller is defined as to "*ensure the safe and efficient movement of aircraft in controlled airspace and aerodromes by directing aircraft movements*." (Air Traffic Controller - ASCO 2541-13).

This definition is not a surprise, as an air service navigation provider is judged on these two criteria: providing safe operations of the airspace and efficient use of airspace. For example, Air Services Australia's vision statement is "*Air Services Australia will be a safe and efficient provider of air traffic management and aviation rescue and fire fighting services with an international reputation for excellence*" (Air Services Corporate Plan 2010-2015, p.5).

Taking 'safety' and 'efficiency' as the two main groups of key performance indicators for an air traffic controller, a usability study needs to link the controller's key performance indicators, duties and the functions performed via the ATI. Eurocontrol defines three basic functions for a controller: Monitoring, Action and Planning. However, the literature does not show how these functions can be mapped to the controller's duties (ASCO 2541-13). In an attempt to close this gap, the following table maps these relationships and represents the strength of the relationship using '+' signs.

Table 1 Mapping out the Controller's duties to the Controller's three generic cognitive functions.

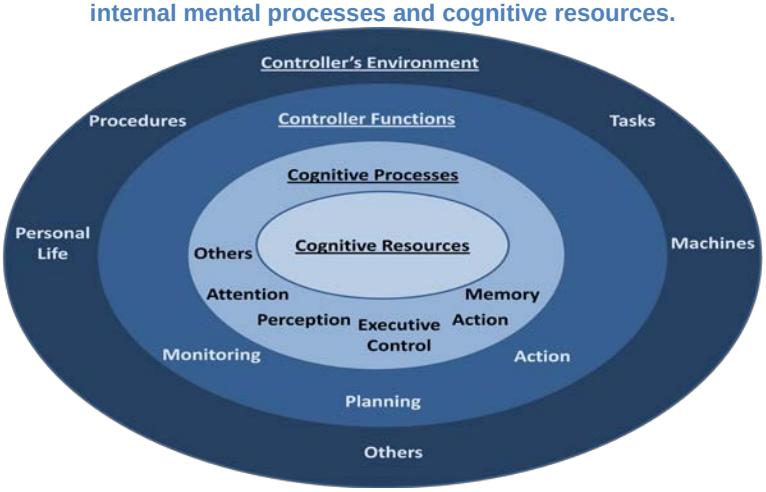
Controller Duty	Monitoring	Action	Planning
oversees the preparation and processing of aeronautical information necessary for the safety, regulation and efficiency of air navigation		+	+++
provides flight information for flight crews and air traffic services staff, such as wind direction and strength, details of cloud cover and temperature and altimeter settings	+++	+	
checks flight plans, position reports, flight levels, estimated arrival times at reporting points or destinations and authorises changes of flight levels and altitudes	+	+	++
controls aircraft movements in the air using radar or non-radar procedures and directing aircraft by radio	++	++	++
controls aircraft movements on aerodromes by issuing runway clearances and directing taxiing, take offs and landings	++	++	++

controls the operation of airport lighting systems such as runway and approach lights and aerodrome beacons	+++	
communicates with other air traffic control units to coordinate activities	++	
alerts airport fire crew and emergency or search and rescue services when aircraft are in difficulty	++	
organises search and rescue assistance to aircraft in distress may instruct air traffic control trainees and train licensed controllers upgrading their ratings Entry Skill Level	++	+++
organises search and rescue assistance to aircraft in distress	++	+++
may instruct air traffic control trainees and train licensed controllers upgrading their ratings Entry Skill Level		

3.5. Operator's cognitive functions

One can argue that all cognitive processes are used in one way or another by an air traffic controller. The following diagram illustrates the general relationship between the controller's job functions and underlying cognitive processes.

Figure 3 An Onion model connecting the Controller's organisational and external environment and his/her internal mental processes and cognitive resources.



Many theories in cognitive science assume the existence of "cognitive resources" that become depleted as cognitive functions are executed. These may be conceived as either a single resource (Kahneman, 1973; Case et.al., 1982; Townse & Hitch, 1995; Barrouillet & Camos, 2001) or as multiple resources (McCracken & Aldrich, 1984; Wickens, 1984; Basil, 1994; Meyer, 1997; Wickens, 2002). The diagram above uses an onion model to demonstrate the relationship between the exogenous factors such as the working environment and the functions the controller performs, as well as endogenous factors contributing to the controller's cognitive processes and cognitive load (CL). It is important to remember that none of the factors in the inner two circles can be measured directly in a usability study. As such, we can only use indicators for these factors as a proxy for the impact of the external environment and tasks on the controller's cognitive resources. An objective measurement of CL is complicated by a number of factors:

- *A lack of consensus as to what constitutes CL:* There is no consensus on what cognitive processes contribute to cognitive workload and how to measure them.
- *Variation across individuals*: The cognitive make-up, motivations, preferences and behavioural patterns of individuals vary widely and controllers or operators may

perform better or worse in certain perceptual and cognitive tasks. This can skew the determination of CL when considered as a uni-dimensional parameter.

- *Extraneous factors*: The cognitive load of the human operator is not only determined by application to monitoring, planning and execution tasks through the ATI, but also in communications with other operators, or in the deployment of cognitive resources to perceive other environmental stimuli.
- *Training and task efficiency*: Finally, the level of training of the operator in the use of the interface can cause variability in cognitive load if the interface is used inefficiently by performing unnecessary and time-consuming actions.

4. Designing the experiment

4.1. Dimensions of the usability study

ISO 9412 – Part 11 is the current standard for usability studies, presenting the extent to which a product can be used by specified users to achieve specified goals within a specified context of use, in terms of the effectiveness (fitness for purpose), efficiency (ease of use) and user satisfaction.

Safety-critical systems have a particular focus on safety. The nature of these systems requires continuous engagement from the user, be it in the form of monitoring, planning or the execution of decisions. Therefore, evaluating the usability of a SC-C2S cannot stop at interface usability alone, the cognitive processes involved in using the interface must be considered as well. The three objectives stated in the Standard intermingle with the different cognitive processes that an operator would deploy during an interaction with the interface. For instance, depletion of an operator's attention resource would adversely impact upon the effectiveness and efficiency of doing a task, and his or her perceived level of satisfaction with the interface would also likely decrease as a result of inattention or mental fatigue.

A traditional approach to the measurement of usability as defined in ISO 9421 Part 11, is through assessment of user performance and experience, in terms of:

- Effectiveness - assessed by successful completion of tasks
- Efficiency - measure by the time to complete tasks
- Satisfaction - a subjective measure of experience as reported by the user

While this approach can produce meaningful results, the objective measurement and comparison of interface solutions can be very difficult, since usage scenarios, task loads and cognitive demands from one usability test session to the next can vary widely.

In addition to an assessment of the above usability criteria, for this COPB we advocate an approach whereby usability is assessed in terms of the operational requirements of tasks, the situational context, cognitive resources and limitations of the human operator. To establish the relationship between these, we need to present some definitions:

a) *Situational awareness* (SA): The literature provides many variations in the way that situational awareness is defined. By consolidating these differing understandings, SA can be defined in terms of either: the process by which one continuously collects, analyses and prioritises usually dynamic behavioural, environmental and task-based information and integrates that information with declarative and procedural knowledge already residing within the operator's long term memory, (Andre, 1998; Dalrymple & Schiflett, 1997; Endsley, 1995a & 1998); or the product or state that one achieves as a result of this process. The product-state of situational awareness is often referred to as the operator's "mental picture" or "mental abstraction" of the context and scenario as it is

unfolding in real time (Adams, Tenney, and Pew, 1995; Billings, 1995; Endsley, 1995). Traditionally, SA is understood to comprise three levels; perception, comprehension and projection (Endley, 1995a & 1995b). In the interest of consistency, we adopt this basic definition here.

b) *Cognitive Load*: This is probably the most confusing concept in the literature. Cognitive scientists debate this concept with many variations– see above for examples. For the sake of simplicity, we will define cognitive load as the amount of cognitive resources needed to perform a task.

c) *Task Load*: In essence, one can see the task load as the component of cognitive load that is caused primarily by the complexity of the task at hand. Since we cannot measure this directly, a complexity measure of the air traffic scenario acts as an indicator of task load.

d) *Interface Load*: The design of the interface impacts the cognitive processes deployed by an operator when performing a task. The “interface load” is the term we coin to indicate the amount of resources depleted because of interface issues when everything else is maintained constant.

e) *Workload* is the concept coined to represent the amount of cognitive resources deployed by an operator as a result of all work related factors. This includes task load, interface load plus any other work-related demands on cognitive resources of the operator.

f) *Environmental Load*: This term refers to all the extraneous factors which attract and divide the attention of the ATI Operator. This may include factors that are directly relevant to the Operator’s job as well unrelated events or distractions.

The assumption here is that task, interface and environmental loads are additive; that is, if we can imagine fixing the task load, minimising possible sources of distraction and only varying the interface, the changes in the operator’s cognitive load is linear in the changes of the interface load. While this assumption can be debated by a few cognitive psychologists, without an acceptable cognitive model to measure these factors it would not be practical to undertake any objective study on usability and cognitive processes without this assumption.

The interaction of the controller and the interface creates a dynamic environment that requires the use of different components of the interface at specific situations. The usability of an interface needs to be evaluated by objective measures and indicators optimised for each of these components. This follows from our hypothesis that the usability of an interface can be assessed in terms of interface load, where this is a measure of the complexity of the human-machine interface *in situ* and depends both upon the quality of the computer interface and of the attributes of the human operator.

In essence, a multi-factorial measure of interface load can be derived by the rigorous objective measurement of a) the complexity of the scenario and demand of the task at hand and b) the various parameters influencing the cognitive load and situational awareness of the human operator. Thus:

Cognitive Load ≈ Work Load + Environmental Load

Work Load ≈ Task Load + Interface Load + Other Work Related Factors

When environmental load and other work related factors are maintained constant, the interface load can be estimated conceptually as:

Interface Load ≈ Cognitive Load - Task Load

One objective of a usability study is to estimate the interface load by understanding the differences in cognitive load when using different interfaces, and while fixing task load. Even

in situations where it is impossible to fix the task because of the dynamic nature of the environment, one can still infer the interface effect by measuring task load for different tasks.

4.2. Experimental variables and measurements

Determining the independent and dependent variables of a usability experience in safety-critical systems is a non-trivial task; especially when most of the variables of interest cannot be measured directly. In building the foundations of the experiment, it is more valuable to think of the concept of measurement in terms of Zeller & Carmines (1980) definition as being the "*process of linking abstract concepts to empirical indicants*" (p.2). From this perspective, we can select variables in a more holistic/comprehensive manner and make use of a range of non-intrusive or minimally intrusive data collection and analysis techniques to elicit a range of data types, minimise threats to the internal and external validity of the experiment and ensure, not only the quality and integrity of the data, but also the interpretations of the data.

In the context of evaluating ATIs, the variables have been divided into three categories: Interface Acceptability, Situational Awareness, and Cognitive Load.

Interface acceptability is a multi-dimensional concept closely related to the notion of usability. Nielsen (1993) and Schneiderman (1998) define five dimensions for acceptability; these are:

- Learnability: The level of ease with which users can accomplish basic tasks the first time they encounter the design.
- Efficiency: Once users have learned the design, efficiency is the speed at which they can perform tasks.
- Memorability: When users return to the design after a period of not using it, memorability is the level of easiness with which they can regain proficiency.
- Errors: The number and type of errors made by users, the severity of errors, and the ease of recovering from these errors.
- Satisfaction: How pleasant it is to use the design.

A mixed method approach that combines traditional qualitative and quantitative methods of inquiry will provide us with the most comprehensive method for evaluating each of these five dimensions. The first four dimensions can be objectively quantified by deploying assessment techniques while the fifth dimension can be evaluated through qualitative methods such as questionnaires or interviews with the user and/or behaviour observation techniques.

There are many *situational awareness* assessment techniques in the literature that are considered to be valid assessment techniques for the elicitation of 'indicant' data. Examples of these are listed in the following table:

Table 2 A summary of some key situation awareness tests and their characteristics.

Metric	Type of Technique
Situational awareness Global Assessment Technique (SAGAT) (Endsley, 1995b) – designed to assess the operators' perception of the elements, the comprehension of their meaning, and their ability to project future statuses. SAGAT-TARCON was devised specifically for ATC.	**Freeze Probe** - a task is frozen and participants are asked to respond to a series of questions based on their knowledge up until the 'freeze' moment (cue recall). For SAGAT-TRACON, it also involves a free recall task.
SALSA (Hauß, Gauss, & Eyferth, 2001) - "*was especially developed to measure SA in the ATC-domain. It pays special attention to the fact that*	**Freeze Probe** - as per above, but ATC-specific questions are asked based upon fifteen aspects of aircraft flight, such as flight level, ground speed, heading, vertical tendency, and conflict types. SMEs are required to rate each simulation to determine the

the relevance the elements of the task environment changes over time."	relevance of the test questions. There is also a cue recall task - questions for each freeze focus on one aircraft only.
Situational awareness Rating Technique (SART) (Taylor, 1990) - developed initially to assess the SA of pilots.	**Self Rating Technique** - administered post-task 7-point rating scale "*uses ten dimensions to measure operator SA: Familiarity of the situation, focussing of attention, information quantity, information quality, instability of the situation, concentration of attention, complexity of the situation, variability of the situation, arousal, and spare mental capacity."* (Salmon, et al, 2006)
Situation Present Assessment Method (SPAM) (Durso et al 1998) -designed to assess the SA of air traffic controllers.	**Real-Time Probe** - task related SA queries based on information appearing within or relevant to the environment are asked via a telephone call during task performance
SASHA (Jeannot, Kelly & Thompson 2003) - developed to assess the SA of Controllers using automated systems	**Real-Time Probe PLUS questionnaire** - consists of a SPAM-like series of scenario-related while-task questions and follows up with a post-task questionnaire.
Situational awareness Rating Scales (SARS) (Waag & Houck 1994) - designed for military aviation	**Self Rating Technique** - similar to SART, uses a 6 point rating scale to gather subjective performance ratings from participants.
Crew Awareness Rating Scale (CARS) **-** developed to measure the SA and workload of C2 commanders (Matthews, Beal & Pleban 2002)	**Self Rating Technique** - based on Endsley's model of SA, it uses a set of 4 questions to elicit information about each of the 3 SA levels, followed by an additional set of 4 questions about the mental workload involved in those SA tasks.
Quantitative Analysis of Situational Awareness (QUASA) (McGuinness 2004)	**Self Rating Technique PLUS Real-Time Probe** - participants are asked to respond to on-task SA true/false questions and then are asked post-task to rate their level of confidence in their responses.
Situational awareness Behavioral Rating Scale (SABARS) (Matthews, et al 2000, Matthews & Beal 2002)	**Observer Rating Technique** - subject matter experts are asked to observe participants on-task and rate their performance according to a number of specified behaviours.
NASA – Task Load Index (TLX) (Hart & Staveland, 1988) - used extensively for C2, cockpit and other mission critical systems to assess operator task performance.	**Self Rating Technique** - is a subjective workload assessment carried out on operator(s) working with various human-machine systems and derives a score based on 6 weighted subscales – mental workload, physical/temporal demands, etc.

The inherent danger of using these types of tests, however, is that the cognitive activities or processes relating to situational awareness and those of task performance can differ greatly (Endsley, 1995b). One can be situationally aware, but be a poor performer (Tenney, et al, 1992). Therefore, looking solely at the operator's output within specific scenarios is not sufficient if a goal of the study is to isolate the operator's situational awareness from the decisions and actions that he or she takes as a result of that "mental picture" or indeed any cognitive or psychological process (Adams, Tenney, and Pew, 1995; Billings, 1995; Endsley, 1995).

It is recommended that one makes use of techniques that elicit or provide greater insight into the cognitive thought processes being used to update this mental picture. When used in conjunction with the previously described techniques, psycho-physiological measurement techniques (e.g. process indices derived from eye tracking, talk-aloud tasks, speech production errors, and sensory equipment such as EEGs) may help researchers paint a more comprehensive picture of an operator's situational awareness and provide them with greater

means for analysing the extent to which an operator will make correct actions and good decisions based on his/her internally generated model of the air traffic control environment.

As noted above, the concept of *cognitive load* is an abstract concept that cannot be measured directly with any precision. Consequently, we are again guided by Zeller & Carmines' definition of measurement (1980) and rely on indicators to estimate the depletion of cognitive resources. Examples of such indicators include mental chronometric measures, which rely on response time in cognitive tests as indicators for measuring activities of cognitive processes.

While the exact nature of mental processes and factors contributing to an overall indication of cognitive load is unknown, a range of metrics and indices have been verified through multiple studies and some are now well established through use over several decades. We adopt a pragmatic approach to the measurement of cognitive processes contributing to CL which also provides scope for both elementary and a more sophisticated analysis in the context of human factors studies and end-user interface usability testing. This is premised upon the following principles:

- Use traditional, well-established and convenient forms of measurement where possible. These will provide the most reliable indications for more basic underlying processes (such as autonomic stress/relaxation response).
- Collect as much information as is feasible given capabilities of available measurement equipment and the level of intrusion of the operator involved in the usability test. Through further, off-line computational analysis of this data, more sophisticated indicators of cognitive processing can then be extracted.
- For reasons of operational simplicity, there may be a preference for a single uni-dimensional measurement of CL. However, evaluating and comparing the 'usability' of an interface is itself a multi-factorial problem which decision makers must evaluate in terms of the risks, costs and potential benefits. Similarly, cognitive load should be understood to comprise several components or dimensions which are influenced by (i) the nature of the task at hand, (ii) the situational awareness and decision-making requirements and (iii) the interface displays, controls, functions and design.

Following these principles and drawing upon current understanding and findings in the human factors, usability testing and related literature, we recommend evaluating some or all of the following physiological and mental processes for which evidence exists to demonstrate a contribution to overall "cognitive load". Measurement techniques for each of these can be found in the literature.

- *Psycho-physiological Arousal*: Measurements of arousal levels has a long history and are typically obtained through detection of changes in heart rate, respiration rate and skin conductance. Psycho-physiological arousal is a general measure of the level of activation or stimulation of the central and autonomic nervous systems, ranging from sleep and relaxed states, to waking states of alertness through to heightened states of stress and excitement.
- *Cognitive Overload and Stress*: Physiological and psychological indicators of arousal levels are also effective in the detection of stress and anxiety associated with "cognitive overload". In conjunction with other explicit or combined indicators of CL, indications of stress may be used to extrapolate from the continuum of normal cognitive load towards a maximum or positive extreme. In addition to measures such as those listed above, a pre-dominance of high frequency cortical Beta waves associated with concentrated thought activity can be indicative of cognitive overload.

- *Underload, Boredom and Fatigue*: It is well known that during highly repetitive tasks and under conditions of sparse stimulation (for example, long-distance driving on straight country roads) apathy and reduced attentiveness can set in. In this case, cognitive load will decrease and specific measurement of a “cognitive underload” state can provide a more complete picture of the entire CL continuum. A theoretical justification for this view is provided by Malleable Attentional Resources Theory (Young & Stanton, 2002) which posits a direct relationship between cognitive load and the availability of attentional resources. This theory provides a unified framework for understanding adverse performance effects due both to excessive cognitive demands and mental overload on the one hand, and insufficient stimulation and mental underload on the other. Slow (Delta, Theta or Alpha) wave activity in frontal regions of the cortex can indicate internalised cognitive processes associated with mental fatigue and disengagement. Increased eye-blink duration and frequency are also considered reliable indicators of drowsiness.
- *Task Engagement and Subjective Interest*: Within a normal operational range, studies have shown that measures of task engagement are correlated with cognitive load. The NASA Engagement Index derived as the ratio of Beta to Alpha plus Theta wave activity at central and parietal regions of the cortex is a well-studied indicator of engagement (Freeman et. al. 1999). In a similar category is “subjective interest”, or more specifically, the detection of ‘non-obligatory’ cortical responses to perceived stimuli, which are believed to indicate degree of subjective interest, or relevance to personal motivations and task orientations. One problem with using EEG based indications of engagement or interest as a measure of cognitive load, is that it may not provide much information for cognitive strain or *work*. These require measurement of additional parameters such as stress and mental fatigue.
- *Attentional Orientation*: The selective deployment of mental resources to endogenous calculations and thought processes, or to the perception and recognition of exogenous stimuli is an important factor influencing mental performance. Measurement of attentional orientation, notably through comparison of posterior Alpha waves during different tasks can give further insight into factors contributing to situational awareness and cognitive load (see for example Ray & Cole, 1985). Temporary in-attentiveness to environmental stimuli due to focused cognitive processing, or alternatively, external distractions to cognitive operations, can both adversely impact situational awareness and result in increased mental exertion to maintain SA levels and rational decision-making capability.
- *Working Memory (WM) Capacity utilisation*: Exhaustion of resources can be tested by WM span tests such as through reading comprehension or letter/code recall in a dual-task test requiring splitting of cognitive resources between memory rehearsal and cognitive processing. Similar experiments have also been conducted to test numerical and spatial memory retention. Current knowledge of functional neuroanatomy supported by fMRI imaging studies strongly suggests this model of dynamic cognitive resource allocation as a fundamental mechanism underlying measurable psycho-physiological phenomena associated with cognitive load.

4.3. Task scenario design

Developing an appropriate real-time scenario design is challenging, as it must balance between being suitably credible for the operators and providing the analyst or researchers with the means to collect the desired data (The NATO Code of Best Practice for Command and Control Assessments, 1998). For Air Traffic Control Officers (ATCOs), an ATM system should support the cognitive processing required to obtain a high state of situational awareness, and help them maintain the “air traffic picture” during task performance. To achieve this end, we use Endsley’s 3-level model of situational awareness (1995b) that can be mapped back to specific KPI and operational task requirements.

- Perception: We define 'P-tasks' within the experimental paradigm to elicit this aspect of situational awareness. In the ATM domain, this represents an awareness of relative positions of planes on different headings and levels within a sector, weather conditions and other events.
- Comprehension: 'C-tasks' are designed to test the user's ability to comprehend the significance of a situation in terms of operational or mission objectives. As an example from ATM, this level of situational awareness is emphasised when a potential separation violation arises between aircraft en-route or during final approach to an airport and the ATCO must immediately comprehend the significance of this situation within the overall context.
- Projection: 'R-tasks' are used to elicit the operators’ ability to estimate flight-paths, predict future conflicts and plan contingencies to these through vectoring and relaying instructions to the pilots.

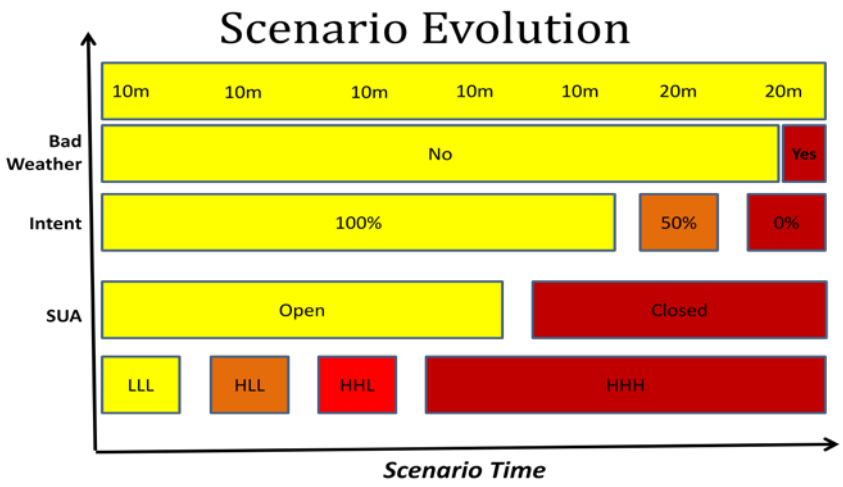
The air traffic scenario(s) that will be used in the experiments need to somehow be encoded for these three levels especially as they are increasing in complexity. Moreover, the level with higher complexity subsumes those with lower complexity. Therefore, when designing P-tasks, we can – to a degree – inhibit comprehension and projection. Realistically, the Controller will still do a level of comprehension and projection within P-tasks, but that would be the minimum level required. Similarly, when designing C-tasks, we need to inhibit projection or maintain it at the minimum possible level. However, as with projection, we realise that we are not able to inhibit the subject from performing monitoring functions. The three type of tasks can be represented as follows:

- P-tasks: low density of traffic with sufficient (large) spacing
- C-tasks: low density of traffic with small spacing
- R-tasks: high level of traffic with small spacing and high conflicts

The scenario may also need to accommodate for other factors such as special use of airspace (SUA) effects, intent information, weather, etc. An example of a scenario design to accommodate for these effects over time is presented below, where the top box represents the length (in minutes) of the time elapsed since the start of a scenario. It demonstrates a scenario with an increasing level of complexity, placing the controller in a critically complex situation during the last 20 minutes of the scenario. The use of ‘L’ and ‘H’ represents low and high: density, spacing, and possible conflicts.

The following figure is an example. Sometimes we can introduce the controller directly to the middle of this scenario, start with high complexity then decay it, or configure other setups that match the objectives of the study. A key aspect that we need to emphasise here is that the scenario design cannot be done in isolation of the measurements. The variables that need to be measured will determine how this scenario is designed and how the situation should unfold.

Figure 4 An example of a scenario design. The x-axis represent time, the y-axis represents block-events, while each box in the figure represents corresponding time-span an event type will take within a scenario.



Another important aspect that should be considered in the design of a scenario is the role of the controller in the unfolding of events. Over time, the controller's interactions can change the dynamic of the scenario. For example, a controller may decide to vector an aircraft which may affect the spacing of existing traffic, causing a P-task to become a C-task early. This aspect takes a lot of time to test possible permutations of actions at the time of scenario design. One cannot guarantee that the intended unfolding of events will proceed as planned, but it is worth spending the time to minimise unanticipated changes in the sequence.

5. Conclusion

This paper presents a methodology to link and position a usability study with and within the organisation. Within the domain of safety-critical command and control systems, usability studies are normally justified based on criticality and safety. This paper supplements this with a methodology that goes beyond a simple justification. The methodology establishes the foundations for making an economic case for usability studies in these systems. We demonstrated how an organisation's strategic objectives can be linked to cognitive measurements within an experimental context.

6. Acknowledgement

Research that is being presented in this paper is part of a wider collaborative project between Thales Australia and the University of New South Wales at the Australian Defence Force Academy in Canberra. We acknowledge the funding from Thales Australia.

References

Adams M.J., Tenney, Y.J., & Pew, R.W., (1995). Situational awareness and the cognitive measurement of complex systems. *Human Factors*, 1995, vol. 38, pp.85-104.

Alberts, D.S., (2002) *Code of Best Practice for Experimentation*. CCRP Publication Series.

Airservices Corporate Plan 2005-2010, (2005), Airservices Australia, http://www.airservicesaustralia.com/media/corporatepubs/docs/corporateplan/corporateplan2010-2015.pdf

Barrouillet, P., and Camos, V., (2001). Developmental increase in working memory span: Resource sharing or temporal decay? *Journal of Memory and Language*, 45, 1-20.

Basil, M.D., (1994) Multiple resource theory I. *Communications Research,* 21(2).

Billings, C. E., (1995). Situational awareness measurement and analysis: A commentary. *Proceedings of the International Conference on Experimental Analysis and Measurement of Situational awareness,* Embry-Riddle Aeronautical University Press, FL.

Bowley, D., Comeau, P., et al., (2006) *Guide for Understanding and Implementing Defense Experimentation* (GUIDEx)-Version 1.1, The Technical Cooperation Program (TTCP).

Case, R.D., Kurland, D.M. and Goldberg, J., (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology,* 33, 386-404.

CDG-CDM, Capability Development Group (2006), *Capability Development Manual*, CDG, DOD, http://www.defence.gov.au/publications/dcdm.pdf

DOD-NCW, (2007). *Network Centric Warfare Roadmap*, Capability Development Group, Australia.

Durso, F.T., Hackworth, C.A., Truitt, T., Crutchfield, J., Manning, C.A., (1998). *Situational awareness as a predictor of performance in en route air traffic controllers*. Air Traffic Quarterly, 6, 1-20.

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

Endsley, M.R., (1995b). Toward a theory of situational awareness in dynamic systems. *Human Factors* 37(1), 32–64.

FAA, (2007). Fact Sheet: *NextGen*. Federal Aviation Authority. http://www.faa.gov/news/fact_sheets/news_story.cfm?newsid=8145.

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

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

Hauss, Y., Gauss, B., & Eyferth, K., (2001). SALSA - A new approach to measure situational awareness in air traffic control. *Focusing Attention on Aviation Safety.*

ISO 9241-11:1998 (1998) *Ergonomic requirements for office work with visual display terminals, Part 11: Guidance on usability*, International Organisation for Standardization.

ISO/TR 16982:2002 (2002) *Ergonomics of Human-System Interaction – Usability Methods Supporting Human-Centred Design*, International Organisation for Standardization (ISO).

ISO/TR 9241-100:2010 (2010) *Ergonomics of Human-System Interaction – Part 100: Introduction to Standards Related to Software Ergonomics*, International Organisation for Standardization (ISO).

Jeannott, E., Kelly, C., Thompson, D., (2003). The development of Situational awareness measures in ATM systems. *EATMP report.* HRS/HSP-005-REP-01.

Kahneman, D., (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Matthews, M.D., Beal, S.A., (2002). Assessing Situational awareness in Field Training Exercises. U.S. Army Research Institute for the Behavioural and Social Sciences. *Research Report 1795*.

Matthews, M.D., Pleban, R.J., Endsley, M.R., & Strater, L.D., (2000). Measures of Infantry Situational awareness for a Virtual MOUT Environment. *Proceedings of the Human Performance, Situational awareness and Automation: User Centred Design for the New Millennium Conference,* October 2000.

Matthews, M.D., Beal, S.A, & Pleban, R.J., (2002). *Situational awareness in a Virtual Environment: Description of a Subjective Measure.* (Research Report 1786). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McCracken, J.H. & Aldrich, T.B., (1984), *Analysis of selected LHX mission functions: Implications for operator workload and system automation goals.* (Technical note ASI 479-024-84(b)), Fort Rucker, AL: Anacapa Sciences, Inc.

McGuinness, B., (2004). Quantitative Analysis of Situational Awareness (QUASA): Applying Signal Detection Theory to True/False Probes and Self-Ratings. *Command and Control Research and Technology Symposium: The Power of Information Age Concepts and Technologies*, San Diego, California, 15th – 17th June 2004.

McMillan, J.H. & Schumacher, S., (1993) *Research in education: A conceptual introduction*. Harper Collins Publishers, New York: NY.

Meyer, D.E., (1997). A computational theory of executive cognitive processes and multiple-task performance: 1. Basic mechanisms, *Psychological review,* 104(1).

NATO COBP for Command and Control Assessment. Washington, DC: CCRP. (1998).

Nielsen, J., (1993). *Usability engineering.* Morgan Kaufmann.

Ray W.J & Cole H.W. EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes, Science. 1985, 228(4700):750-2

Salmon, P., Stanton, N., Walker, G., and Green, D., (2006) Situational awareness measurement: A review of applicability for C4I environments, *Applied Ergonomics*, 37(2), 225-238.

SEASAR, (2011). The European ATM Master Plan, https://www.atmmasterplan.eu/

Schneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Addison Wesley.

Taylor, R.M., (1990). Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design. In *Situational Awareness in Aerospace Operations* (AGARD-CP-478) 3/1 –3/17, Neuilly Sur Seine, France: NATO-AGARD.

Tenney, Y.J., Adams, M.J., Pew, R.W., Huggins, A.W., and Rogers, W.H., (1992). *A principled approach to the measurement of situational awareness in commercial aviation*. NASA contractor report 4451, Langley Research Center: NASA.

Towse, J.N. And Hitch, G.J., (1995). Is there a relationship between task demand and storage space in tests of working memory capacity? *Quarterly Journal of Experimental Psychology*, 48A(1), 108-124.

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

Wickens, C.D., (1984). "Processing resources in attention", in Parasuraman, R. & Davies, D.R. (Eds.), *Varieties of attention*, 63-102. New York: Academic Press.

Wickens, C.D., (2002). Multiple resources and performance prediction. *Theoretical Issues in* Ergonomics Science, 3(2), 159-177.

Young, M.S. & Stanton N.A. (2002) Malleable Attentional Resources Theory: A New Explanation for the Effects of Mental Underload on Performance, Human Factors 44: 365

Zeller, R.A. & Carmines, E.G., (1980) *Measurement in the social sciences: the link between theory and data.* CUP Archive.